SPACE, TELECOMMUNICATIONS AND RADIOSCIENCE LABORATORY

STARLAB
DEPARTMENT OF ELECTRICAL ENGINEERING / SEL
STANFORD UNIVERSITY • STANFORD, CA 94305

PHASE MEASUREMENTS OF VERY LOW FREQUENCY SIGNALS FROM THE MAGNETOSPHERE

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

By Evans Paschal

January, 1988

PHASE MEASUREMENTS OF VERY LOW FREQUENCY SIGNALS FROM THE MAGNETOSPHERE

A DISSERTATION

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING

AND THE COMMITTEE ON GRADUATE STUDIES

OF STANFORD UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

By

Evans Wayne Paschal

January 1988

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

(Principal Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Approved for the University Committee on Graduate Studies:

Dean of Graduate Studies

ABSTRACT

The usual methods of spectrum analysis applied to analog tape recordings of very low frequency (VLF) signals extract only magnitude information and ignore phase information. A digital signal processing system using a recorded constant-frequency pilot tone has been developed which can correct tape errors due to wow and flutter, and reconstruct the signal phases. Frequency shifts are corrected during analysis by interpolating between spectral points in the windowed Fourier transform, and the output phases of the synthesized filters are corrected for timing errors. Having signal component phases as well as magnitudes doubles the available information.

Whistler-mode signals from the VLF transmitter at Siple Station, Antarctica, are analyzed as received at Roberval, Quebec. The phase of a non-growing signal is found to give a less-noisy measure of duct motion than Doppler frequency shift, with improved time resolution. Correlations are seen between variations in the whistler-mode phase delay and the earth's magnetic field component D. They are interpreted as Pc 2 micropulsation transients, short compared to the length of the field line, which propagate from equator to ground as Alfvén waves.

Pulses with temporal growth show an advance in relative phase with time, indicating a positive frequency offset from the transmitted signal. This offset is sometimes seen even at the beginning of a received pulse, an effect not explained by any current model of cyclotron-resonant wave-particle interactions. Pre-termination triggering of an emission always occurs after a phase advance of 1.5–3 revolutions. Instantaneous frequency measurements show that all emissions, even termination fallers, begin above the frequency of the triggering signal, and that the transition from a signal to a termination emission 100 Hz higher may occur in less than 5 ms. Other phase effects give clues to the mechanisms of sideband generation, suppression, entrainment, and whistler precursors.

Magnetospheric signals with line structure are analyzed, and found to have little or no connection with possible power line harmonic radiation. A model is developed to explain them based only on magnetospheric processes—growth, suppression, echoing, and multipath coupling.

PREFACE

The idea of building a digital analysis system for VLF data first occurred to me in 1974. At that time I was interested in retrieving weak, coherent signals from a noisy background. We had started looking at recordings made at our receiving site in Roberval, Quebec, of signals from the Siple Station VLF transmitter, which had just begun operation the previous year. The transmitter had already produced some exciting effects, but its capabilities were still largely unknown. We hadn't learned the best frequencies and times of day to transmit, and much of the time we couldn't see any signals at Roberval. Dyke Stiles used a Sigma 5 computer (long since gone) to analyze a small amount of VLF data and demonstrated the power of digital signal analysis, even though his computer was not particularly well suited to this task. Why not build a special system for VLF data analysis, one that could average weak but repetitive signals and pick them out of the noise?

A proposal was submitted to the NSF in 1976 for the purchase of equipment and parts for the system. I thought it would take about a year to put together. Work started early in 1977. As is often the case, the work was found to take much longer than expected, even excluding time spent on other projects along the way.

The impact of the digital analysis system is likely to be quite different from what was envisioned in the original proposal. At that time the value of phase measurements was not fully appreciated. The phase information present in VLF data is mentioned in the proposal, but rather parenthetically, and only in regard to its need when integrating weak signals. The use of signal phase to identify and classify signals, to make accurate frequency measurements, and in the study of wave-particle interaction mechanisms is the most significant contribution of the present system.

Phase analysis is a valuable tool for certain kinds of signals—those from a coherent source such as a VLF transmitter, for example. Phase information gives a new perspective from which to view these signals, effectively doubling our information about them. Many of the signals I have analyzed show new and interesting, and in some cases quite unexpected, behavior when their phase information is included. However, if any new phenomena are found in these pages, it is only because I have had the good luck to be one of the first to look.

I would like to thank those who have had faith in me and encouraged me in various VLF projects over the years. I would especially like to thank Professor Emeritus John Katsufrakis, under whose able management of field programs I first became interested in these studies; and Bill Trabucco, Old Antarctic Explorer and longtime friend with whom I have learned receivers and transmitters, field operations and why not to trust data taken on New Year's Day. I thank Professors Robert Helliwell and Donald Carpenter, and spectrum analyst Jerry Yarbrough, for many stimulating and illuminating conversations. And I thank my wife Freddie, whose patience and encouragement has helped to bring this work about.

Equipment for the digital analysis system was purchased under grant DPP76-15678 from the Division of Polar Programs of the National Science Foundation. The NSF also funds the operation of Palmer, Siple, and South Pole Stations in Antarctica, and Roberval and Mistissini in Quebec, Canada. The magnetometer data used in Sec. 3.4 was kindly provided by L. J. Lanzerotti and C. G. Maclennan of Bell Laboratories.

Ev Paschal
Dec. 6, 1987

CONTENTS

CONTENTS

TABLES

ILLUSTRATIONS

ILLUSTRATIONS

1. INTRODUCTION

1.1 The Nature of VLF Research

Stanford University has been engaged in the study of very-low-frequency (VLF) radio wave propagation since the mid 1950's. These VLF waves include both those that travel in the earth-ionosphere cavity (sub-ionospheric waves) and those that penetrate the ionosphere and travel through the magnetosphere above (*whistler-mode* waves).

Passive Studies. Many of these signals are of natural origin. The archetypal signal is the *whistler*. A whistler starts when a lightning stroke at the surface of the earth creates an electromagnetic impulse. Some of its energy travels under the ionosphere and is heard in a distant radio receiver as a click or *spheric*. Some of the energy from the stroke may penetrate the ionosphere and, if conditions are right, will be guided along a path parallel to the earth's magnetic field, coming down to the surface again in the opposite hemisphere at the magnetic conjugate point. Because of dispersion along the path, the magnetospheric signal is heard as a gently falling tone or whistler. When the whistler reaches the opposite hemisphere some of its energy may not penetrate the ionosphere but be reflected back along the magnetospheric path. In this way whistler *echoes* are generated. The whistler is guided approximately along a field line because of the anisotropic refractive index of the magnetospheric medium, an effect due to the presence of thermal (low-energy) electrons trapped by the magnetic field. However, field-aligned density irregularities or *ducts* also seem to be necessary in order for a whistler to be able to penetrate the opposite ionosphere and be heard on the ground. See *Helliwell* [1965] for the history and theory of whistlers.

Much of the early work in the VLF field involved the study of whistlers. For instance, by measuring the frequency of minimum dispersion or *nose frequency* of the whistler and the time delay between the causative lightning stroke and the arrival of the whistler we can determine the magnetic latitude of the path and the electron density along it. If several whistler ducts are present it is possible to map magnetospheric plasma density as a function of magnetic latitude. If whistlers continue for several hours it is possible to plot the motion of the ducts and infer the presence of large-scale electric fields in the magnetosphere. Whistler studies have followed the motion of plasma in the magnetosphere during magnetic storms, and have defined the flow of plasma between the ionosphere and magnetosphere.

There are other types of naturally-occurring signals besides whistlers. Whistlers sometimes trigger the generation of *emissions*, signals of magnetospheric origin due to the interaction of the whistler with energetic electrons. These *wave-particle interactions* are thought to be due to cyclotron (transverse field) resonance between energetic electrons spiraling along the magnetic field in one direction and a whistler-mode wave (circularly polarized) going in the opposite direction. Sometimes an emission can echo and trigger another emission. Sometimes the interaction may be strong enough to be self-sustaining. *Periodic emissions*, *chorus*, and *hiss* are magnetospheric signals probably due to these processes. See *Park and Carpenter* [1978] for a review of studies using natural signals.

Active Studies. Much of our current work involves the use of man-made signals from VLF transmitters. With these signals we can actively probe the magnetosphere. We control what goes in and monitor what comes out, and we don't have to wait for nature to provide a signal. These active experiments have benefitted especially from the establishment of the VLF transmitter at Siple

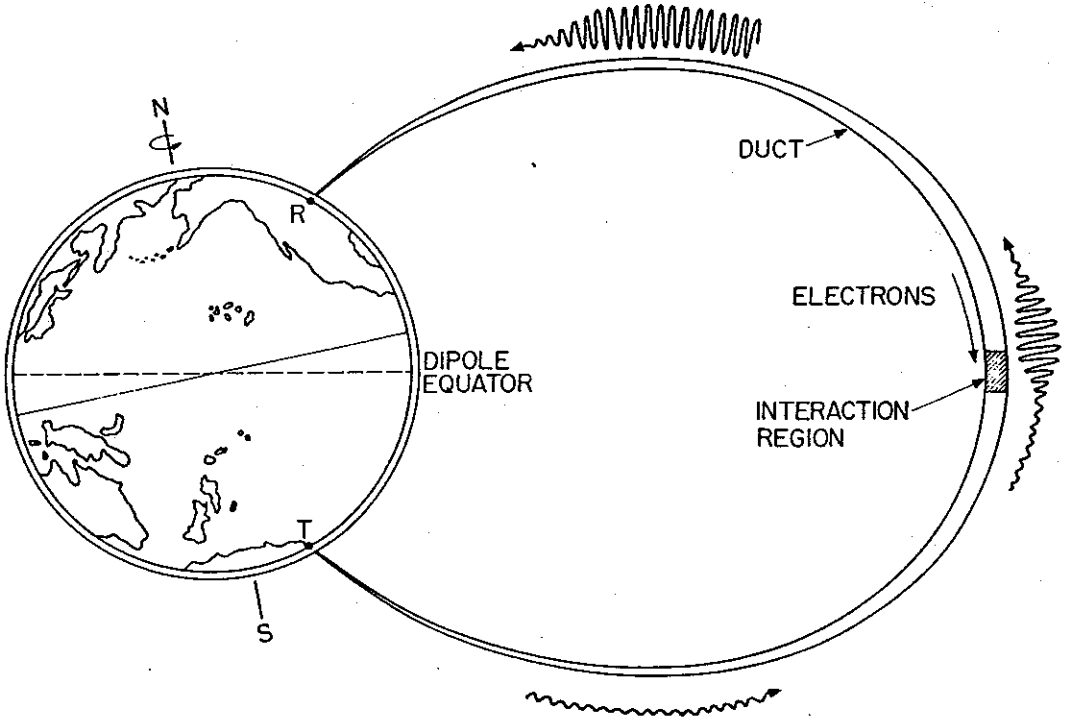

Figure 1.1. A whistler-mode signal from the transmitter, T, at Siple Station, Antarctica, moves through the magnetosphere on a field-aligned tube or duct of enhanced ionization. In the interaction region at the top of the path it undergoes cyclotron resonance with energetic electrons moving in the opposite direction, extracting some of their energy and becoming amplified. The amplified signal then travels back to the surface and is heard at the receiver, R, in Roberval, Quebec.

Station, Antarctica. This transmitter is used mostly at frequencies from 1 to 6 kHz, and can radiate several kilowatts of power. (The transmitter can put up to 170 kW into the antenna, but antenna efficiency is only a few percent.) The transmitter is also relatively broadband, with a full-power bandwidth of about 500 Hz.

We are particularly interested in studying wave-particle interactions. A whistler-mode signal from the Siple transmitter may resonate with energetic electrons it encounters, as shown in Figure 1.1. The *interaction region* where the resonance occurs is at the top of the path, near the equatorial plane. The interaction can cause the particles to give up energy to the wave, which has several effects. First, the wave may be amplified by 30 dB or more. Second, if conditions are right, the magnetospheric amplifier may become self-excited and turn into an oscillator, generating an emission at a frequency different from the input signal, or continuing after the termination of the input signal. Finally, the pitch angles of interacting electrons, normally trapped in the magnetosphere by the earth's magnetic field, may be decreased sufficiently that the electrons fall into the ionosphere and create a patch of enhanced ionization, x-rays, and light.

At the present time we have considerable information on the frequencies and power levels of signals that produce wave-particle interactions. Using the Siple transmitter we can stimulate whistler-mode signal amplification relatively easily and reliably. The details of the amplification process are still poorly understood and are the object of current study. The SEEP satellite experiment has observed particle precipitation into the ionosphere caused by signals from VLF transmitters [*Imhof*

et al., 1985]. We are still looking for ionospheric effects from this man-made precipitation, though we have good evidence for effects caused by whistlers. See *Helliwell and Katsufrakis* [1978] for a review of some of the work using VLF transmitters.

On-Site Signal Analysis. In a few cases it is possible to build special equipment to measure specific signal properties at a receiving site and answer specific scientific questions. For instance, signals from VLF transmitters have been used for many years at field stations to ensure the accuracy of local clocks. The phase of the sub-ionospheric signal is the important property here. The transmitted frequencies of some VLF stations are derived from primary frequency standards. By comparing the phase of the received signal against the phase of a similar signal synthesized from a local standard, any drift in the local standard can be measured and errors in local clocks can be kept to a few microseconds. (Initially setting a local clock to this accuracy is another problem.) As an extension of this technique, we are now starting to use improved phase comparators to detect phase changes in sub-ionospheric signals produced by particle precipitation from the magnetosphere (Trimpi events), and so identify the times and locations of such events.

Another example is *Leavitt's* [1975] frequency-tracking direction finding system, which can estimate the direction of arrival of whistler-mode signals. The amplitude and phase relationships of signals received on two perpendicular loop antennas and a vertical whip antenna are used here. Leavitt compares the amplitudes of signals from the loops to estimate signal bearing. However, only those loop components which are in phase with the signal from the vertical whip are used, and quadrature components are rejected. This tends to reduce azimuth errors caused by the elliptical polarization of incident waves. His system has been used to try to identify the exit points of whistler ducts.

Finally, systems have even been invented to automatically detect and classify whistlers on the spot. The frequency dispersion or df/dt characteristic is the important signal feature here. As an example of a recent effort, *Okada et al.* [1977] have developed an instrument that uses two frequency discriminators covering the ranges 4.66–5.86 kHz and 3.79–4.66 kHz. When a whistler is received, first the upper discriminator and then the lower one emits a characteristic response. The two responses are cross-correlated, and the delay time from the first to the second gives the dispersion of the whistler. They have used this instrument in a system that measures the direction of arrival and polarization of whistlers [*Okada et al.*, 1981].

Field Recording and Laboratory Analysis. In most cases, however, signals at a field station must be recorded *in toto* with the plan of analyzing their content at some later time. There are two reasons for this. First, it is not often known ahead of time what signal characteristics will prove scientifically interesting. Second, even if some interesting characteristics are known, it is usually impractical to build complicated, special-purpose equipment to measure only them. We have two tasks: to record the signals in the field with sufficient fidelity to preserve the information they contain, and then to return to the laboratory and extract that information using some general-purpose analysis system.

I have developed an analysis system which can process field recordings in a new way. For the first time, signal phase measurements can be routinely made from analog tape recordings, in addition to the usual magnitude measurements. Finding the phase of signal components, *phase analysis,* doubles the information available in certain cases. In this report I will describe the techniques used in phase analysis, and apply them to some current problems.

1.2 Fundamentals of Spectrum Analysis

Before we discuss the measurement and use of signal phase information, we need to introduce the general topic of *spectrum analysis*. This is the method most commonly used to study VLF signals. Spectrum analysis involves the mapping of a one-dimensional signal waveform onto a two-dimensional frequency-time space in order to isolate, identify, and study individual signal features. It might be more properly called "signal" analysis, since it involves more than just looking at the Fourier transform of the waveform, the "spectrum." However, spectrum analysis is the common name, and the machines that do the work are called spectrum analyzers. To get started, let's look at two complementary aspects of the signal: the waveform in the time domain, and the spectrum in the frequency domain.

The Time Domain and the Waveform. An obvious way to study a signal is to graph its value (voltage, pressure, position, etc.) versus time, and examine the graph for relevant features. Waveform analysis is used to study brief, usually non-repetitive, events such as transients in electric power distribution systems, or the movement of prices on the stock market. There are even reliable reports [*Randi*, 1982a, 1982b] of a man who can identify the title and composer of a work of classical music by examining the grooves (waveform) in a phonograph record.* However, direct examination of the waveform doesn't tell us everything.

Let us take as a problem the identification of the pitch of a given note in a recorded symphony on a phonograph record. Assume we have identified that place in the waveform where the given note was played, perhaps by listening to the record and stopping it when we reach the required passage. We can look at the waveform at this point and may be able to identify our note by eye, depending on the number of other sounds present at the same time. However, unless the note was played by a solo flute (whose sound has a simple structure), the waveform will be very complex, full of ripples of various periods all interfering with each other. Estimating the pitch or frequency of this note by, say, counting zero-crossings in the waveform is likely to be pretty inaccurate.

The Frequency Domain and the Spectrum. As an alternative to waveform analysis, we can look at the *spectrum* of the signal. Given a signal waveform $x(t)$, its spectrum $X(f)$ is defined as the Fourier transform of $x(t)$, given by [*e.g., Bracewell,* 1965]:

$$X(f) = \int_{-\infty}^{+\infty} x(t) e^{-j2\pi ft}\, dt. \tag{1.1}$$

While the real function $x(t)$ describes the waveform as a function of time, the complex function $X(f)$ specifies the spectrum of the signal as a function of the frequency variable f. When we express $X(f)$ at a given frequency in the form $X = A\exp(j\phi)$, we will refer to the modulus $A = |X|$ as the *magnitude* of the signal, and the argument $\phi = \angle X$ as the *phase* of the signal, at that frequency. We may also use the word *amplitude,* but the reader should know that this last term is somewhat ambiguous. An engineer may say "amplitude" when he means the magnitude of a spectral component, but he may also mean some general measure of the size of a waveform, its rms

* There is nothing mysterious about this. The man, Dr. Arthur Lintgen, is apparently a music buff and can identify post-Mozart, classical, fully-orchestrated compositions from his familiarity with them. He was tested with a number of such records and identified them successfully. He declared an additional record to be "gibberish" and "disorganized"—it was an Alice Cooper number inserted as a control.

or peak voltage, for example. (Mathematicians sometimes use "amplitude" to mean the argument of a complex number; that is, the phase ϕ in our case. We will never use it with this meaning.)

Since the waveform $x(t)$ is a real-valued function, the spectrum is hermitian; that is, spectral values at negative frequencies are just the complex conjugates of those at corresponding positive frequencies, $X(-f) = X^*(f)$. We can restrict ourselves to only positive frequencies with no loss of generality. For a signal of finite length such as a phonograph recording we may also limit the domain of integration to the times between the beginning and end of the record, defining the waveform $x(t)$ to be zero at other times.

The squared magnitude of the spectrum $|X(f)|^2$ shows the distribution of power versus frequency summed over the entire recording. If we calculate the spectrum of the waveform on our phonograph record we may be able to tell the key in which the music was played by looking for the presence or absence of signal components at specific frequencies (specific notes of the scale). We may even be able to tell if there was a bassoon in the orchestra by looking for signals at the lowest frequencies. However, we cannot find the frequency of a given note by examining the spectrum of the entire recording because it is impossible to separate the contribution of this one note, which occurred for only a brief interval at a given time, from all the other notes which were played before and after.

Conversely, even if we were sure that there was a bassoon in the orchestra playing a note at a given pitch we would not be able to tell exactly when it was played. The problem here is that the time of occurrence of a signal element is encoded in the phase information in the spectrum, and this is hard to interpret in any but the simplest cases. For instance, assume that we have a short signal element $v(t)$, representing a certain note played by a given instrument, and its spectrum $V(f)$. If this particular note had been played a time τ later, then its spectrum would be given by the function $V(f)\exp(j2\pi f\tau)$, the transform of $v(t+\tau)$. Note that the magnitude of each spectral component is unchanged but the phase of a component at frequency f has been advanced by $2\pi f\tau$ revolutions. The phase advance is proportional to frequency, and thus the shift in time of the given note is translated into helicity in the spectrum, a winding up of the spectral phase as a function of frequency. In the frequency domain we may be able to tell the key and mode of a piece of music, but the time that a given note was played is represented in the spectrum phase structure in a complicated and largely indecipherable way.

The Frequency-Time Plane and the Spectrogram. We see that these two representations, waveform and spectrum, give us different views of the signal. Examination of the waveform is useful in those cases where brief events such as transients are important. Examination of the spectrum is useful where the structure of some repetitive, continuing process is to be extracted. Neither approach allows us to easily uncover much information about those signal elements which last long enough to have an interesting structure in frequency (which structure may change with time) and yet are brief enough that their time of occurrence is also important.

Many interesting signals show such dynamic, non-stationary behavior. How can we analyze them? There are two possible solutions to this problem. One approach is to truncate the waveform $x(t)$ in time, eliminating all notes occurring before the desired one, and all music played after it, and calculate the spectrum of this truncated signal. This process is called *windowing*. To do this we choose some truncating or *weighting** function $w(t)$ which is zero (or suitably small) outside an interval of appropriate duration $[-\frac{1}{2}\Delta t, \frac{1}{2}\Delta t]$, multiply $x(t+t_0)w(t)$ to select a signal segment

* We will follow the convention of *Nuttall* [1981], calling the temporal function $w(t)$ the *weighting* function, and its transform, the spectral function $W(f)$, the *windowing* function.

around the desired time t_0 (note that we view the window as stationary in time, and the signal as moving through it), and then calculate the windowed spectrum as the Fourier transform

$$S(t_0, f) = \int_{-\infty}^{+\infty} x(\tau + t_0) w(\tau) e^{-j2\pi f \tau} \, d\tau. \tag{1.2}$$

The windowed spectrum $S(t_0, f)$ of the truncated waveform is a function of two variables since its value depends on the segment time t_0 as well as the frequency f. Since we can calculate the windowed spectrum for any given value of t_0, it is proper to think of it as a variable and not just a parameter. The squared magnitude $|S(t_0, f)|^2$ at the given time t_0 shows the distribution in frequency of the music power of our isolated note. If the timbres (harmonic structures) of the instruments playing at this time are not too complex we may hope to identify the pitch of the note.

We can also express the windowed spectrum as a convolution in the frequency domain. If the *windowing* function $W(f)$ is the Fourier transform of the weighting function $w(t)$, then we can write $S(t_0, f)$ as

$$S(t_0, f) = \int_{-\infty}^{+\infty} X(\nu) e^{j2\pi\nu t_0} W(f - \nu) \, d\nu \equiv X(f) e^{j2\pi f t_0} * W(f) \tag{1.3}$$

where the asterisk $(*)$ stands for convolution. Written in this form it is easy to see the effect windowing has on our ability to resolve signal components closely-spaced in frequency. If a signal $x(t)$ is of the form $\cos(2\pi f_1 t)$, then (the positive-frequency half of) its spectrum $X(f)$ will be $\delta(f - f_1)$; that is, zero everywhere except for an impulse at frequency f_1. However, this sharp line spectrum in $X(f)$ gets smeared out in the windowed spectrum and we find $S(t_0, f) = \delta(f - f_1) * W(f) = W(f - f_1)$ (again considering only positive frequencies and ignoring the phase $j2\pi f_1 t_0$). The sharp line from $X(f)$ now has the shape of $W(f)$ moved up to frequency f_1. If there are other signals near f_1 we may be unable to separate them. For better frequency resolution we would need a window function $W(f)$ which is narrower in frequency, which in turn entails a weighting function $w(t)$ that lasts longer in time. Thus there is a tradeoff between frequency resolution and time resolution. We will return to windows and their effect on frequency resolution in Section 2.5.3.

The second approach to analyzing the recorded signal is to pass it through a bank of bandpass filters covering the frequencies of possible pitches and watch their outputs when the desired note goes through. The filters serve to separate the different frequency components of the music, and by comparing their responses at the proper time we can isolate and measure the pitch of our particular note.

Consider the output of the particular filter centered at frequency f_0. If the impulse response of this filter is $h_{f_0}(t)$ then its output $s_{f_0}(t)$ in response to the waveform $x(t)$ is

$$s_{f_0}(t) = \int_{-\infty}^{+\infty} x(\tau) h_{f_0}(t - \tau) \, d\tau. \tag{1.4}$$

The transfer function $H_{f_0}(f)$ of the filter is the Fourier transform of its impulse response $h_{f_0}(t)$. If the passband response of the filter is symmetrical about the frequency f_0, that is, if $H_{f_0}(f_0 + f) = H^*_{f_0}(f_0 - f)$ for $|f| < f_0$, then we can write $H_{f_0}(f) = \frac{1}{2} H_l(f - f_0) + \frac{1}{2} H_l(f + f_0)$, where $H_l(f)$ is the transfer function of the *equivalent low-pass* filter. $H_l(f)$ is obtained by shifting the (positive frequency) passband of $H_{f_0}(f)$ down to zero frequency. In this case, the impulse response of the bandpass filter centered at f_0 can be expressed in terms of that of the equivalent low-pass filter as $h_{f_0}(t) = h_l(t) \cos(2\pi f_0 t)$.

Now comes the interesting part. Let us choose a weighting function as in Equation (1.2) that is just the time inverse of the equivalent low-pass filter impulse response here; that is, let

$$w(t) = h_l(-t). \tag{1.5}$$

If we write

$$S(t, f_0) = \int_{-\infty}^{+\infty} x(\tau + t) w(\tau) e^{-j 2\pi f_0 \tau}\, d\tau = A_w(t, f_0) e^{j \phi_w(t, f_0)} \tag{1.6}$$

from Equation (1.2), evaluating the windowed spectrum at a fixed frequency f_0 and viewing it as a function of time t, then we can show:*

$$s_{f_0}(t) = \int_{-\infty}^{+\infty} x(\tau) h_l(t - \tau) \cos[2\pi f_0 (t - \tau)]\, d\tau = A_w(t, f_0) \cos[\phi_w(t, f_0)]. \tag{1.7}$$

That is, the envelope of the filter output A_w is the same as the magnitude of the windowed spectrum, and the instantaneous phase of the filter output ϕ_w is the same as the phase of the windowed spectrum, when evaluated at the center frequency of the filter f_0 and the same time t. Note that the phase is a rapidly-increasing function of time. We will often express it as $\phi_w(t, f_0) = 2\pi f_0 t + \phi_{rel}(t, f_0)$, where $2\pi f_0 t$ is the advance in phase with time at the center frequency of the particular filter, and the remaining term $\phi_{rel}(t, f_0)$ is the *relative phase,* a slowly-varying function of time. The relative phase shows signal phase behavior with respect to an oscillator running at frequency f_0.

The windowed spectrum and bank of filters approaches are equivalent. They are equivalent in the formal sense that for any given weighting function $w(t)$ there is a corresponding set of bandpass filters whose impulse responses are related to the weighting function. Both approaches have useful features. When we process sampled data using the discrete Fourier transform as described in Sec. 2.5 we are calculating the digital equivalent of the windowed spectrum. Yet it is convenient to regard the spectral values so calculated as the outputs of some fictitious bank of filters. We will refer to these "synthesized filters" as if they were the actual source of our information.†

The two approaches are also equivalent in the more general sense that they both provide a means of mapping the one-dimensional waveform $x(t)$ (or the one-dimensional spectrum $X(f)$) onto a two-dimensional space, the frequency-time or *f-t* plane. They allow us to assign to each signal element a time of occurrence and a frequency. A common procedure when analyzing a signal is to

---

* To prove this, separate Eq. (1.6) into its real and imaginary parts, and compare to Eq. (1.7) evaluated after substituting Eq. (1.5). The symmetrical-passband case is particularly simple, but a similar correspondence exists between more general filter shapes and other windows. See *Papoulis* [1962, Sec. 8-1] for further discussion.

† I presented the windowed transform approach first since it is the one we will actually use when analyzing data. Historically, the bank of filters method was used first for whistler-mode signals (in *Potter's* [1951] sound spectrograph). In fact, it is only in the last two decades, since the invention of the FFT algorithm [*Cooley and Tukey,* 1965], that it has become practical to calculate windowed transforms in quantity. And only for the last ten years have the best windows been available [*Harris,* 1978; *Nuttall,* 1981]. For a view of spectrum analysis at the beginning of the modern era the reader may wish to see *Bingham et al.* [1967], *Welch* [1967], and the introduction to *Childers* [1978] (which contains reprints of the preceeding two papers). Welch's "modified periodogram" is, of course, the magnitude of the windowed transform.

plot the signal over a lattice of points in the *f-t* plane as various shades of gray or various colors, where density or color at each point is proportional to the magnitude $|S(t,f)|$ at that time and frequency. The resulting plot is called an *f-t spectrogram* (see Figure 2.2 for an example).

This mapping procedure is the essence of spectrum analysis. In effect, spectrum analysis allows us to reconstruct from the recorded waveform the two-dimensional score which specified the times and pitches of the notes in the music. The analogy between *f-t* spectrograms and music scores extends even to a similarity in layout, with frequency or pitch increasing toward the top and time increasing to the right in the usual presentation. Here the analogy ends. A score, of course, is merely a recipe for the performance of music, which is the desired product. Spectrum analysis works in the other direction, starting with a signal and generating a spectrogram in an effort to understand the mechanisms that created the signal. It is as if music were incidental to a study of the process of composition.

There is another major difference between a music score and a spectrogram. While there is usually only one authentic score (in many copies) for a given piece of music, there are many different ways of mapping a given signal into a frequency-time spectrogram. By changing the duration of the weighting function, or the bandwidths of the corresponding bandpass filters, we will produce a different spectrogram. Short weighting functions/wide filters are useful when analyzing brief or quickly-changing signal components where time resolution is more important than frequency resolution. Conversely, long weighting functions/narrow filters provide maximum frequency discrimination at the cost of lower resolution in time. When analyzing data in later chapters we will often try various combinations of time *vs.* frequency resolution to tailor the analysis to a particular signal.

1.3 Benefits of Phase Information in Spectrum Analysis

The *f-t* spectrogram described above shows the magnitude of the signal $S(t,f)$, but not its phase. Most whistler-mode research has used the *f-t* spectrogram to study signal magnitudes as functions of frequency and time. However, VLF signal phases have received scant attention. There are two reasons for this. First, many naturally-occurring signals such as whistlers and chorus have a very complicated structure and it is not clear that their phase information would be easy to interpret. Second, until recently there have not been any machines capable of making relevant signal phase measurements. Signals require special recording techniques in the field to preserve phase information, and special processing in the laboratory to extract and display it.

Yet signal phases contain information independent of signal magnitudes. If the signals are simple enough, such as phase-coherent signals from VLF transmitters, their phases may be easy to interpret and very interesting, as we will see in Chapters 3 and 4. The uses of phase information fall into three broad catagories as follows:

1. *Phase information can help identify and classify signals.* The phase of a signal from a VLF transmitter reveals its modulation format, and may even say something about the complexity of its keying equipment. By comparing the phases of constant-frequency signals in a recording we can sort out those that are mutually phase-coherent, such as harmonics from local power lines, from those that are not, such as magnetospheric line emissions.

2. *The phase of a signal may be an interesting parameter in itself.* Phase changes in sub-ionospheric signals may be correlated with whistler-mode waves (Trimpi events), indicating electron precipitation into the ionosphere from wave-particle interactions. Phase changes in coherent whistler-mode signals can be used to measure path length changes (duct drift) caused by large-scale motions of

the magnetosphere. And the phase changes that occur in whistler-mode signals during wave-particle interactions may help to validate theoretical models of the interaction.

3. *Phase information allows us to measure instantaneous frequency.* The *instantaneous frequency* of a signal is defined as

$$\tilde{f} = \frac{1}{2\pi} \frac{\partial \phi(t,f)}{\partial t} \tag{1.8}$$

where $\phi(t,f) = \arg\{S(t,f)\}$ is the phase of the signal at time t and frequency f. The instantaneous frequency clearly corresponds to the common-sense notion of frequency in signals where $\tilde{f}$ varies slowly with time. *Ackroyd* [1970] shows that it provides a measure of the frequency at which the power of the signal acts at a given time even for signals where $\tilde{f}$ varies rapidly. In practice, we will approximate the instantaneous frequency by the finite difference

$$\tilde{f} \approx \frac{1}{2\pi} \frac{\phi(t+\Delta t, f_0) - \phi(t, f_0)}{\Delta t} \tag{1.9}$$

where $\phi(t,f_0)$ is the phase of the signal as passed by an analysis filter centered at frequency f_0. If we can make Δt sufficiently small we can measure the frequency of a signal very rapidly. The minimum length of Δt is limited by the signal/noise ratio. If there is a lot of noise present, a measurement made with a small Δt will be inaccurate; we must use a longer interval to average out some of the noise. As an example, with a typical VLF signal we might need an interval of 10 ms in order to make a frequency measurement accurate to 10 Hz.

"Wait a minute," I hear you say. Isn't this a violation of the uncertainty principle that $\Delta t \cdot \Delta f \approx 1$? That is, shouldn't it take about 100 ms to measure the frequency of a signal to within 10 Hz? The answer is no, *as long as there is only one signal component present.* It is true that if we were trying to separate two signals spaced Δf apart, a filter which could discriminate between them would have to have a response time no less than approximately $1/\Delta f$. (As engineers know, in a practical filter with 10–90% risetime τ and 3-dB bandwidth ν, the product $\tau\nu$ is usually from 0.30 to 0.35.) However, separating two signals is a different problem from trying to measure the frequency of one, providing we know there are no interfering components at nearby frequencies. To take a related problem, suppose we know that a given function is a sinusoid of the form $A\sin(2\pi ft + \phi)$. We need its value at only three closely spaced values of t, not necessarily over a full cycle, to detect its curvature and determine the three unknown constants A, f, and ϕ. In the case of our signal, the magnitude $|S(t,f_0)|$ (corresponding to A) is known separately, and only two measurements of $\phi(t,f_0)$ are needed to determine $\tilde{f}$.

Instantaneous frequency measurements provide a lot of information about VLF signals that cannot be obtained from *f-t* spectrograms. For instance, when whistler-mode signals from the Siple transmitter show growth due to wave-particle interactions, the output signal from the magnetosphere is usually higher in frequency, say by 1 to 5 Hz, than the transmitted input signal [*Paschal and Helliwell,* 1984]. This offset sometimes appears almost at the beginning of a received pulse. When a triggered emission occurs at the end of a pulse, the emission always starts at a frequency well above the input frequency, say 50 Hz higher. What is more, the transition from the frequency of the growing input pulse to that of the emission seems to take place almost instantaneously, perhaps in less than 5 ms in some cases. These sorts of measurements cannot be made without phase information.

Dowden's Analog "Phasogram" Technique. The earliest attempt I know of to explicitly examine the phase information in magnetospheric signals is that of *Dowden et al.* [1978]. They recorded a phase reference tone along with whistler-mode signals from a 6.6 kHz VLF transmitter using an

analog recorder with FM modulation. In the laboratory, a frequency synthesizer was phase-locked to the reference tone and its output (which could be tuned in frequency) was compared in phase to the whistler-mode signal. The result was a "phasogram," plotting relative phase (1 revolution full-scale) versus time. They were able to measure the positive frequency offset of a growing whistler-mode signal, as well as other effects such as the Doppler shift in frequency due to duct motion [*Rietveld et al.*, 1978]. *Rietveld* [1980] has since used the phasogram technique to measure the frequency of whistler precursors.

In some ways the system I will describe parallels Dowden's phasogram system, though they were developed independently. However, there are many differences. One is that we use digital rather than analog signal processing. This adds flexibility to the analysis and presentation of the data. Also, while Dowden uses his reference tone to correct for timing errors and reconstruct signal phases, he cannot unshift the spectrum to correct for rate (speed) errors. We do, and are able to use narrower analysis filters as a result.

1.4 History of VLF Spectrum Analysis at Stanford

The pace of scientific discovery in many fields has historically been driven (or limited) by the invention of machines that allow men to observe nature in new ways. Even in such an esoteric field as VLF research and over such a short history as that of the past thirty years, our ability to recognize and understand new phenomena has been determined by the observational tools in existence at any given time. In order to appreciate the features and limitations of the current analysis system, we need to see it in its proper context. This section describes the various spectrum analyzers that have been used at Stanford. (I am glad to be able to preserve some of the details of these instruments before they are completely forgotten.)

Table 1.1 lists the characteristics of spectrum analyzers used by the VLF Group at Stanford University during the thirty years up to 1986. There are a few other units that have also been used, particularly at field stations, but these six have done almost all of the laboratory data analysis. They are listed in chronological order. The UA-6B/H, SD350-6, and Paschal system (the subject of this report) are currently still in use.

Looking at this table it may be hard to tell the direction of the arrow of progress. There is no steady movement to increased input bandwidths or faster analysis, for instance, though there is some change in that direction. Nor (though it isn't shown in the table) have spectrum analyzers become simpler or easier to use. Instead, these machines have evolved toward extracting more information from the signal, through higher-quality analysis (such as greater uniformity in the filter responses), better presentation (such as computer plots), and additional features (such as using phase information).

Sona-Graph. Early VLF recordings made by Stanford in the mid 1950's were analyzed with a Kay Electric Company (now Kay Elemetrics Corp.) Sona-Graph spectrum analyzer. This sound spectrograph was originally invented for the analysis of speech; it was first applied to the study of whistlers by *Potter* [1951]. The Sona-Graph has a single bandpass filter which is scanned through the input signal bandwidth as follows: A short interval (2.4 s) of signal is recorded on a magnetic drum. Connected to the drum is a cylinder wrapped with a piece of specially-treated paper. During analysis, the drum rotates and the signal is played back through the filter. The detected output of the filter controls the darkness of a line being drawn on the paper by a pen. With each rotation of the drum, the whole 2.4 s interval of signal is analyzed. After each rotation the frequency of the filter is increased, the pen is moved along the axis of the cylinder, and another trace is drawn representing signals at an adjacent frequency. After many revolutions, the paper is unwrapped to

TABLE 1.1

Spectrum Analyzer Characteristics

	Sona-Graph	Rayspan	UA-6B/H	Stiles	SD350-6	Paschal
Input BW[1]	8 kHz	10.5 kHz	.01–40 kHz	2.1 kHz	.01–300 kHz	10.6 kHz
Filter BW	45 Hz	32 Hz	30–120 Hz[4]	33, 65 Hz	19–600 Hz[4]	20–640 Hz
Speed[2]	1 : 125	real-time	real-time	$< 1 : 15$[5]	real-time[6]	$< 1 : 10$
Duration[3]	2.4 s	∞	∞	minutes	∞	400 s
Output	thermofax paper	35 mm film	35 mm film	line plots	35 mm film[7]	dot-matrix plots
Phase Info?	no	no	no	no	no	yes
Technique	single filter	bank of filters	time compression	computer FFT	special FFT	computer FFT

Notes:

1. The Rayspan, UA-6B/H, and Stiles system allow the input signal to be translated in frequency. The SD350-6 has a zoom function which has a similar effect. The Sona-Graph and Paschal system analyze baseband signals only. Effective data bandwidths can also be changed by playing back tape-recorded data at different speeds.
2. Speed is the processing time for a unit time of signal. Real-time means a ratio of 1 : 1 or greater.
3. Duration is the maximum continuous interval of signal which can be processed.
4. Filter bandwidths are listed for a 10 kHz input bandwidth. Other filters are available with different input ranges.
5. FFT calculation only. Output plotting is very slow.
6. Real-time for input bandwidths to 50 kHz.
7. Digital output of spectrum magnitudes also available.

reveal a "Sonagram," an *f-t* spectrogram which displays the magnitudes of signal components at different frequencies and times through varying shades of gray.

The Sona-Graph was the first instrument to make a spectrogram. It was a major improvement over the swept-frequency analyzers (such as the Panoramic analyzer) that were also used about this time. A swept-frequency analyzer generates a graph of signal magnitude versus frequency by slowly tuning a bandpass filter through the frequency interval of interest. If the signal is not stationary the swept-frequency analyzer may fail to observe some signal feature by being at the wrong frequency at a given time. By playing the signal over and over again while the filter is swept slowly, the Sona-Graph is sure to catch all signal features, no matter at what time or frequency they occur.

However, the Sona-Graph suffers from two major limitations. First, it can only analyze a short segment of data at one time. Second, and more important, it is very slow. On each rotation of the drum the signal is played back and analyzed at one frequency. To generate a complete spectrogram requires about five minutes [*Helliwell*, 1965, p. 88].

Despite its limitations, the Sona-Graph has remained a popular spectrum analyzer. In 1984 Kay came out with their model 7800 Digital Sona-Graph, a solid-state machine that stores the signal digitally and has a number of features not found in the original machine. It still produces "Sonagrams" with the same thermofax paper on a rotating drum.

Rayspan. The slow speed of the Sona-Graph was overcome in the Raytheon Rayspan, introduced about 1961. The Rayspan has a bank of 420 magnetostrictive rod (mechanical) bandpass filters

spaced every 25 Hz from 167,000 to 177,475 Hz, each with a bandwidth of 32 Hz. The input signal is translated in frequency (by mixing with a 167 kHz local oscillator, say, for 0 to 10.5 kHz analysis) and applied to all the filters in parallel. The outputs of the filters are then sampled in sequence by a rotating capacitive commutator and detected. The commutator can make up to 150 scans of the complete filter bank each second. The output of the Rayspan is recorded on 35 mm photographic paper or film by varying the brightness of a trace moving across an oscilloscope with each scan (in the frequency direction) while the paper is slowly pulled through the camera (in the time direction). The resulting spectrogram is very similar to that produced by the Sona-Graph, except that the scan lines run in frequency for each moment in time, rather than in time at each increment in frequency. The Rayspan operates at real-time speed, analyzing one second's worth of data in one second. It is also capable of analyzing data segments of unlimited length, rather than the few seconds at a time processed by the Sona-Graph.

The Rayspan greatly increased the rate of data analysis. Whereas a trained operator in one working day might generate spectrograms for about one minute's worth of data using the Sona-Graph, now he could process several hours of field recordings. Also, the output of the Rayspan could be displayed momentarily in spectrogram form on a long-persistance oscilloscope without having to be filmed. For the first time interesting events suitable for detailed study could be identified from field tapes by their spectral characteristics as analyzed by machine rather than just by ear.

Still, the Rayspan has its shortcomings. The most serious is the lack of uniformity between the responses of the 420 individual filters. Some are more sensitive than others; the manufacturer's specifications permit a variation of ± 3 dB. This can create horizontal lines in the spectrogram indistinguishable from true constant-frequency signal components. In fact, magnetospheric signals with line structure were not routinely seen until the Rayspan was replaced because lines in the spectrogram were taken to be artifacts of the analyzer rather than real signals.

Another shortcoming is the inability to change the bandwidth of the analysis filters to match the signals being analyzed. By changing the playback speed of the tape recorder, the effective bandwidth of the analysis filters can be changed, but only by a limited factor. (This technique can be used with all spectrum analyzers, of course.)

Ubiquitous UA-6B/H. The next spectrum analyzer was the Federal Scientific Corporation (now Nicolet Scientific) Ubiquitous UA-6B/H, acquired in 1972 and still in use. This is a so-called "time compression" spectrum analyzer. A digital delay line is used to make a recirculating buffer that can hold about 1500 samples. The input signal is sampled and digitized. As each sample is taken it is entered into the buffer, replacing the oldest sample stored there so the buffer always contains the latest 1500 samples. At the same time that new samples are being taken in, the buffer is scanned at a much higher rate and the samples read out are converted back into an analog signal. This output signal is just a reproduction of the input signal stored in the buffer, but since it is read out faster it is compressed in time. All components in the input signal are multiplied in frequency by the compression factor. An input signal from 0 to 10 kHz will be reproduced from the buffer as a signal spanning 0 to 5 MHz, a speed-up of a factor of 500. The sped-up signal is then analyzed by a sweeping filter. The bandwidth of the sweeping filter is wider by the compression factor and its response time correspondingly short, with the result that a single sweeping filter can now provide complete real-time coverage. In fact, the UA-6B/H uses a bank of 5 sweeping filters positioned at adjacent frequencies for a further increase in analysis speed, and a complete scan of 500 spectral lines is output every 10 ms. The bandwidth of each filter is about 1.6 times the filter spacing, so 10 kHz data will be analyzed by 500 filters spaced every 20 Hz, each with a bandwidth of 32 Hz.

The innovation introduced with the UA-6B/H analyzer is the ability to change the overall analysis bandwidth, and thus the bandwidth of each of the 500 effective spectral filters. This is done by changing the input sampling rate but not the buffer readout rate. For example, by sampling the (appropriately filtered) input signal at half of the rate for 10 kHz analysis, we will store only those components lying between 0 and 5 kHz; but we still generate a 500-line spectrum every 10 ms, whose spectral filters are now only 16 Hz wide. By sampling at even lower rates we can analyze with even narrower filters. (Of course, as spectral filters become narrower there is less independence between one spectrum and the following one 10 ms later. The spectral data are becoming increasingly smooth and redundant.) The UA-6B/H is used in conjunction with a frequency translator so a total analysis bandwidth of 1, 2, 5, or 10 kHz can be centered about any arbitrary input frequency.

The UA-6B/H is currently the most popular analyzer at Stanford, primarily because of its simple and straightforward operation. However, it still suffers a bit from filter non-uniformity. It is difficult to keep the gains of the 5 sweeping filters exactly equal, though they are much more uniform than the 420 filters of the Rayspan.

Stiles' Digital Analysis. G. Stiles [1974] was the first person at Stanford to make extensive use of a general-purpose computer to analyze VLF signals. In his work, a 2.1 kHz segment of input data was translated in frequency, filtered, digitized, and stored on magnetic tape. Then the signal was analyzed using digital signal processing techniques on a mainframe computer, an XDS Sigma-5. Some of the programs used, including the FFT procedure, were previously described by *Cousins* [1971]. (The FFT or Fast Fourier Transform procedure is an algorithm for calculating the discrete Fourier transform, the counterpart for sampled signals of the integral Fourier transform of Eq. (1.1). We will return to this in Section 2.5.)

Stiles was primarily interested in the small-scale structure of VLF emissions triggered by pulses from VLF transmitters. In particular, he was interested in what happens at the very beginning of an emission as it starts to separate from the triggering signal. The question is whether an emission starts at the transmitted frequency and then drifts up, or does it begin already above the input frequency. His conclusion was that the emission starts at the transmitted frequency, though, as he noted, any model in which the emission starts within about 50 Hz of the input signal would agree fairly well with his results. Stiles was not able to use the phase information in his signals, and this limited to some extent his ability to resolve the fine structure of emissions. Emissions are a subject of continuing interest, as we will see in Chapter 4.

Stiles' work is important in several ways beyond his direct contributions to magnetospheric physics. He showed that VLF spectrum analysis could be done on a computer as well as on special-purpose spectrum analyzers. He demonstrated the power of the computer to examine small details of signal structure. One of the benefits of computer processing here is the variety of plotting routines that can be devised to display spectral information. It is much easier to show the detail in a small region of the *f-t* plane with computer graphics than it is with the more restricted 35 mm film format of the analog analyzers.

SD350-6. The Spectral Dynamics (a subsidiary of Scientific Atlanta) SD350-6 spectrum analyzer arrived in 1980. This is a general-purpose unit that uses digital signal processing techniques internally, but still has analog input and output ports. The input signal is digitized and stored in memory. Special-purpose hardware is used to calculate the FFT of overlapping input segments. The magnitudes of filtered spectral components are then converted back to analog form for external display or filming. The outputs of the synthesized spectral filters can also be averaged (magnitudes only) before being displayed, if desired. Though it is presumably present internally, no signal phase

information is available at the output.

The bandwidth of the signal to be analyzed can be varied, as with the UA-6B/H, by changing the input sampling rate. For a fixed number of synthesized filters, this changes their bandwidths while the spectrum output rate remains constant. SD350-6 input bandwidths range from 10 Hz to 300 kHz. However, a new dimension of control has been added with the SD350-6. Not only can the input sampling rate, and thus the analyzed bandwidth, be changed, but the FFT transform size, and thus the number of filters synthesized over that bandwidth, can also be selected. For instance, a 10 kHz input bandwidth can be analyzed with 25 filters spaced every 400 Hz (transform size = 64, one output spectrum generated every 300 μs), or with 50, 100, ..., on up to 800 filters spaced every 12.5 Hz (transform size = 2048, one output every 14 ms).

The SD350-6 also has a "zoom" function which allows the analyzer to concentrate on a limited portion of the input bandwidth. Zoom magnifications range in powers of two from 2 to 128. For example, with an input bandwidth of 10 kHz, using a magnification of 8 will analyze a 1.25 kHz section of the input signal. This section can be centered anywhere in the 10 kHz range of the input. With a transform size of 2048, the 800 filters synthesized will now be placed every 12.5/8 or 1.5625 Hz. The zoom feature is similar in effect to frequency translation of the input signal as used in the previous three analyzers, though it is a bit harder to adjust.

The SD350-6 is a complicated machine and represents the present peak of stand-alone analyzer development. Despite its complexity, it is being used more frequently as its virtues become better known. Recently its digital output was hooked up to a small computer in the analysis lab. Output spectra can now be recorded on digital tape or sent to other computers around campus for further processing and plotting. This is starting to supplement the standard use of 35 mm filmed spectra.

Paschal's Digital Analysis System. As we gather from the previous discussion, the system that is the topic of this report is merely the latest in a long line of spectrum analyzers. This system uses digital signal processing techniques performed on a computer. To this extent it is a descendant of Stiles's system. However, the hardware is dedicated especially to VLF analysis, and represents a step away from the large central systems of the past towards the smaller individual computers of the present. Chapter 2 describes this system in more detail.

Table 1.1 shows that this digital analysis system is not as wide-band or as fast as some of the other systems. It is not a general-purpose spectrum analyzer like, say, the SD350-6. Its main innovation is to make signal phase information available for use. In this it represents another stage in the evolution of techniques to extract more and more information from the signal. It is not, however, the final answer. Chapter 5 has some suggestions for future analysis systems.

1.5 Digital versus Analog Analysis

Finally, I want to say something about the differences between the analog and digital analyzers mentioned above as we use them for the study of VLF signals. These differences are not really about the way signals are processed, since any analog filter can be mimicked arbitrarily closely by a corresponding digital filter, and *vice versa*. Rather, the important differences lie in the presentation of spectral data.

All of the recent analog analyzers above (Rayspan, UA-6B/H, and SD350-6 using its analog output) make spectrograms on 35 mm film or paper. This medium has both some advantages and some severe limitations. The visual examination of filmed (analog) spectrograms is relatively fast and inexpensive, and is the best method for quickly identifying the gross features of the recorded VLF signals. However, converting these features into quantitative data by scaling the film on an optical digitizer is a time-consuming and labor-intensive task. Even more serious is the fact that many important spectral properties cannot be readily extracted from the image. These are as follows:

1. *Magnitude Information.* It is very difficult to extract quantitative magnitude information from photographed spectra. Magnitude information is present as variations in optical density, and there is no easy way to scale the density to determine absolute (or even relative) signal magnitudes. Most of our analysis is limited to saying "yes, there is a signal present at this frequency and time," or "no, there is no signal present." Displays that show magnitude directly (such as the A-scan format, which produces a series of line graphs of magnitude versus frequency at successive times) are expensive in terms of the amount of film used, and are awkward to generate.

2. *Dynamic Range.* The dynamic range of the analyzer-film system is only about 20 dB. Within this range, signals show variations in density in the film proportional to signal magnitude. Outside this range, stronger signals appear uniformly black and weaker signals are not visible. Also, if there is a strong signal present it may be very difficult to detect a weak signal (such as the leading edge of an emission) which is near it in frequency.

3. *Precise Time and Frequency.* Even if a signal stands out strongly in the filmed spectrogram, and is easily identified, it is still difficult to measure accurately its frequency and time of occurrence. Our present accuracy is about 30 Hz over a 10 kHz bandwidth, and 30 ms over a time interval of a few seconds. This is due to difficulties in measuring and compensating for various errors in the recording and analysis process. Fundamental problems are variations in tape speed during recording and playback, variations in camera film speed when photographing the spectra, and limitations in the bandwidth and spectral output rate of the analog analyzers. At present we cannot use the recorded pilot tone to compensate for tape speed variations when using an analog analyzer, though it is conceivable that such a system could be built.

The digital analysis system overcomes some of the above limitations and is a powerful tool for the analyzing certain types of VLF data. Some measurements, such as phase and instantaneous frequency, can only be made using digital techniques. Some types of output, such as A-scan plots, can be made more easily with the digital system. However, analog analysis also has distinct advantages, and the two methods are best used when their capabilities complement each other.

For routine data surveys and for studies where sufficient information can be gleaned from filmed *f-t* spectrograms, analog analysis wins hands down. The advantage of analog analysis here comes from two factors. First, our analog machines are capable of analyzing data in real time; that is, the spectrum appears almost immediately after the tape is played into the system. The digital analysis

system requires a significant amount of time just to digitize the data, and the subsequent analysis proceeds at much less than real-time speed. Generating an *f-t* spectrogram on the digital system typically takes about 10 seconds of processing for each second of data, and phase analysis is even slower. Second, the analog analyzers are capable of processing larger amounts of data at one time. For data surveys where the operator sits at the screen and watches for interesting signals, our analog analysis systems can process data segments whose duration is limited only by operator endurance. Even if the results are to be filmed, data may be processed an hour or so at a time. The digital analysis system limits data segments to a length that will fit on the computer disk (say 30 seconds of 10.6 kHz data) or on one digital tape (about 7 minutes of data).

Because of these differences, digital analysis is currently only useful for short, well-defined data segments. The first step in using the digital analysis system is to locate an interesting data segment in a field recording, either by monitoring the tape on an analog analyzer or by examining existing analog spectrograms. Only then is it profitable to take the field tape, digitize, and analyze it.

The types of signals that can be usefully examined with the digital system are also limited. Generally speaking, the simpler the signal, the more the information that can be uncovered by digital analysis. For phase analysis, relatively constant-frequency signals are needed. When very narrow-band signal structure is to be analyzed, even the narrowest filters available in the digital system may be too wide, and frequency translation followed by analog analysis may be the better choice. The analysis of more complicated signals such as chorus and whistlers is beyond the capabilities of the current digital system.

1.6 Outline of the Thesis

General Philosophy. This is primarily an observational, one might almost say experimental, thesis. Part of the work describes the methods of observation; the remainder uses these methods, primarily phase analysis, to study in turn a variety of VLF phenomena. Some phenomena reveal new and unexpected features while others are merely viewed in greater or more accurate detail than before. When there is an existing body of relevant theoretical work, it is usually discussed at the end of each section of observations. However, while discrepencies between the observations and the predictions of current theories are noted where they occur, the development of new theories is in most cases beyond the scope of this thesis.

Chapter 2 describes the various algorithms used in the digital analysis system, with an emphasis on those pertaining to phase measurement. Particularly important are the techniques that have been developed to measure and correct for analog tape timing errors (wow and flutter). I hope to give enough detail that the reader who sits down to develop the next-generation spectrum analyzer will have a head start.

Chapter 2 also describes the instruments that are used at field stations to record data so that the reader may have some understanding of the data's limitations. The computer hardware used in the digital analysis system to run the programs and plot the results is briefly described. However, I do not go into detail here. Computer hardware has changed so rapidly in the past decade that almost all of the analysis equipment described is already obsolete.

Nor do I give any software listings, not out of secrecy but because the analysis programs are not usable on other systems. In order to attain reasonable program efficiency, in memory use as well as in processing time, much of the software had to be written in Data General Nova/Eclipse assembly language, not exactly the *lingua franca* of the computer world. Much of the FFT algorithm is written in Eclipse S/230 microcode, an even more restricted dialect. I am sorry about this, as the

major investment in a computer system is the effort needed to write the software, and it would have been nice to pass on some of my work. We can only hope the situation improves in the future as more powerful processors enable more programs to be written in high-level languages. (Computer manufacturers have been forecasting this for years.)

Chapter 3 analyzes several examples of signals without growth. These are signals that move from source to receiver unchanged in form by the medium through which they travel, except possibly for some general attenuation and time delay. With some we can use their phase characteristics to classify them, such as the various modulations of naval VLF transmitters. With others, phase is a means of monitoring their path of propagation. We will see phase changes that occur during Trimpi events, perturbations in sub-ionospheric paths due to magnetospheric particle precipitation. And we will see changes in phase as whistler-mode ducts stretch and compress due to large-scale motions of the magnetosphere.

Chapter 4 is concerned with growing signals, whistler-mode signals changing because of wave-particle interactions. Here we encounter a wide variety of phenomena. We will look at some old events from a new perspective, such as the growth of transmitted pulses (particularly fruitful signals) and the generation of sidebands. We will see some events in greater detail than before, such as the behavior of emissions at the moment of separation. And we will see some completely new effects, such as the phase-locking that occurs during entrainment. Chapter 4 includes references to theoretical models that make predictions relating to the observations.

Chapter 5 summarizes the results and presents a few suggestions for future work. Improvements should be made in the equipment used at field stations to record VLF signals. A new generation of analysis systems needs to be developed with smaller, individual workstations, perhaps based on personal computers. Analysis algorithms need improvement to give better measurement and correction of analog tape errors, as well as providing a greater range of analysis filter bandwidths and output formats. Finally, some new and exciting observations, such as the correlation between whistler-mode phase delay and transient magnetic field disturbances, are very tedious to reduce from broadband data and warrant the use of dedicated hardware in the field.

The three appendices present some incidental results. *Appendix A* gives a mathematical derivation not found in textbooks, the discrete Fourier transform of a sinewave with arbitrary frequency *and* phase. This is used in Chapter 2 when discussing the phase errors in the analysis procedure. *Appendix B* extends the group delay calculations of *Park* [1972] and finds the ratio of phase delay to group delay over a whistler-mode path. This turns out to be a simple function of the ratio of signal frequency to equatorial gyrofrequency, at least inside the plasmapause. *Appendix C* gives a procedure using the chirp z-transform algorithm to perform data resampling for the correction of tape timing errors. This procedure was too time-consuming to use in the current system but may be useful in some future design.

1.7 Contributions of the Present Work

1. A technique has been developed to correct analog tape recordings for errors due to wow and flutter, for the first time enabling signal phase information to be extracted on a routine basis. During spectrum analysis, the phase of a recorded pilot tone provides a time reference, and its frequency gives the tape speed error. We correct for frequency shifts caused by the speed error, and correct the phase at the output of each analysis filter for timing errors. (All tapes recorded by Stanford since 1973 have included a phase-reference pilot tone in preparation for this technique.) Knowing the phase as well as the magnitude of signal components effectively doubles the available information. Phase enables the measurement of instantaneous signal frequency, and facilitates the study of coherent and almost-coherent signals.

2. Non-growing whistler-mode signals from the Siple Station transmitter are found to show changes in phase delay which are correlated with the earth's magnetic field at the ends of the duct. Previous studies found similar correlations between whistler-mode Doppler frequency shifts and continuous micropulsations. Phase is found to give a less-noisy measure of path motion than Doppler shift, with much better time resolution. For the first time it is possible to study relatively rapid events, and we find new evidence for transient micropulsations which move from equator to ground as Alfvén waves.

3. The relative phase of a growing whistler-mode signal is found to advance with time, often parabolically. That is, the growing signal at the output of the interaction region exhibits a positive frequency offset from the input signal; this offset generally increases with time. Such behavior places an important new constraint on the predictions of theoretical models of wave-particle interactions. Occasionally the instantaneous frequency is offset even at the beginning of a pulse, an effect which is not explained by any current model.

4. There are as yet no completely satisfactory models for emission triggering. Phase analysis now offers some clues for future models. When pre-termination triggering occurs, it is always after a phase advance of 1.5–3 revolutions. When triggering occurs at the end of a pulse, the frequency can change from the growing pulse to the emission 100 Hz higher very rapidly, sometimes in a few milliseconds or less. When growth resumes after a pre-termination emission, the signal phase restarts at its initial value.

5. Phase analysis shows that the mechanism which causes a two-tone transmitted signal to develop sidebands on a whistler-mode path is the mechanism of growth, supporting the model of *Helliwell et al.* [1986a]. Each beat in a two-tone signal behaves like a growing pulse, showing a phase advance accompanying an increase in magnitude. Yet successive beats remain phase-locked, resulting in discrete sidebands at harmonics of the two-tone separation.

6. It is known that the components in a two-tone signal suppress each other's growth when their frequencies are separated by 10 to 50 Hz. Phase analysis now shows that the suppressed components are slightly advanced (≈ 0.15 rev) over their input phases, evidence for linear amplification. When the amplitude of one component is reduced by 20 dB while the other remains constant, the output phases show an additional advance as suppression weakens, but the components still remain coherent.

7. Falling emissions are generated when the interaction region moves upstream (down-wave) of the equator. When a faller is entrained by a constant frequency pulse, the phase behavior in this off-equatorial interaction is seen to be similar to that during equatorial growth. Entrained signals show a phase advance accompanying an increase in amplitude, and a brief phase wrap-up at the end of the entraining pulse.

8. Whistler precursors are seen for the first time on a transmitted signal. Amplitude, phase,

and sideband behavior indicates the precursors are due to a momentary increase in growth activity. They cannot be explained merely by the presence of a triggering signal, as proposed in some models of natural precursors.

9. Magnetospheric signals with line structure have been observed for some time, often with components spaced roughly 60 Hz apart. They have been thought to be caused by radiation of harmonics from power lines. Phase analysis now shows that magnetospheric lines, at least in the cases studied, are not spaced by 60 Hz, are incoherent, and have no relation to power line components. A model is developed to explain them based on magnetospheric processes only—growth, suppression, echoing, and multipath coupling.

2. METHOD

2.1 Field Station Recording Equipment

In this section I briefly describe the equipment used to receive and record signals at a typical field station. While it is not necessary to understand this equipment in great detail, a basic knowledge will help explain some of the characteristics and limitations of the recorded data. Figure 2.1 shows a simplified block diagram of a typical VLF receiving system. The various items in the system are as follows:

Loop Antenna and VLF Receiver. The VLF receiver uses a loop antenna to receive very-low-frequency radio waves. It generates an output voltage proportional to the magnetic field B of the incident wave. The frequency response of the receiver is flat from 300 Hz to about 50 kHz. That is, over this frequency range the output voltage depends only on the strength of an incident signal component, and not on its frequency. (Note that the voltage induced in the loop is proportional to dB/dt, and rises with frequency. We overcome this effect by designing the receiver so that the input circuit impedance is dominated by the loop inductance L. Since the loop reactance $j\omega L$ also rises with frequency, the input current to the receiver is constant with frequency.) A flat frequency response gives the receiver the greatest dynamic range since the peak amplitudes of signal components at the low-frequency end (like whistlers) are about the same as those of components at the high end (like VLF communications transmitters).

Frequency Standard. The frequency standard at most stations is an ovenized quartz oscillator with a good short-term frequency stability of 1×10^{-10} or better. However, depending on the time elapsed since the oscillator was last calibrated, the output frequency, though stable, may be offset from its desired value by as much as 1×10^{-8}. The frequency of the output signal from the standard is divided down by the clock to provide the station local time reference, against which signals are compared when making phase measurements. An error in frequency of 1×10^{-8} will give a clock rate error of about one microsecond per minute. Were it not for other factors, this error would limit the accuracy with which we could measure changes in the phase of a received signal over time and thus, say, the exact frequency of the signal. In practice, timing accuracy during analysis is limited by the signal-to-noise ratio of the recorded signals and the pilot tone, and we may regard the station frequency standard as being essentially perfect.

Clock. The clock counts cycles from the frequency standard to keep track of time. The clock is set by the operator according to standard time signals from shortwave stations such as WWV, and is usually accurate to within a few milliseconds. The clock generates the pilot tone and time mark signal which is recorded on the analog tape. The pilot tone is a constant-frequency signal, synthesized from the output of the frequency standard and possessing its stability, that serves as a phase reference in the recorded data. Pilot tone frequencies we have used are 1, 9, and 10 kHz. The time mark signal is encoded by amplitude modulation of the pilot tone. The pilot tone is increased by 10 dB for 40 ms on each second and for 1040 ms on the minute to generate time ticks. Station identification, day of year, and time of day are also inserted in Morse code during the first 20 seconds of each minute.

Mixer. The mixer combines the data signal from the VLF receiver and the pilot tone and time mark signal from the clock. The combined signal is fed to the analog tape recorder. The mixer also

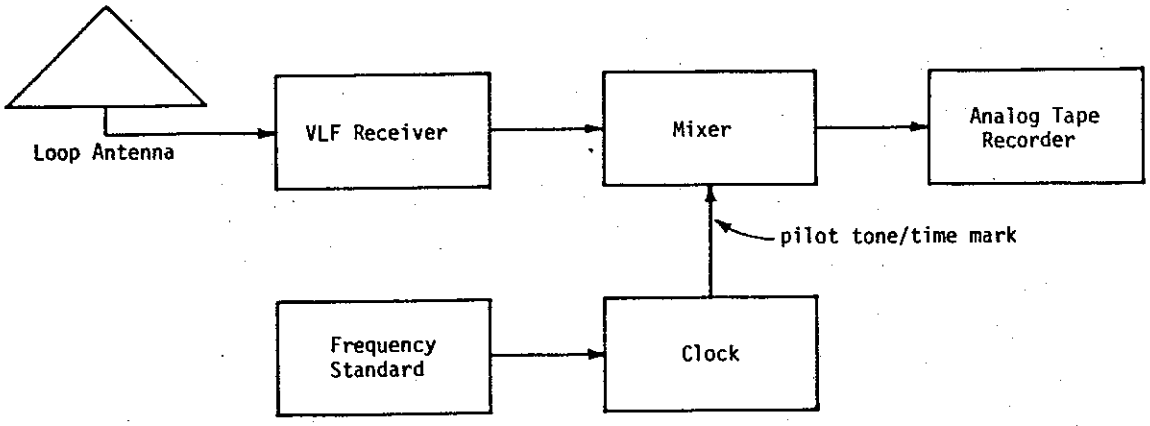

Figure 2.1. Block diagram of a typical VLF receiving system. The pilot tone/time mark signal is added to the recording to enable reconstruction of signal phases during analysis.

allows the operator to make voice annotations at the beginning of each tape recording, identifying the station, date and time, and nature of the recording.

Analog Tape Recorder. The recorders we use are high-quality professional recorders such as Ampex 350, AG-440, and ATR-100 models. They record the output of the mixer on 1/4-inch analog magnetic tape using direct (AM) recording. Usual tape speeds are 3.75, 7.5, or 15 inches per second (ips). Almost all recordings are made with half-track heads, recording one channel along half the width of the tape in one direction and then swapping tape reels to continue recording in the other direction along the other half. Used in this way, a 3600-foot reel of tape at 7.5 ips can hold 96 minutes of single-channel data on each side. We have also occasionally used quarter-track heads and made 4-channel recordings (one direction only) for direction-finding experiments.

One of the special characteristics of our recorders is that they have all been modified to use constant-current recording equalization, so-called because the recording head current (and thus the magnetic flux left on the tape) depends only on signal amplitude but not on signal frequency. Like the flat frequency response of the VLF receiver, constant-current equalization is used to maximize the dynamic range of the tape recording since the peak levels of VLF signals at high frequencies (over 10 kHz) are as large as those at middle and low frequencies. This equalization differs from standard audio practice; standard audio equalizations emphasize high frequencies during recording so they can be attenuated, along with tape noise, on playback.

Despite the care taken in the selection and maintenance of the analog tape recorder, it remains the weakest link in the system. The recorder has the least dynamic range of any signal-handling component, and limits the weakest or strongest signals which can be analyzed. The frequency response of the recorder is limited, especially at slower tape speeds, and it may not record higher frequency signals such as the various VLF transmitters from 15 to 25 kHz. Particularly bothersome are tape recorder wow and flutter. These are periodic variations in tape speed. Variations that repeat at intervals of a second or longer are called wow, and faster ones are flutter. Wow and flutter cause small timing variations when tapes are reproduced, and are the main source of error when measuring signal phase (causing much larger phase errors than those typically due, say, to noise in the recorded signal). Using a pilot tone to correct these tape timing errors is the major innovation of the present analysis system.

We are now experimenting with digital recording in the field using a Sony PCM converter and a video cassette recorder. This gives a larger dynamic range than with analog recording. It also greatly reduces, though it does not completely eliminate, the wow and flutter problem. Future field stations may record all signals digitally to overcome the limitations of analog recorders.

VLF Signal Characteristics. Now I will discuss some of the characteristics of the VLF signals as received and recorded. The minimum detectable received signal level is limited by the inherent noise of the loop antenna and the input stages of the receiver. The equivalent noise density of our receivers is typically less than 3×10^{-17} T/Hz$^{1/2}$ (equivalent to an incident wave electric field E of 1×10^{-8} V/m-Hz$^{1/2}$) from 3 to 30 kHz, rising slightly at lower and higher frequencies. The maximum signal which can be received without clipping is on the order of 1 nT (300 mV/m). The only signals stronger than this are impulsive spherics from nearby lightning strokes, and we are not particularly interested in them. The VLF receiver is capable of responding to a very large range of signal amplitudes, typically about 150 dB over a 1 Hz bandwidth.

The range of signals that can be recorded on an analog tape is considerably less than this. The minimum signal level is limited by the level of tape noise when the tape is played back (caused by the magnetic granularity of the recording medium). The maximum recordable signal is limited by the saturation flux density of the tape. Over the frequency range from 300 Hz to 20 kHz the typical 1/2-track analog tape has a broadband dynamic range between 50 and 60 dB. This is a range in a 1 Hz bandwidth (assuming white noise) of 93 to 103 dB, or about 50 dB less than that of the VLF receiver.

We see immediately that we cannot record the full dynamic range of the signals from the receiver. Instead, we can only record some portion of this range, determined by the amount of gain in the VLF receiver and the sensitivity of the recorder. The maximum gain we should use is that which brings receiver noise up to the level of tape noise; with any more gain we are just making a higher fidelity recording of receiver noise. In practice we usually use less gain than this to limit the intermodulation products that appear when strong signals (particularly VLF transmitters) saturate the tape.

All stations suffer to some extent from interference generated by local power lines. This appears as harmonics of 60 Hz that may extend to frequencies as high as 3 or 4 kHz. Usually the odd-numbered harmonics are the strongest. This interference is not from freely-propagating electromagnetic waves, of course, but is a local problem caused by currents induced in the antenna by the magnetic fields surrounding nearby power lines. This type of interference is easy to identify in the record because all of the lines in the spectrum are harmonically related and track each other in phase. At some stations this interference is very strong and it is necessary to filter out signals below, say, 1 kHz to prevent overloading at the tape recorder. One of the main criteria in selecting a receiving site is freedom from local power-line hum.

At frequencies above 10 kHz the spectrum is dominated by the various VLF transmitters used for navigation and communication. The signals from these transmitters can be very strong, especially at night, and at some stations it has been necessary to use filters to attenuate or eliminate certain ones. Two different schemes have been used. One is to use narrow notch filters to knock out one or two specific signals, such as NAA at 24.0 kHz and NSS at 21.4 kHz, which are quite strong in the eastern US and Canada. The rest of the spectrum is left as is in this case. This technique is good when only one or two transmitters are a problem, but the notch filter destroys any phase or amplitude information that those signals carry. That is, a notched signal cannot be examined for things like Trimpi events. The second scheme is to use a filter that attenuates all signals above

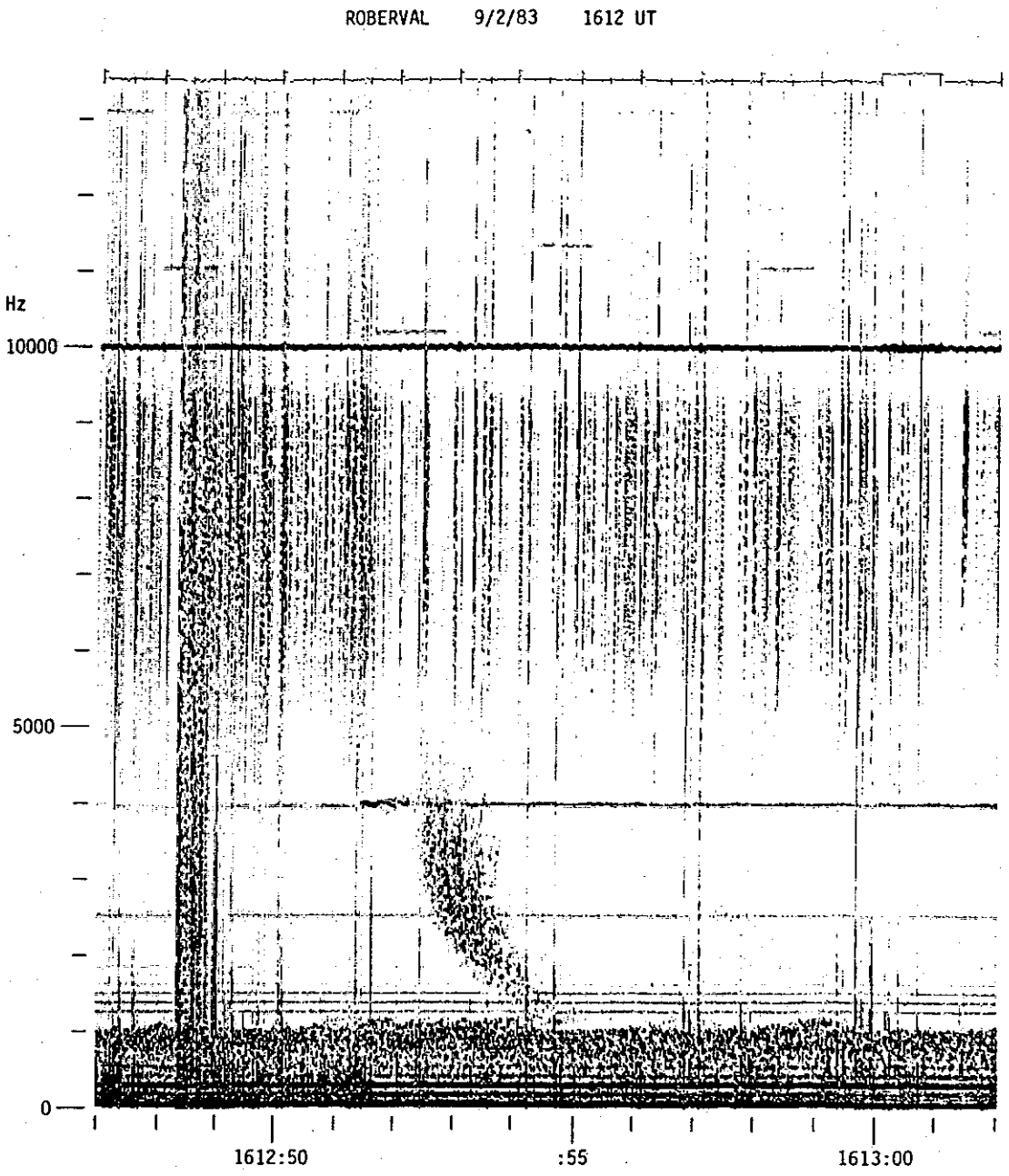

Figure 2.2. Spectrogram showing signal characteristics at Roberval, Quebec. Vertical lines are spheric impulses due to lightning. Horizontal lines below 2600 Hz are harmonics of the local power-line frequency. Chorus occurs around 1 kHz. At 3980 Hz there is a whistler-mode signal transmitted from Siple Station, Antarctica. A strong spheric causes a whistler echo, and a whistler precursor on the Siple signal (very unusual). The pilot tone is at 10 kHz, and above it are pulses from the Omega North Dakota transmitter.

10 kHz by, say, 20 dB. The advantages of this technique are that all VLF transmitters that might cause overloading are attenuated, yet their phase information is preserved. However, this type of filter (at least as we have implemented it) introduces a little ripple in the frequency response below 10 kHz.

Typical Spectrogram. Figure 2.2 is an *f-t* spectrogram of signals recorded at Roberval, Quebec, and shows some typical signals (as well as one very unusual one). The vertical lines are impulsive noise or *spherics* due to lightning strokes. Note that the intensity of all of the spherics drops above 9.5 kHz. This is due to a low-pass filter that attenuates signals above this frequency by 20 dB to limit interference from strong VLF stations. Many of the spherics also decrease in intensity below 5 kHz. This is not due to the receiver but to attenuation in the earth-ionosphere waveguide. Some spherics show no attenuation within the receiver passband. These are from lightning strokes relatively close to the station. Note in particular the very strong group of spherics at 1612:48.3. This group is strong enough to cause a whistler. The two-hop whistler echo is seen starting about :52.2, and there is a very faint four-hop echo about :57. Nose frequencies of different components of the whistler range from 3 to almost 4 kHz.

Horizontal lines in the spectrogram represent constant-frequency signals. Local power-line inteference is present as odd-numbered harmonics of 60 Hz from 60 to about 2600 Hz. There is also a constant-frequency signal just below 4000 Hz. This is not a power-line harmonic. It is a whistler-mode signal from the VLF transmitter at Siple Station, Antarctica, which was sending a two-tone signal at 3950 and 3980 Hz. An interesting feature is the short, variable-frequency signal that occurs on top of the Siple tones just before the whistler echo, at about :51.4. This is a *whistler precursor,* a very unusual signal that is discussed in Sec. 4.5.3.

The fuzzy signal on top of the power-line harmonics from about 700 to 1100 Hz is a band of *chorus.* This is a naturally-occurring signal of magnetospheric origin. The processes that generate chorus are not completely understood at the present time.

The strong signal at 10 kHz is the pilot tone, recorded to provide a phase reference for data analysis. Note that there is a little ripple at about a 6 Hz rate in the frequency of the pilot tone, due to flutter in the tape recording. This recording was processed to take out any average tape speed error (tracker time constant = 1 s), but not the faster speed variations. Above the pilot tone are approximately one-second pulses from the Omega navigation system transmitter in North Dakota. Signals are seen at 10.2, 11.05, 11.33$\overline{3}$, and 13.1 kHz. At the very top of the plot is a strip giving time ticks and showing the amplitude of the pilot tone. Time marks are recorded by amplitude modulation of the pilot tone. There is a 40 ms tick on each second (except seconds :09, :19, ..., :59) and a 1 s tick on the minute.

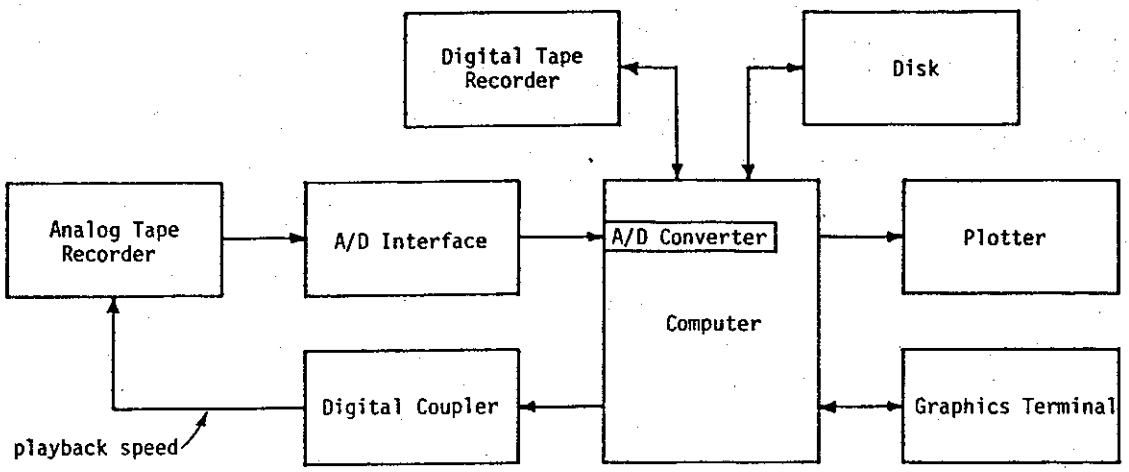

Figure 2.3. Block diagram of the digital analysis system. Analysis takes place in two steps. First the analog recording is played back, sampled, and digitized. Then the digital samples are processed and the results plotted. The playback speed control to the tape recorder is used to change the bandwidth of digitized signals, not as a servo to correct speed errors.

2.2 Digital Analysis System Components

In this section I describe the hardware of the digital analysis system. The reader will note that this system was designed in 1976, and there have been revolutionary changes in the power and cost of computing and signal-processing equipment since then. The digital analysis system is now obsolete in many ways. However, it is still useful to describe the system at hand since it incorporates many features that will be required in any more modern signal processing system. Chapter 5 has some suggestions for such a future system, and I will make a few comments here as well. Figure 2.3 shows the components of the digital analysis system, which are as follows:

Computer. This is the heart of the system. The Data General Eclipse S/230 computer is a general-purpose minicomputer well suited to low-cost dedicated data analysis. It has 64 Kbytes (32 Kwords) of core memory, just (barely) sufficient for our needs. Two features of the Eclipse particularly useful for us are a very fast floating-point unit and a writable-control-store (WCS). The WCS is a special section of one of the processor boards that allows us to program using microcode, special low-level and very fast (200 ns cycle time) computer instructions. The FFT algorithm has been written in microcode and runs much faster than would a similar program written in a high-level language or even in assembly language. This FFT routine is not as fast as those found in special-purpose signal processors or in computer systems with array processors, but is it fast by minicomputer standards.

The computer also contains interface boards which allow it to communicate with external devices such as the terminal, printer/plotter, disk drive, digital tape drive, and the digital coupler.

The use of a Data General machine was dictated by several factors. First, this computer and its associated peripherals are compatible with other computers in the VLF Group at Stanford. Second, Data General machines provided (at least when this system was purchased) the most number-crunching power per dollar, especially if the user was willing to do low-level programming.

To give the reader some idea of the speed of this machine, here are some instruction execution times. The Eclipse S/230 has a basic instruction time of 0.6 μs. This is the time to perform

most register-to-register arithmetic and logical instructions. Load and store register instructions (memory access) take 1.0 μs. A 16 × 16-bit integer multiply takes 7.2 μs. Floating-point operations are very fast—a single-precision multiply takes about 4.5 μs. (The floating-point unit occupies two 15-inch square boards, which is pretty big.) The WCS has a micro-instruction cycle time of 0.2 μs. Microcoded programs will run maybe half again faster than those written in assembly language. For purposes of comparison, an 8-MHz IBM PC/AT style machine (a personal computer using the Intel 80286 microprocessor) takes about 0.5 μs for register-to-register arithmetic, 0.75 μs to move a word between a register and memory, and 3.0 μs for a 16-bit multiply. If used with an Intel 80287 floating-point chip (at 5.33 MHz), a single-precision multiply might take 12.5 μs. Thus the Eclipse S/230 is comparable in power to a modern high-end personal computer.

Analog Tape Deck. An Ampex AG-440C analog tape deck is used to reproduce field recordings when digitizing data. This deck is a high-quality commercial unit just like those used at field stations. The analog deck takes 1/4-inch tape reels up to 10-1/2 inches in diameter (3600 feet of tape) and has half-track heads. It can reproduce two channels of data at once, but is usually used to play back one channel only. The AG-440C uses constant-current equalization to reproduce field tapes with maximum fidelity.

The transport has a servo-controlled capstan motor. The motor speed can be set by an external signal. The speed control signal is generated by the digital coupler (described below) under control of the computer, and allows a wide variety of tape speeds to be used, ranging from 1.172 ips up to 30 ips. Field tapes are all recorded at either 3.75, 7.5, or 15 ips, but using other speeds during playback allows us to digitize different effective data bandwidths without changing the digitizer sampling rate.

Tapes recorded with other equalizations or other head configurations can not usually be reproduced accurately on this machine. In particular, tapes made on standard stereo equipment can only be played if the tape has been recorded on one side only (since one quarter-width track for the reverse direction will also be picked up by the playback head). Also, most stereo equipment uses NAB equalization, where the tape flux at high frequencies is emphasized with respect to that at low frequencies, and these tapes will be reproduced with improper frequency response. Tapes made using frequency modulation and 1/2-inch audio tapes (such as NASA analog data tapes) cannot be played back at all.

A/D Interface. This is an analog instrument built at Stanford. The interface takes from one to four audio input signals in the range of 0.1 to 5 V peak, and amplifies them to the level needed by the A/D converter. Next, the signals are clipped to help eliminate impulsive noise due to spherics. Finally, the clipped signals are filtered to prevent aliasing during the sampling process. The frequency cutoff of the anti-aliasing filters is 10.6, 5.3, or 2.65 kHz for 1, 2, or 4-channel sampling, respectively. (Multi-channel sampling is done by interleaving samples from different inputs while the total sample rate remains fixed; hence the lower cutoff frequencies.)

The A/D interface also contains circuits to generate various types of calibration signals. A very pure 1 kHz tone can be generated to test for distortion effects during sampling or analysis. A comb signal with equal-level components every 200 Hz can be used to measure the system frequency response. Two pseudo-random signals can be used to check the performance of certain analysis routines. The short-period random signal repeats every 320 ms and generates a spectrum with components every 3.122 Hz. The long-period signal repeats only every 82.0 s, and has components every 0.01220 Hz.

The A/D interface also allows all input and output signals to be monitored with either a

loudspeaker or earphones, and their broadband amplitudes can be measured on a meter.

Digital Coupler. This is a Stanford-built unit. It contains a clock, and a set of input and output buffers to control the speed and functions of the analog tape transport. It also has buffers to receive tape deck motion commands and sampling start and stop commands from the operator via a small hand-held control box. The coupler has a series of LED indicators which are used during analysis to indicate such things as the elapsed data time and the instantaneous tape rate error.

The coupler contains a 1 MHz quartz crystal oscillator whose frequency is divided down to provide the 25.6 kHz sampling signal which triggers the A/D converter. This oscillator has an accuracy of only about 1×10^{-6}. However, this is so much more accurate than the speed of the analog tape deck that sample timing variations due to this oscillator are insignificant. In any case, a small rate error in the 25.6 kHz sampling signal will be corrected when the pilot tone signal is measured.

Analog to Digital Converter. The sample-and-hold and analog-to-digital (A/D) converter board is in the computer. The unit has eight differential inputs, though we use only one most of the time. The aperature uncertainty (jitter) of the sample-and-hold unit is specified as 5 ns peak. The A/D converter is a 12-bit unit with an accuracy of 0.25% $\pm 1/2$ LSB (least significant bit). The maximum system conversion rate is 28 ksamples/s.

The A/D converter board also contains address registers, word counters, and interrupt logic to perform direct memory access when digitizing. That is, it is only necessary to load an initial memory destination and a sample count, start the first conversion, and the rest will follow and be stored automatically in the computer memory.

Digital Tape Drive. The digital tape drive is used during data analysis to store the digitized signal data. This drive takes 10-inch reels of 1/2-inch digital tape, and writes at 800 bpi in NRZI format. One reel of tape can hold about 6.5 minutes of 10.6 kHz sampled data, either in one continuous stream or separated into shorter tape files. Because of the speed needed during digitizing (25600 samples or 51200 bytes per second), the tape drive is a vacuum-column unit that can read or write at 75 ips. Because the tape transport is working very close to its maximum capacity when digitizing data, there is no time for tape errors to be corrected. Tape performance must be perfect (no tape dropouts) and only high-quality magnetic tape can be used.

When this system was first proposed, 800 bpi was a common tape density. Since then other higher-density standards have come along, and one priority for modernizing the system would be to get a higher-density tape drive. Other types of data storage are also becoming available. For instance, cartridge tape drives may make more sense for desk-top analysis systems. Optical disks may soon be an even better choice.

Disk. This disk is a 10 megabyte top-loading unit that holds the various programs used as well as small sections of data. Most of the system and user programs reside on the lower fixed disk, which is always present. The upper removable disk (5 Mbyte) contains particular programs for different applications, and different disk cartridges may be fitted for special uses. All of the programs used in data analysis are contained in one disk cartridge, which also has about 4 Mbytes of blank space for temporary data and plot files.

Printer/Plotter. All permanent data output is produced on a Versatec D900A printer/plotter. This unit uses an electrostatic technique to write images on specially coated paper. Our unit uses fan-fold paper in 8-1/2 by 11 inch sheets. The writing head contains an array of 1600 writing nibs (each of which can write one tiny dot) covering a line 8 inches wide across the paper. When printing

or plotting, the paper is slowly advanced over the writing head and individual nibs are turned on or off as needed. The nibs leave small charges on the paper which is then flushed with toner. Toner particles settle out where the paper is charged, and the toner solvent is dried. The result is a high-quality half-tone plot with a resolution of 200 dots/inch in both the horizontal and vertical directions. In the data analysis system, the plot is always oriented so that time runs along the long edge of the paper (vertically), and frequency, magnitude, or phase is plotted along the short edge (horizontally). Plots are viewed sideways, much like photographic spectrograms on film or paper.

This plotter has the virtues of being relatively fast, say 10 seconds per page, depending on the type of plot; and cheap, about 3 cents per page. However, the paper is definitely not of archival quality. A better modern alternative would be a laser printer which uses ordinary xerographic paper and also has higher resolution.

Graphics Terminal. The operator communicates with the system via a Tektronix 4012 graphics terminal. The terminal has a keyboard for typing in commands and a display screen for reading the results. The terminal has the capability of drawing graphs on a matrix of dots 780 high by 1024 wide, sufficient for a quick look during analysis but not providing quite the detail available from the plotter.

This terminal provided the most graphics power at a reasonable cost when the analysis system was designed. However, it has several drawbacks. First, the terminal uses a storage display rather than a raster-scan display, so changing a display means erasing and then rewriting the whole screen. Second, graphic output is written by sending the beginning and ending coördinates of each vector to be drawn. If only a single pixel is to be turned on, it is viewed as a vector of zero length and still requires at least two bytes to specify. This process (as well as the limited speed of the writing gun in the display tube itself) limits the rate at which random graphs such as spectrograms can be drawn. Finally, the brightness, contrast, and actual resolution of the display are not as good as desired. A better choice at the moment would be a raster-scan display using directly-addressable bit-mapped graphics. Raster-scan displays are brighter and have more contrast than storage screens. Bit-mapped displays (where individual pixels are represented by individual bits in the main computer memory) are much faster to generate. A color display can show more data for a given-size picture matrix than a monochrome display, but more about this in Chapter 5.

2.3 Tape Timing Errors

Time and Rate Errors. Analog tape speed variations cause two related errors which must be corrected before accurate phase measurements can be made. First, and most important, speed variations cause the phases of recorded signals to vary. For instance, if a constant-frequency signal were recorded on a tape deck running below its correct speed, upon playback we would find that signal zero-crossings were occurring before the expected times, and this phase error would increase the further down the tape we looked. We will refer to this as the time error. Second, if the tape speed change is large enough, the signal frequency on playback will be shifted significantly, and may even be shifted outside the passband of the analysis filter. We will call this effect the rate error.

These two errors are obviously related, since the time error is just the integral of the rate error, but it is convenient to think of them separately since they require different techniques for correction during analysis. First, the analysis procedure must correct for the rate error so the desired signal is filtered correctly, and then the time error must be corrected to reconstruct the phase of the signal as originally received.

The rate error will also cause the amplitude of the signal on playback to vary, since playback head response increases with rising signal frequency, but this is compensated by the equalization electronics in the recorder and is not a problem. Short-term speed variations in quality recorders are small, on the order of 0.1%, and the amplitude variations they cause are insignificant. Long-term (average) speed errors may be larger, as much as 2 or 3%. In any case, because of amplifier gain uncertainty and variations in tape quality, to make absolute measurements of amplitude it is necessary to periodically record a calibration tone of known level.

To illustrate these concepts, let's look at an example. Consider a pilot tone at frequency f_p that is recorded on a field tape. The phase of the pilot tone as generated from the station frequency standard is given simply by

$$\phi_p(t) = 2\pi f_p t. \tag{2.1}$$

Assume that the tape recorder in the field was started at time $t_0 = 0$. At time t, the tape has moved a distance x and is recording a pilot tone phase of $\phi_p(t)$. Let's assume that the tape is moving with a velocity v_{rec} at this point. When the tape is done, we take it to the laboratory and play it back. Time in the laboratory will be called t'. The playback tape deck is started at time $t'_0 = 0$, and after running a time t' reaches the point x again. Let's assume that the tape on the playback deck is moving with velocity v_{pb} at this point. We measure the phase of the pilot tone, and find the value $\phi'_p(t')$. This will be the same value as was recorded at point x in the field, so $\phi'_p(t') = \phi_p(t)$. Now the question is, what was the time in the field when this signal was being recorded. The answer, of course, is

$$t = \frac{\phi_p(t)}{2\pi f_p} = \frac{\phi'_p(t')}{2\pi f_p}, \tag{2.2}$$

which is how we will reconstruct field time from the pilot tone phase. We will call the difference in time $t - t' = [\phi'_p(t')/2\pi f_p] - t' = t_{err}$, the *time error*. This is the difference between the tape time, as indicated by the pilot tone, and time on a stopwatch which was started when we began playing back the tape.

The frequency of the pilot tone $f'_p(t')$ reproduced from the tape may be different from what was recorded unless the record and playback deck speeds are the same. The frequency on playback is given by

$$f'_p(t') = \frac{1}{2\pi}\frac{d\phi'_p(t')}{dt'} = \frac{v_{pb}}{v_{rec}} f_p = r(t') \cdot f_p \tag{2.3}$$

where

$$r(t') \equiv \frac{v_{pb}}{v_{rec}} = \frac{f_p'(t')}{f_p} \tag{2.4}$$

is the *relative data rate*. Since the speeds of the two tape decks are generally not constant, the relative data rate is a function of time. It shows at any given moment the ratio of playback frequency to original frequency of all signals on the tape, not just the pilot tone. The data rate r is usually fairly close to 1.00. We will call the difference $r(t') - 1 = r_{err}$, the *rate error*.

The pilot tone phase as measured in the laboratory is given from Eq. (2.3) by

$$\phi_p'(t') = \int_0^{t'} 2\pi f_p'(\tau)\, d\tau = \int_0^{t'} 2\pi r(\tau) f_p\, d\tau \tag{2.5}$$

so we can write

$$t_{err} = \frac{\phi_p'(t')}{2\pi f_p} - t' = \int_0^{t'} r(\tau) - 1\, d\tau = \int_0^{t'} r_{err}\, d\tau. \tag{2.6}$$

Thus the time error is the integral of the rate error, as stated above.

Example of Typical Tape Errors. Figure 2.4 shows the magnitude and phase of the pilot tone in a typical recording. These plots were generated by turning off the pilot tone tracker in the analysis program and plotting the actual values observed. The top plot shows the time code amplitude modulation. The pilot tone level is increased 10 dB for one second on the minute, for 40 ms on the second, and also to give the Morse code station identification and time information. The station is -·- ––– or "RO" for Roberval, and the next three characters are ·–––– ·–––– ––··· or "117" giving the day of year for April 27. Two-digit codes for the hour and minute will follow the day number.

The middle plot shows the relative phase of the 1 kHz pilot tone with respect to a reference signal at exactly 1 kHz. What is plotted is the time difference $(\phi_{pilot}/1\text{ kHz}) - t'$, where ϕ_{pilot} is in units of revolutions, and t' is time in the reference frame of the analysis laboratory. The phase is plotted in units of time, 1000 microseconds full-scale. This is, of course, exactly one revolution of phase at 1 kHz. Time units are used to emphasize that this plot shows the time error of the recording. Note that the relative phase of the pilot tone is advancing with time with respect to the 1 kHz reference. Time on the tape as played back is running by faster than it was in the field when the tape was recorded (when the pilot tone was exactly 1 kHz). Either the field tape deck that made this recording was running slow, or the one in the lab that played it back was running fast, or both. The speed-up in time is seen to be about 10.5 ms in 12 seconds, giving an average rate error of +0.088%. Looked at another way, the pilot tone is seen to gain 0.88 revolutions of phase per second with respect to the 1 kHz reference, so it was reproduced at a frequency of 1000.88 Hz.

The bottom trace shows the pilot tone phase to an expanded scale. The uniform phase advance due to the mean rate error has been removed (by using a phase reference at 1000.85 Hz) and what remains are some smaller phase variations about the mean. The structure of these remaining phase errors is somewhat complicated, but there are two obvious regularities. There is a roughly sinusoidal error component with a peak-to-peak amplitude of 85 μs at a frequency of 2 Hz. And there is a smaller but faster variation about 14 μs p-p at a frequency of 20 Hz.

We have expressed these small fluctuations as time errors. We can differentiate and express them as rate errors as well. We find that the 2 Hz fluctuation corresponds to a rate error of $85 \times 10^{-6} \cdot 2\pi \cdot 2 = 0.107\%$ p-p, and the 20 Hz one to a rate error of $14 \times 10^{-6} \cdot 2\pi \cdot 20 = 0.176\%$ p-p.

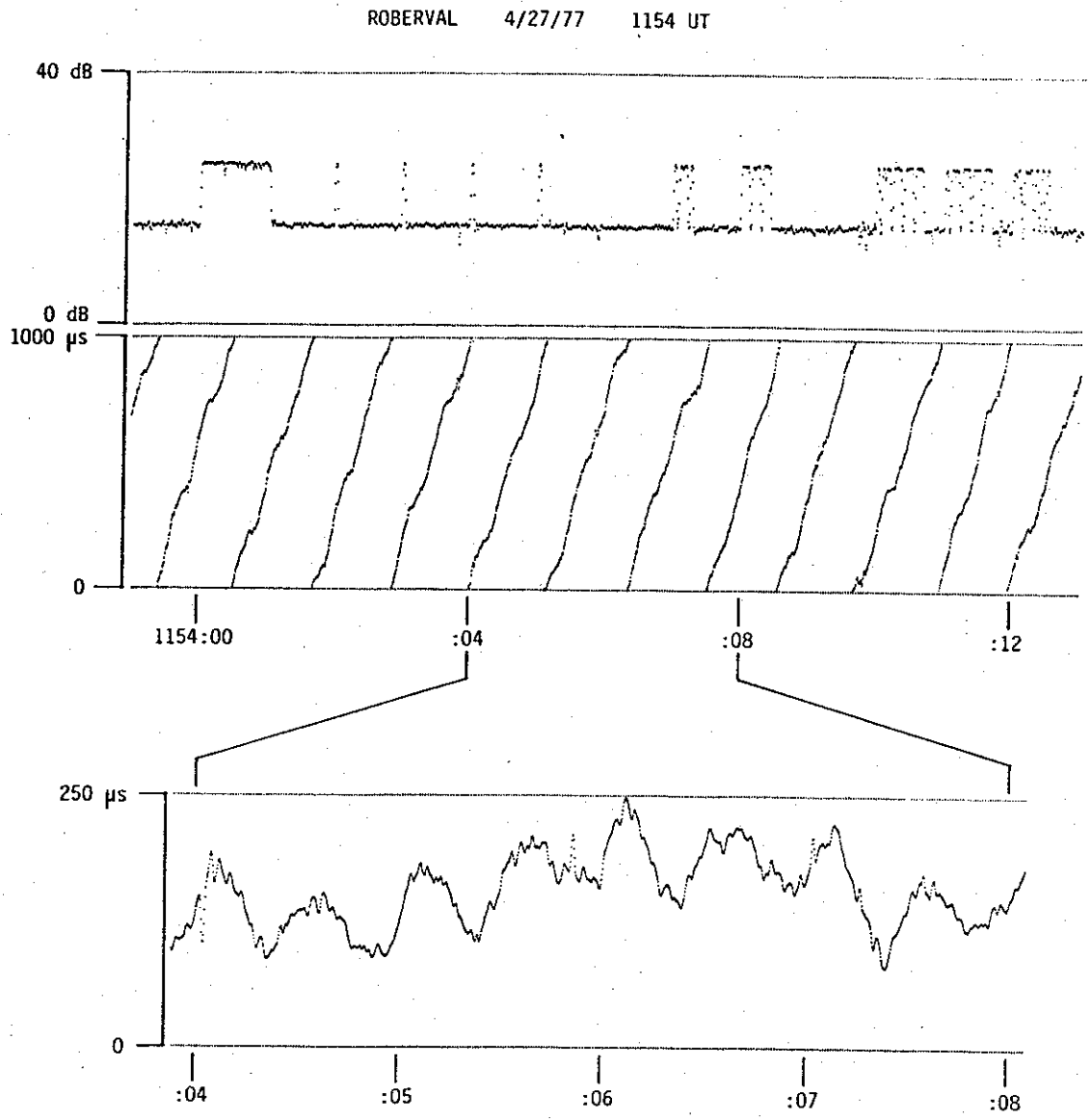

Figure 2.4. Magnitude and phase of the 1 kHz pilot tone on a typical analog recording. The top trace shows magnitude of the pilot tone, with 10 dB amplitude modulation due to time ticks and station and time codes. The middle trace shows the phase relative to a 1 kHz reference. There is an average rate error of +10.5 ms in 12 s, or +0.088%. The bottom trace shows the phase with the average error removed. The residual errors are due to the record deck idler pulley (2 Hz ripple) and capstan (20 Hz ripple). Analysis filter bandwidth is 40 Hz.

Assuming that these two components are independent and make up the bulk of the short-term errors, the total peak-to-peak fluctuation about the mean rate error is the sum 0.107+0.176 or 0.283% p-p.

Causes of Tape Timing Errors. The +0.088% average speed error may be caused by the record deck, the playback deck, or some combination of the two. Possible causes of static speed error include an undersize capstan drive shaft due to wear, and tape slippage due to improper tape tension or improper capstan pinch roller pressure. Tape decks with synchronous capstan motors (all Ampex

TABLE 2.1

Rotational Sources of Tape Timing Fluctuations

Source	Diameter	Frequency of Fluctuations *vs.* Tape Speed		
		3.75 ips	7.5 ips	15 ips
3.75/7.5 synchronous capstan	0.118 in	10. Hz	20. Hz	
7.5/15 synchronous capstan	0.235 in		10. Hz	20. Hz
AG-440C servo capstan	0.380 in	3.1 Hz	6.3 Hz	12.5 Hz
small supply idler	1.2 in	0.99 Hz	1.99 Hz	3.98 Hz
large supply idler	1.5 in	0.80 Hz	1.59 Hz	3.18 Hz
pinchroller	2.0 in	0.60 Hz	1.19 Hz	2.39 Hz
tape reel, empty	4.5 in	0.27 Hz	0.53 Hz	1.06 Hz
tape reel, full	9.5 in	0.13 Hz	0.25 Hz	0.50 Hz

350 series and the AG-440B model) are also affected by power-line frequency variations, which can be as much as 0.1% in North American power grids. Susceptible tape decks at sites with power from less-stable local generators (such as at Antarctic stations) must be run from frequency-stabilized supplies to avoid this problem. The AG-440C and ATR-100 decks with servo-controlled capstan motors are immune to line-frequency problems.

The residual phase errors in the lower plot of Figure 2.4 are caused in this case by the record deck. This tape was recorded at 7.5 ips on an Ampex tape deck using a synchronous 3.75/7.5 ips capstan motor. Overall tape tension is controlled by the torque of the supply and takeup reel motors. However, the actual motion of the tape is controlled by friction with the supply idler pulley (a free-running shaft with a large flywheel) on one side of the heads and the capstan motor drive shaft on the other. The supply idler on this deck has a diameter of 1.2 inches. At a tape speed of 7.5 ips it rotates at 2 revolutions per second (rps). The capstan motor rotates at exactly 20 rps in its 7.5 ips mode, and the capstan shaft has a nominal diameter of 0.118 inches. The 2 Hz and 20 Hz phase ripples are caused by small eccentricities in the idler and capstan.

If the nominal tape speed is v and the diameter of a circular shaft over which the tape runs is d, then the shaft rotates at a uniform rate of f rps where $f = v/\pi d$. If the shaft is circular but its geometrical center is offset from its center of rotation by an amount ξ, then with each rotation of the shaft the tape will at some time be advanced from its mean position vt by ξ, and at another time retarded by $-\xi$. If the offset ξ is small, then the displacement of the tape at the shaft will be sinusoidal in time, with peak-to-peak amplitude 2ξ. If the record/playback heads are midway between the idler and capstan shafts, then a p-p displacement of 2ξ at one shaft will cause a displacement over the head of about half that amount or ξ, since the elastic tape is anchored at the other end by the other shaft. This displacement at the tape head will cause a time error of ξ/v p-p. Thus the 85 μs and 14 μs p-p pilot tone phase errors are caused by eccentricities of only 0.00064 and 0.00011 inches (16 and 2.7 μm), respectively, in the idler and capstan shafts.

Table 2.1 lists the various rotating parts of the analog tape decks that can cause tape phase fluctuations. Because of its relatively small diameter, the 3.75/7.5 ips capstan motor is particularly susceptible to damage in the field. Tape flutter with this capstan motor can be especially bad. Also, mean rate errors are often 1% or more. The 0.088% error in Figure 2.4 is actually surprisingly small for this capstan motor. The 7.5/15 ips and AG-440C capstan motors are much better than the 3.75/7.5 ips type.

Damage to the pinchroller usually involves dimples or bumps on its surface. These cause sharp phase irregularities rather than the sinusoidal variations due to off-center shafts.

Since the torque of the reel motors is constant while the tension exerted on the tape depends on the distance from the center of the reel to the edge of the wrap, an off-center reel can cause phase errors by causing variations in tape tension. The frequency of any effect depends on the amount of tape on the reel. Minimum and maximum diameters and corresponding fluctuation frequencies are listed. Noise due to a reel scraping some external object and tape stretch due to a bad wrap on a reel also occur with these fluctuation frequencies and may be important in some cases.

Inter-Channel Jitter on Multi-Channel Recordings Due to Tape Skew. One final source of tape timing errors needs to be mentioned. This error, inter-channel jitter, is an error in timing that occurs between two tracks of a multi-channel recording and is caused by azimuthal tape skew. Tape azimuth is the angle which tape motion makes with respect to a line drawn through the head gaps. If the head gaps are not exactly perpendicular to the motion of the tape, then signals in one track will cross their head gap slightly before or after those in some other track, and signals recorded simultaneously but on different tracks will not be played back at exactly the same time. If the error in tape azimuth is constant then the delay from one track to another will be constant and may not pose a problem. However, if the azimuth changes then the delay will change as well.

I have found when trying to process 4-track recordings that there is typically a jitter between channels of about 25 μs peak-to-peak at a tape speed of 7.5 ips. This corresponds to a peak inter-channel displacement of 0.000094 inches (2.4 μm). If the two channels are 1/8 inch apart across the width of the tape, then this represents a peak shift of only 2.6 minutes of arc in the azimuth of the tape about its average value. This sort of shift can occur over the 3-inch distance between the tape guides on either side of the tape head if the clearance at each guide is only 0.001 inch. This error probably cannot be reduced much.

A similar problem may be caused by distortion of the tape surface. Unless the tape is wrapped very evenly on the reel (which does not happen when the tape deck fast-forward or rewind functions are used), the tape will tend to roll sideways over inner layers. Since the tape is wound under a slight tension, this may cause the tape to stretch differentially from edge to edge at different portions of the wrap. When the tape is played back this stretching will cause azimuth errors.

Another potential source of inter-channel jitter is excessive head scatter. This is the displacement of the different gaps in the head stack from their mean position on a line perpendicular to the tape, either forward or backward in the direction of tape motion. This gives a delay between signals on different tracks. For a constant tape speed the constant delay produced may not be a problem, but if the tape speed varies the delay will also vary. However, unless the head scatter is more than a few thousandths of an inch, the jitter caused by typical speed variations is unlikely to exceed 1 μs.

The problem with inter-channel timing jitter is that it makes it impossible to compare the phases of signals recorded on different tracks of a given tape. A pilot tone recorded on one track cannot be used to rectify timing errors on a different track. Though the average rate error can be corrected, the inter-channel jitter cannot. If phase measurements are to be made from multi-channel recordings, each channel must include its own pilot tone.

2.4 Sampling and Digitizing

The conversion of an analog signal to digital form entails first sampling the signal, that is, measuring its waveform at periodic intervals; and then digitizing or converting these samples into digital numbers. In this system, signal sampling takes place at a rate of 25600 samples per second. The sample-and-hold circuit on the A/D converter board in the computer is triggered by a 25.6 kHz pulse train from the analog interface. At each pulse, the instantaneous input signal voltage is captured (a capacitor is charged) and held temporarily (the charge switch is opened) while being converted into a digital number by the A/D circuit. After the conversion, the A/D board stores the number in the computer memory. When enough samples have been accumulated, the computer writes out a block of samples to either the digital magnetic tape or the disk.

The process of sampling and digitizing is necessary in order to convert the analog signal from the tape recorder into a form that can be processed by the computer. However, this conversion can introduce various errors. Frequency aliasing, due to improper filtering before sampling, can cause components at one frequency to appear somewhere else. Phase distortion in the anti-aliasing filter can cause different phase shifts between different signal components if there is wow and flutter in the tape recording. And sample-and-hold jitter and A/D converter quantization error can introduce noise. We will consider each of these potential problems in turn.

Sampling and Aliasing. The sampling theorem [*e.g., Bracewell,* 1965, Ch. 10] states that if a signal is known to contain components only at frequencies below some maximum cutoff frequency f_c, then it is possible to precisely reconstruct the signal from evenly-spaced samples of it taken $2f_c$ times per second (the Nyquist rate). However, if the signal contains components at frequencies above f_c, then there is not enough information at $2f_c$ samples/s to reconstruct the original signal. If such a reconstruction is nevertheless attempted, what will happen is that higher frequencies (above f_c) in the original signal will appear as lower frequency components (below f_c) in the reconstructed signal. In fact, if the frequency is above f_c but less than $2f_c$, say at $f_c + x$, it will be reconstructed at $f_c - x$. This error in reconstruction is known as *frequency aliasing.* Thus, for a given sampling rate it is necessary to ensure that the bandwidth of the input signal is limited to at most one-half of that rate.

With a sampling rate of $2f_c = 25600$ samples/s, it is theoretically possible to retain information about all input signal components from zero frequency up to $f_c = 12.8$ kHz. Since the analog signal reproduced from the field recording typically contains components above this frequency, it is necessary to filter it before sampling. In the ideal case we would process the analog signal by feeding it through a filter that passed components at frequencies below 12.8 kHz but rejected those at higher frequencies. In practice, it is impossible to make a filter with an infinitely sharp cutoff. A real low-pass filter is specified by a passband from 0 Hz to a maximum frequency f_p, throughout which there is a maximum permissable attenuation (passband ripple) A_{max}; and a stopband from frequency f_s on up, where the attenuation is always at least some minimum value A_{min}. The region between f_p and f_s is the transition region where the attenuation is unspecified, though it will generally be between A_{max} and A_{min}. The optimum low-pass filter is specified for a given sampling bandwidth f_c when $f_s - f_c = f_c - f_p$. When this is true, signals in the passband below f_p will be passed, and signals above f_s aliased into the passband will be attenuated by at least A_{min}. However, the region from f_p to f_c is subject to unknown attenuation, and to unknown amounts of aliasing from signals in the region f_c to f_s. This optimum design minimizes the contaminated region $f_p < f \leq f_c$.

In this system we use a *7th*-order Cauer (elliptic) low-pass filter designed following *Zverev* [1967]. The filter used is a CC072045 design in his notation. It is flat (within 0.1 dB) from dc to 10.6 kHz,

but attenuates signals above 15.0 kHz by more than 70 dB. This ensures the fidelity of sampled data up to 10.6 kHz. However, signals from 10.6 to 12.8 kHz will be subject unknown attenuation and to aliasing by signals from 12.8 to 15.0 kHz. The analysis program automatically discards spectral points in the contaminated band above 10.6 kHz.

(The above discussion refers to the sampling of a single-channel signal reproduced at the same tape speed as recorded. Playing back tapes at different speeds changes the effective bandwidth of the sampled data, while the actual sampling rate in the laboratory remains the same. Also, multi-channel sampling involves interleaving samples from different inputs. In this case, different anti-aliasing filters are used, with cutoffs of 5.3 or 2.65 kHz, for 2- or 4-channel sampling, respectively.)

Phase Distortion in the Anti-Aliasing Filter. In Eq. (2.2) above we showed how to find the tape time t given a measurement of the pilot tone phase ϕ_p. In fact, what we measure is not just the phase ϕ_p but also phase shifts introduced by all of the various circuits through which the signal passes on its way from the tape to being digitized. Most of these additional phase shifts are small and are of no concern. Some phase shifts are proportional to frequency and just introduce constant time delays, again of no concern. However, the phase shift introduced by the anti-aliasing low-pass filter is neither small nor linear with frequency and may be a source of error.

Let us call the phase shift of the filter $\theta(f)$. At a given frequency, this phase shift introduces a *phase delay* $t_{ph}(f) = \theta(f)/2\pi f$ [*Papoulis*, 1962, Eq. (7-58)]. This is the steady-state delay of a constant-frequency signal passing through the filter. If the delay were some constant value for all signals it would pose no problem. Even if the delay were different for different signal components, but still constant for each, it would not be a concern. Ultimately we are interested in measuring the relative phases of narrow-band signals as functions of time, and not in comparing the absolute phases of different components, so a constant phase shift between one component and another at a different frequency is not a problem. The problem occurs when the data rate changes and the frequencies of all signal components change. If the phase delay at each frequency does not change by exactly the same amount for a given data rate change, then we will be unable to use the phase of the pilot tone to reconstruct the phase of some other signal since the other signal will have suffered a different and unknown delay. Our task here is to estimate the differential phase delay between components at two different frequencies for a given change in data rate.

Consider a component at frequency f. If the data rate changes from r to $r' = r + \delta r$ then the filter phase shift will change from $\theta(rf)$ to

$$\begin{aligned}\theta(r'f) &= \theta((r+\delta r)f) \\ &\approx \theta(rf) + \delta r \cdot f\frac{d\theta}{df} \qquad \text{for small } \delta r.\end{aligned} \tag{2.7}$$

This will give a delay in time of

$$\delta t = \frac{\delta r \cdot f}{2\pi f}\frac{d\theta}{df} = \frac{\delta r}{2\pi}\frac{d\theta}{df} = \delta r \cdot t_{gr}(rf) \tag{2.8}$$

where

$$t_{gr}(f) = \frac{1}{2\pi}\frac{d\theta(f)}{df} \tag{2.9}$$

is the *group delay*. If the group delays are different at the pilot frequency f_p and at some signal frequency f_s, then a rate change of δr will give a differential time delay

$$\Delta t = \delta t_p - \delta t_s = \delta r[t_{gr}(rf_p) - t_{gr}(rf_s)]. \tag{2.10}$$

The response of the anti-aliasing filter is determined by the locations of the poles and zeroes of its transfer function, which are placed as follows (in units of Hz):

$$\begin{array}{lrll} \text{Poles at} & -4257.070 & & = \sigma_0 \\ & -3397.053 & \pm j5766.480 & = \sigma_1 \pm j\nu_1 \\ & -1815.235 & \pm j9346.958 & = \sigma_2 \pm j\nu_2 \\ & -532.103 & \pm j10822.336 & = \sigma_3 \pm j\nu_3 \\ \text{Zeroes at} & & \pm j15261 & = \pm j\nu_4 \\ & & \pm j18040 & = \pm j\nu_5 \\ & & \pm j30270 & = \pm j\nu_6 \end{array}$$

The group delay of the filter is given from the locations of its poles by [*Zverev*, 1967, Eq. (5.3.3)]:

$$t_{gr}(f) = \frac{1}{2\pi}\left(\frac{\sigma_0}{\sigma_0^2+f^2} + \frac{\sigma_1}{\sigma_1^2+(f-\nu_1)^2} + \frac{\sigma_1}{\sigma_1^2+(f+\nu_1)^2} + \cdots + \frac{\sigma_3}{\sigma_3^2+(f+\nu_3)^2}\right). \qquad (2.11)$$

Note that the zeroes ν_4–ν_6 are all on the jf axis in the stopband, and have no effect on the group delay in the passband.

TABLE 2.2

Anti-Aliasing Filter Group Delay *vs.* Frequency

f	t_{gr}	f	t_{gr}	f	t_{gr}
0 kHz	69.3 μs	9.5 kHz	159.4 μs	10.9 kHz	365.5 μs
1	68.7	10	193.0	11	338.5
2	68.0	10.1	206.5	11.1	301.5
3	69.6	10.2	224.0	11.2	262.4
4	74.6	10.3	246.3	11.3	226.4
5	81.8	10.4	273.8	11.5	169.8
6	87.8	10.5	305.4	11.7	131.6
7	94.4	10.6	337.6	12	96.0
8	111.6	10.7	363.4	13	48.5
9	142.9	10.8	374.5	14	31.7

Table 2.2 shows the filter group delay t_{gr} as a function of frequency. The group delay has a minimum value of 68 μs around 2 kHz, and a maximum value of 375 μs near 10.8 kHz. The highest analyzed signal frequency is 10.6 kHz (higher-frequency components may be aliased and are thrown out) where the group delay is 338 μs.

Let's assume that the variation in data rate about the mean is $\delta r = 0.3\%$ peak-to-peak, a little bit worse than that seen in Figure 2.4. Then the worst-case differential time delay, which will occur between a signal at 2 kHz and one at 10.6 kHz, will be $\Delta t = 0.003 \cdot (338 - 68)$, or 0.81 μs p-p. This is not zero, but it is not very great, either. For signals below 10 kHz the error will be at most half of this, and even less at lower frequencies, since the filter group delay flattens out rapidly as we move away from the edge of the passband.

Most of the interesting effects we will study involve phase shifts of hundreds of microseconds or more, and distortion introduced by the anti-aliasing filter will be unnoticable. There may be some cases, though, such as when looking for Trimpi event phase shifts (only a few microseconds in many

cases) on VLF transmitter signals (which may be at the upper end of the filter passband where distortion is worst) where the anti-aliasing filter together with wow and flutter will limit the effects we can see.

Sample-and-Hold Jitter. When the filtered input signal is sampled, we must ensure that the samples taken are evenly spaced in time. The sampling jitter, or "aperture uncertainty," is the difference between the time when a sample is supposed to be taken (a multiple of 1/25600 second in our case) and the time it actually is taken. The problem here is that a changing signal sampled at the wrong time will have a different voltage than it had at the specified time. If the maximum frequency of the sampled signal is f and its peak voltage V, then the maximum rate of change of voltage with time is $2\pi fV$. The maximum relative error is then $2\pi fV\tau/V = 2\pi f\tau$, where τ is the jitter in seconds.

The aperture uncertainty of the sample-and-hold unit in our system is specified as ± 5 ns. If the maximum frequency of the sampled signal is 10.6 kHz, then the peak sampling error due to jitter will be $2\pi 10.6 \times 10^3 \times 5 \times 10^{-9} = 0.033\%$, or 70 dB below the peak sampled signal. Since only rarely will we sample signals with such large high-frequency components, most of the time the errors due to sampling jitter are smaller than those due to quantizing, which are discussed next.

Quantizing Error and Dynamic Range. Once signal samples have been taken, they are converted into digital numbers (integers) by the analog-to-digital (A/D) converter. However, there is an error of approximation called the *quantizing error* involved in this conversion. The error arises because each sample of analog signal, which can take on a continuous range of values, must be represented by a digital number, which can have only a finite number of values. The quantizing error limits the smallest signal that can be converted, and thus determines the dynamic range of the signals that can be represented digitally. If the quantizing error is statistically independent from one sample to the next (as is usually the case with real signals) then the effect of quantizing is to add a certain amount of white noise to the sampled signal.

The precision of an A/D converter is usually given as so many "bits." A p-bit converter approximates an input voltage as being one of $Q = 2^p$ equally-spaced levels. For instance, our 12-bit converter assigns an input sample to one of $Q = 4096$ different integers from -2048 to $+2047$, representing signals from -5 to $+5$ V in increments of 10 V/4096 = 2.44 mV. However, unless the input sample falls right at one of the quantized levels it will not be represented exactly. Instead, the input will be approximated as being at the nearest level, with some error e. If the input range of the converter is $-V$ to $+V$, then the quantizing interval is $2V/Q$ volts. The error e ranges uniformly from $-V/Q$ to $+V/Q$ (that is, never more than half the interval), with probability density $Q/2V$. The mean squared error is the expected value of e^2 or $V^2/3Q^2$, and the rms error is then $V/\sqrt{3}Q$.

We will define the dynamic range of the A/D converter as the ratio of the largest undistorted sinewave which can be converted, to the rms error (noise) in the conversion. (Note that this is not the common definition, which is usually something like the ratio of the peak convertible value to the rms noise, a definition reminiscent of hi-fi amplifier ads.) The maximum sinewave signal which can be digitized has a peak value of V, or an rms value of $V/\sqrt{2}$. The dynamic range of the converter is thus $\sqrt{3/2} \cdot Q$ or $\sqrt{3/2} \cdot 2^p$.

The dynamic range of a 12-bit converter is found to be 74 dB. This assumes, of course, that the A/D converter has no other errors besides the quantizing error, such as non-linearity or missing codes. The dynamic range of our A/D converter is not quite 74 dB because of these additional errors, but it is still quite a bit greater than the broad-band dynamic range of the analog tape recording (50–60 dB). Thus the quantizing noise of the A/D converter is not significant compared

to the inherent noise of the analog tape.

2.5 Analysis Algorithms

Types of Algorithms. The algorithms I will discuss fall into three catagories. First are those procedures used in most signal processing tasks. Some of these, such as the FFT algorithm, are widely known, though not always optimally used. Others, such as windowing, are not known as widely as they should be. The presentation of these algorithms is in the nature of a review. Anyone who sets out to build an analysis system must know this stuff.

Second are those algorithms such as the tracking of the pilot tone and interpolation of the output spectrum that deal with the correction of analog tape timing errors. These procedures have been developed especially for this system to enable us to make signal phase measurements from analog field recordings. In the future we will have recorders without the wow and flutter of analog tape decks and future analysis system designers won't have to worry about these algorithms.

Third are those algorithms such as the calculation of relative phase and formatting for output that have been developed to display the results of analysis. Some of these, such as the plotting of *f-t* spectrograms, are merely the translation to the digital world of techniques that analog analysis systems have used for a long time. Others, such as the gray-scale or width-modulated phase format, are new. This is an area ripe for further invention.

As well as the spectrum analysis program itself, the digital analysis system has three other programs that are used during data analysis. These are a program to digitize data and record the samples on digital tape or in a disk file, a program to copy sampled data files, and a program to generate synthetic data samples mathematically for testing various analysis procedures. The design of these programs is straightforward and will not be described here.

Basics of Digital Spectrum Analysis. First I want to introduce some of the basic ideas of digital signal processing to lay a foundation for the discussions of algorithms in the following sections. This is basic stuff and the knowledgeable reader can skip ahead. However, the presentation here is from a slightly different viewpoint than usual. A common presentation of the order-N discrete Fourier transform is as the Fourier series coefficients of a bandlimited repetitive waveform. This is true, of course, but it's of little help when we're trying to understand signals which are not repetitive. It is more appropriate, being familiar with analog spectrum analyzers, to view the DFT as an approximation to the windowed spectrum or the output of a bank of filters. For additional details about spectrum analysis and the windowed Fourier transform the reader is referred to *Rabiner and Gold* [1975, Ch. 6], *Rabiner and Schafer* [1978, Ch. 6], or *Cadzow* [1987, Sec. 3.9–3.14]. For practical windows the reader must consult *Harris* [1978] and, especially, *Nuttall* [1981].

In the digital world we represent an analog waveform as a series of equally-spaced samples. How can we represent and calculate the digital equivalent of its spectrum? Assume we have a signal $x(t)$ and its Fourier transform $X(f)$, given before as

$$X(f) = \int_{-\infty}^{+\infty} x(t) e^{-j\,2\pi f t}\, dt. \tag{1.1}$$

We sample $x(t)$ at increments of time T to generate the sequence of samples $\{x_n\}$, where

$$x_n = x(nT) \qquad \text{for } n = 0, 1, \ldots. \tag{2.12}$$

We define the *infinite-time* discrete Fourier transform of this sequence to be

$$X_D(f) = \sum_{n=-\infty}^{+\infty} x_n e^{-j2\pi fnT}. \tag{2.13}$$

Now, if $x(t)$ is bandlimited to the sampling cutoff frequency $f_c = 1/2T$, that is, if $X(f) = 0$ for $|f| > f_c$, then we can show [*e.g., Rabiner and Gold,* 1975, Sec. 2.12]:

$$X_D(f) = \frac{1}{T} X(f) \qquad \text{for all } |f| \le 1/2T. \tag{2.14}$$

We see a direct relationship between the spectrum of an analog waveform as given by its Fourier transform and the corresponding discrete transform of its samples given by Eq. (2.13). The only difference is that the discrete spectrum $X_D(f)$ is periodic in f with period $1/T = 2f_c$ (since $\exp(j2\pi m) = 1$ for any integer m). That is, $X_D(f+m/T) = X_D(f)$. We are only interested in $X_D(f)$ at frequencies $|f| \le f_c$.

As we saw in Chapter 1, spectrum analyzing non-stationary signals involves calculating the windowed spectrum $S(t_0, f)$, given by Eq. (1.2), at various times t_0. We need to do the analogous thing here. Given a weighting function $w(t)$, we will sample it at times nT to get a sequence $\{w_n\}$. We define the discrete windowed spectrum S_D, evaluated at time $t_0 = mT$ and frequency f, to be

$$S_D(mT, f) = \sum_{n=-\infty}^{+\infty} x_{n+m} w_n e^{-j2\pi fnT}. \tag{2.15}$$

Note again that the window is stationary in time, and the signal moves through it. We can also express S_D by analogy to Eq. (1.3) as a convolution in the frequency domain as

$$S_D(mT, f) = \int_{-1/2T}^{1/2T} X(\nu) e^{j2\pi\nu mT} W_D(f-\nu)\, d\nu = X(f) e^{j2\pi fmT} * W_D(f) \tag{2.16}$$

where

$$W_D(f) = \sum_{n=-\infty}^{+\infty} w_n e^{-j2\pi fnT}. \tag{2.17}$$

We see that $S_D(mT, f)$ is almost the same as the windowed spectrum $S(t_0, f)$. There is a slight difference, however, because the discrete window function $W_D(f)$ is periodic in frequency with period $1/T = 2f_c$. This causes frequency aliasing in those synthesized filters near the cutoff frequency f_c. For example, consider a window function $W(f)$ that represents an equivalent lowpass filter with bandwidth b. $W(f)$ is zero outside the interval $[-b, b]$. We sample $W(f)$ to get the weighting sequence $\{w_n\}$. When we evaluate Eq. (2.17) we find that the corresponding discrete window function $W_D(f)$ is bandlimited to intervals $\ldots, [-b-2f_c, b-2f_c], [-b, b], [-b+2f_c, b+2f_c], \ldots$, and so on. Now let's look at the discrete windowed spectrum at some frequency f_1 within b of the band edge f_c. We find from Eq. (2.16) that $S_D(mT, f_1)$ involves $X(f)$ not only at frequencies in the interval $[f_1 - bf, f_c]$ as in the continuous-signal case, but also at frequencies $2f_c$ lower than this in the interval $[-f_c, -2f_c + f_1 + b]$. It is as if components in this lower interval were aliased up to the interval $[f_c, f_1 + b]$ where the sampled signal should have no power. Because the signal is real, its spectrum is hermitian and these aliased components are just the conjugates of $X(f)$ from the

interval $[f_1 - b, f_c]$. Instead of being symmetrical, the passband of a filter near the sampling cutoff is folded back on the conjugate of itself about the frequency f_c.

In practice aliasing near f_c is not usually a problem. Useful window functions are narrowband compared to the input signal (else we'd only synthesize a few, very broad analysis filters); and we have to throw away spectral points close to the sampling limit f_c anyway because of the finite transition bandwidth of the sampling lowpass filter as discussed in Sec. 2.4. However, a similar problem occurs in *both* the continuous and discrete worlds for filters near zero frequency. The discrete windowed spectrum $S_D(mT, f)$ evaluated at a low frequency f_1 within the lowpass bandwidth b will depend not just on signal components at positive frequencies $[0, f_1 + b]$ but also at negative frequencies $[f_1 - b, 0]$, the conjugates of those components at $[0, b - f_1]$. The analog circuit designer faces this problem when trying to design a bandpass filter with arithmetic symmetry when the bandwidth is a large percentage of (or even larger than) the center frequency. We will refer to these effects, at both the high and low ends of the analyzed spectrum, as *passband aliasing*.

Finally we are led to the primary tool of digital spectrum analysis, the *order-N discrete Fourier transform* or DFT defined as [*e.g., Rabiner and Gold,* 1975, Sec. 2.21]:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi nk/N} \qquad \text{for } k = 0, 1, \ldots, N-1. \tag{2.18}$$

We can think of this as $S_D(0, k/NT)$, the discrete windowed spectrum evaluated at time $t_0 = mT = 0$ and discrete frequencies $f = k/NT$, using a uniform weighting sequence $\{w_n\}$ of length N, where

$$w_n = \begin{cases} 1, & \text{if } n = 0, 1, \ldots, N-1, \\ 0, & \text{otherwise.} \end{cases} \tag{2.19}$$

In practice we will use more sophisticated weighting functions than Eq. (2.19), but we postpone their discussion until Sec. 2.5.3. We have also left out any explicit segment time origin t_0 in Eq. (2.18). At time mT we actually evaluate the DFT of the sequence $\{x_{n+m}\}$ to find $X_k = S_D(mT, k/NT)$.

When the DFT is calculated, each segment $\{x_n\}$ of N real data points generates a transform sequence $\{X_k\}$ of complex spectral points at frequency multiples of $f_D = 1/NT$. Note that X_k is periodic with period N, that is, $X_{k+N} = X_k$ so we only need to calculate spectral points for $0 \le k \le N-1$. In fact, since $\{x_n\}$ is a real sequence, the transform points $X_{N/2+1}, \ldots, X_{N-1}$ are just the conjugates of the points $X_{N/2-1}, \ldots, X_1$ and are not needed. (The spectral point $X_{N/2}$ is a special case and is independent of those below, but we don't need it either.) If the transform size N is adjustable we can change the segment length NT and spectral line spacing f_D, and thus trade off frequency resolution for time resolution as desired. In the present system, N must be a power of 2 from 64 to 2048, and we can generate spectra with from 32 to 1024 synthesized filters.

For example, using the maximum transform size $N = 2048$, at 25600 samples/second we will transform a segment 80 ms long to get a spectrum with 1024 points spaced every 12.5 Hz from 0 to almost 12.8 kHz. As mentioned above, the input anti-aliasing filter allows some frequency aliasing above 10.6 kHz. We will throw away spectral points above that frequency, and only use points $X_0, \ldots, X_{848}$, which cover 0 to 10.6 kHz.

Data Analysis Procedure. Figure 2.5 shows schematically the various steps that are taken when analyzing a signal. We start out with a long string of data samples representing up to 400 seconds of a 10.6 kHz bandwidth analog signal. The data analysis program runs in an endless loop, processing a small windowed segment of data samples at each pass and then moving ahead to the next segment

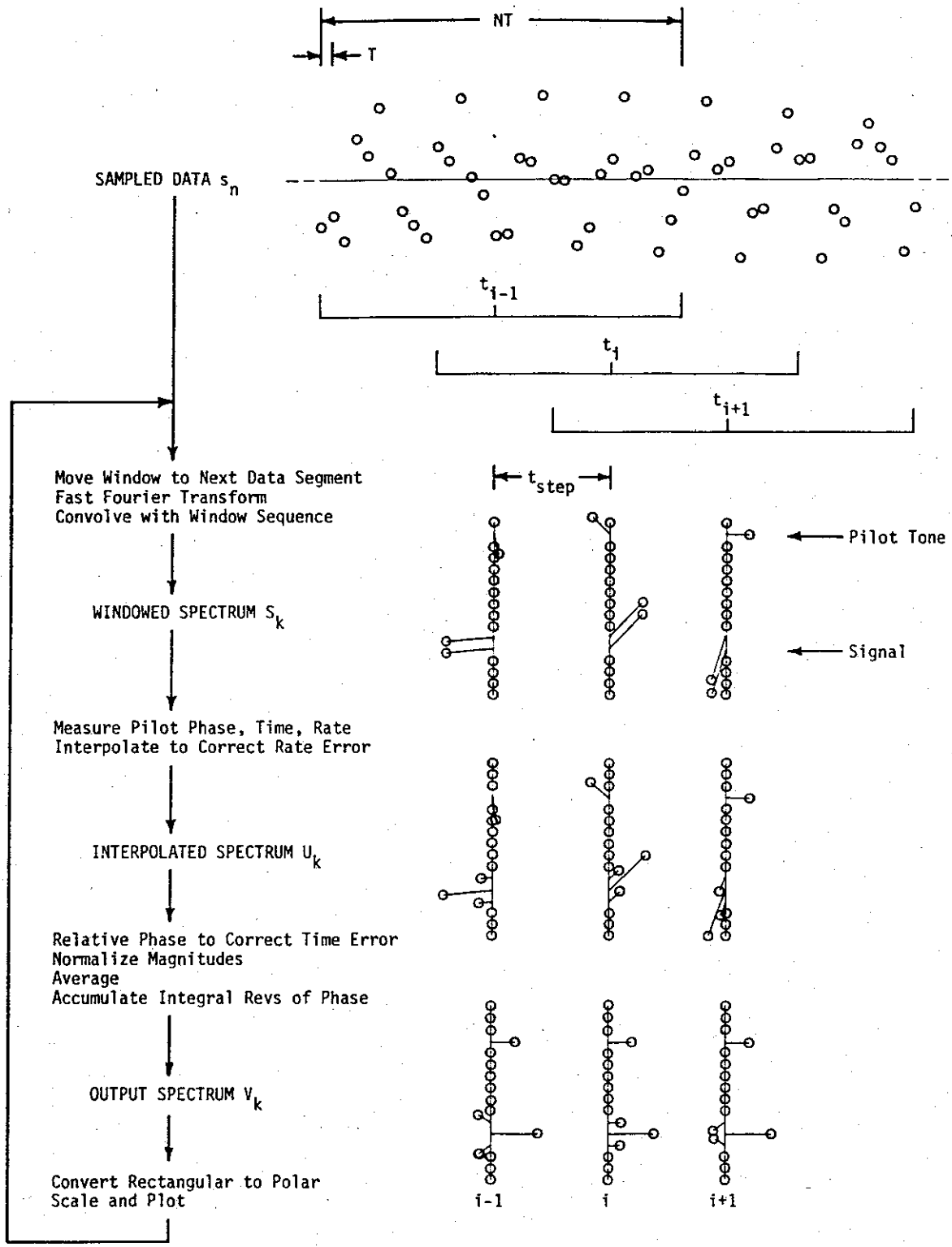

Figure 2.5. Data analysis procedure. Successive overlapping segments of input samples are transformed via the FFT. The pilot tone phase and frequency are measured and used to calculate the actual data time and relative data rate. Spectral points are interpolated in frequency to correct for rate error, and a reference phase is subtracted from each filter output. The resulting magnitude and relative phase values are then plotted.

TABLE 2.3

Program Variables and Parameters

$\{s_n\}$, where $s_n = s(nT)$, up to 10^7 real samples of the input waveform $s(t)$
T = nominal time between data samples = 1/25600 s for 10.6 kHz data
N = number of data points in a data segment to be analyzed, 64, 128, ..., 2048
$\{x_n\}$, a sequence of N samples from $\{s_n\}$, the current segment of $s(t)$ to be analyzed
$\{X_k\}$, $N/2$ complex points in the DFT spectrum of $\{x_n\}$
$\{S_k\}$, complex points in the windowed spectrum of $\{x_n\}$
$\{U_k\}$, complex points in the interpolated spectrum of $\{x_n\}$
$\{V_k\}$, complex points in the output spectrum of $\{x_n\}$
$\{\Phi_k\}$, integral revolutions of accumulated phase in the output spectrum
$f_D = 1/NT$, frequency spacing of filters in the DFT and windowed spectra
f_I = frequency spacing of filters in the interpolated and output spectra
f_{lo} and f_{hi}, frequency bounds of the interpolated and output spectra
j_{scale} = number of halving steps for overflow prevention in the FFT
t_{step} = nominal advance in data time between successive data segments
j_{step} = actual number of samples advanced from the previous data segment to the current one
$\hat{t}_i$ = estimated time of the center sample ($x_{N/2}$) of the i-th data segment
t_i = measured center time of the i-th data segment
f_p = pilot tone frequency as recorded
f_{pi} = smoothed pilot tone as measured at the i-th data segment
$\bar{r}_i = \bar{f}_{pi}/f_p$ = smoothed relative data rate at the i-th data segment
G_p = pilot tone gain
τ_p = pilot tone tracker smoothing time constant
G_{out} = output spectrum gain
τ_{avg} = output spectrum averaging time constant
M_{span} = magnitude plot range in dB (log magnitude plots only)
P_{span} = phase plot range in revs or fraction of a rev

until the entire string is processed. At each pass the spectrum for that data segment may be plotted. In the following sections we will discuss each of the processing steps in turn. Table 2.3 lists the more important program variables that are mentioned in the discussion of the various analysis algorithms.

2.5.1 Move Window to Next Data Segment

The first step in each pass through the analysis program is to move the data window forward a time t_{step} to the next data segment to be analyzed. The usual procedure in moving-window analysis is to advance the data window by the same, fixed number of sample points at each pass. This has two disadvantages. First, this means we must always advance by a time which is an integral multiple of the sample time T, and this limits the choice of spectrum output rates. We may want to make plots to time scales unrelated to the sampling rate and will need to generate spectra at arbitrary intervals. Second, if we are dealing with tape-recorded data and possible timing errors, we may not know exactly what the time scale is at any particular point, and it may change as we move through the data.

What we really want to do is calculate output spectra at equal increments in data time. If t_{start} is the time of the first data segment analyzed, then we want successive output spectra at times $t_i = t_{start} + it_{step}$, for $i = 0, 1, \ldots$. Since the data rate r may change, an increment of time t_{step} will not always represent the same number of samples. And it won't necessarily represent an integral number of samples in any case. What we will do is measure the data time and calculate the instantaneous data rate as we go, and try to advance in even increments t_{step}. Of course, while we can advance our clock t_i by some arbitrary amount we are constrained to move the data window forward only by integral numbers of samples. We will move it to the nearest sample.

The purist will note that by advancing the window only by integral samples we can not generate spectra at *precisely* uniform increments of time $t_i = t_{start} + it_{step}$, even though that is how we shall plot them. However, the error in time between the clock and the data window need not be more than half the distance between samples or $0.5\bar{r}T$, which is about 20 μs for 10.6 kHz data. This error will not affect the plotted phases of analyzed signals since relative phases are calculated with respect to the measured data time t_i, and not the nominal time $t_{start} + it_{step}$.

We proceed as follows on the *i-th* pass:

1. We desire the time of the center of the data segment on the *i-th* pass to be $t_{start} + it_{step}$. This represents an advance from the previous segment by a time of $it_{step} - t_{i-1}$, where t_{i-1} is the time of the previous segment as determined from the pilot tone phase using Eq. (2.2). Each sample in the data at this point represents an increment in data time of approximately $\bar{r}_{i-1}T$. (We have to use the data rate $\bar{r}_{i-1}$ measured from the previous segment, since we have not yet calculated $\bar{r}_i$.) Therefore, we will move our window ahead by $j_{step} = (it_{step} - t_{i-1})/(\bar{r}_{i-1}T) + 0.5$ points, where j_{step} is an integer, and the term $+0.5$ rounds j_{step} to the nearest whole number.
2. If $j_{step} < 0$ then use $j_{step} = 0$. This is an exceptional case that may occur on the first few passes only, but it is important not to try to step backward through the data, especially at the start. On the initial pass we assumed a data time $t_0 = t_{start}$, but we may actually have measured a time t_0 that was greater than $t_{start} + t_{step}$ by a small amount, depending on the initial phase of the pilot tone and the size of t_{step}. This would cause us to move backward to get to time $t_1 = t_{start} + 1 \cdot t_{step}$. This exceptional condition will disappear shortly.
3. If the index in $\{s_n\}$ of the first sample of the previous data segment $\{x_n\}$ was some integer, say a (*i.e.*, $x_0 = s_a$), then assign $a = a + j_{step}$ and assign $x_n = s_{a+n}$, for $n = 0, 1, \ldots, N-1$ to get the segment to be analyzed on this pass. Note that the integer a here is fictitious. The entire sequence $\{s_n\}$ is not stored in memory as an array, and the actual assignment of s_{a+n} to x_n involves reading in data records from external files and so on.
4. Having moved ahead j_{step} samples we estimate the current segment time as $\hat{t}_i = t_{i-1} + j_{step}\bar{r}_{i-1}T$. The segment time will be refined when the pilot tone phase is measured, but it

is important for the pilot tone tracking routine that we have a good estimate in order to avoid skipping whole cycles of pilot tone.

Each data segment contains N sampled data points $x_0, x_1, \ldots, x_{N-1}$, though we will think of the segment as $N+1$ points with the right end point omitted so that it is symmetrical about the sample $x_{N/2}$. The data time of a segment and the phases of its signal components will be calculated with respect to this center point. If the sample time is T (say, 1/25600 second for 10.6 kHz data) then a segment represents NT seconds of input data. In this system the number N must be a power of 2 from 64 to 2048, so the length of the data segment NT will be from 2.5 to 80 ms.

The data segment step time t_{step} determines how fast we advance through the data. In this system it can be set arbitrarily over the range $0.01NT \leq t_{step} \leq 10NT$. The segment step time is usually chosen to give some particular time scale to the output plot. It also controls the redundancy and smoothness of the output plot. If $t_{step} > NT$, then some input data points will be skipped as the analysis window is moved ahead, and we will obviously lose some signal information. However, if t_{step} is small compared to NT, then we will transform many of the same data points over again and cannot expect that the output spectra will show much change from one segment to the next. We will return to the choice of t_{step} *vs.* spectrum redundancy when we talk about windows.

2.5.2 Fast Fourier Transform

After moving ahead in the input data to the next segment $\{x_n\}$ to be analyzed, we want to calculate its discrete Fourier transform as given by Eq. (2.18). We need to calculate $N/2$ components $X_0, \ldots, X_{N/2-1}$, each of which involve the summation of N terms. Each term involves the product of a real number x_n by a complex number $\exp(-j2\pi nk/N) = \cos(2\pi nk/N) - j\sin(2\pi nk/N)$. This will take two actual multiplications, one for the real component and one for the imaginary one. We are looking at making $N \cdot N$ multiplications, as well as a like number of additions and table look-ups. On the Eclipse S/230, an integer multiplication is a time-consuming operation taking 7.2 μs. For $N = 2048$, $N \cdot N$ comes to 4.19×10^7 multiplications which will take over 30 seconds, not counting the other operations. And this just to process 80 ms worth of data. Data analysis will be even slower than with a Sona-Graph.

Fortunately there is a way out. The *fast Fourier transform* or FFT algorithm is a procedure to calculate an order-N DFT in $N \cdot \log_2 N$ operations. For $N = 2048$ this involves $2048 \cdot 11 = 22528$ multiplications, which will take only 162 ms. The algorithm, invented by *Cooley and Tukey* [1965], involves factoring Eq. (2.18) using the fact that $\exp(j2\pi nk/N)$ is periodic in $n \cdot k$ with period N. The reduction in the number of operations needed comes about because these various factors are common to different X_k's, and the calculation of the entire array $\{X_k\}$ occurs in parallel. See *Cooley et al.* [1967] for a brief history of this technique. The FFT is well covered in the literature, for example by *Gold and Rader* [1969], *Rabiner and Rader* [1972], and *Rabiner and Gold* [1975], and it is unnecessary to describe the algorithm itself in much detail here. However, there are a few details regarding its implementation when using integer arithmetic and processing real data that do need to be mentioned.

Overflow Detection and Handling in the FFT. The FFT algorithm takes an array of N complex data points and transforms it in $\log_2 N$ stages. At each stage the array is processed in $N/2$ elemental steps, where data points are operated on in pairs and returned to the array where they remain until the next stage. At the end of the last stage the array which was originally complex data has now become the complex transform of that data. (There is another step which involves reordering the array points but it needn't concern us here.)

The basic numerical operation that manipulates each pair of data points is called a *butterfly*. The butterfly takes two complex values from the array at one stage, say X_j and X_k, combines them with a complex phase or *twiddle factor* of the form $\exp(j2\pi n/N)$, and generates two values X'_j and X'_k which are stored back in their respective places in the array. These new values will be combined (with different partners) by butterfly operations in the next stage of the algorithm, until the final stage is reached and the transform is complete. There are two kinds of butterfly, *decimation-in-time* and *decimation-in-frequency*, which are used in two slightly different varieties of the FFT algorithm, involving different factorings of Eq. (2.18). However, they are similar as far as arithmetic overflow is concerned. The decimation-in-time butterfly is as follows:

$$\begin{aligned} X'_k &= X_k + X_l e^{j2\pi n/N} \\ X'_l &= X_k - X_l e^{j2\pi n/N}. \end{aligned} \tag{2.20}$$

If we represent the real and imaginary parts of the complex number X by R and I, then the actual arithmetic operations in the butterfly are

$$\begin{aligned} R'_k &= R_k + R_l \cos(2\pi n/N) + I_l \sin(2\pi n/N) \\ I'_k &= I_k \ - R_l \sin(2\pi n/N) + I_l \cos(2\pi n/N) \\ R'_l &= R_k - R_l \cos(2\pi n/N) - I_l \sin(2\pi n/N) \\ I'_l &= I_k \ + R_l \sin(2\pi n/N) - I_l \cos(2\pi n/N). \end{aligned} \tag{2.21}$$

Depending on the phases of the various complex numbers involved, it is possible for the magnitude of one or both of the X's to double as they are transformed into X''s. Since each datum will go through $\log_2 N$ butterflies, its magnitude could conceivably double that many times, to $2^{\log_2 N} = N$ times its original size before the FFT is completed. In fact, this kind of behavior will (almost) occur if the original waveform is a sinewave tone at a frequency, say f_m, corresponding to the center frequency of one of the synthesized filters. At each stage in the FFT the signal becomes concentrated in fewer and fewer elements of the array until at the end everything resides in the two array elements corresponding to $\pm f_m$. This follows from the fact that the order-N DFT of $x_n = A\cos(2\pi mn/N)$ (a signal at $f_m = m/NT$ Hz) is $AN/2$ at spectral points X_{N-m} and X_m, and zero everywhere else. However, equation (2.20) shows that even with random data the magnitudes of array elements will tend to increase.

The FFT routine in this system uses 16-bit integer arithmetic. All numbers in the signal/spectrum array are constrained to take on integral values between -32768 and $+32767$. Because of the potential doubling in magnitude that can occur we may generate a value at some stage that is less than -32768 or greater than $+32767$, and cannot be represented with only 16 bits. Such a value is said to overflow. When this happens we have an error of unknown size that can propagate to additional array points in each remaining stage of the FFT. To avoid arithmetic overflow we will have to scale back data values before they become too large.

On the other hand, to preserve the maximum dynamic range in our signal we want to keep array values as large as possible so the relative contribution of roundoff noise will be small. The sines and cosines in Eq. (2.21) are actually 16-bit integers of the form $32768\sin(\cdot)$, and the multiplication of an array component by a sine or cosine is actually the multiplication of two 16-bit integers followed by a right shift of 15 bits (an implicit division by 32768). There is a potential error here as low-order bits are discarded which we will minimize by rounding rather than just truncating the product. These rounding errors tend to contribute a fixed amount of random noise to the spectrum at each stage.

We have two constraints here. We need to scale intermediate results to prevent overflow, yet keep them as large as possible to minimize processing noise. *Welch* [1969] has a good discussion of overflow and roundoff noise, and presents some scaling strategies. *Rabiner and Gold* [1975, Sec. 10.5] also review these problems. The approach taken in this system is to scale intermediate results by halving all array values on the last few stages of the FFT. We define an integer, j_{scale}, that specifies how many times halving will be performed. Scaling operates as follows:

1. The FFT routine accepts a data sequence $\{x_n\}$ in one array, and returns the spectrum $\{X_k\}$ in a second array. The original array is not destroyed by the FFT call. Intermediate results are scaled by halving on the final j_{scale} stages of the FFT. The FFT routine returns an overflow flag which is set if an arithmetic overflow occurred.
2. If an overflow occurred, j_{scale} is incremented and we go back to step 1 and call the FFT routine again. There will be one more stage of halving this time. This process is repeated as long as an overflow is detected.
3. If no overflow occurred, the spectrum is okay and program operation continues. Magnitude data in this particular spectrum will later be multiplied by a factor $2^{j_{scale}}$ to correct for scaling in the FFT before they are used.
4. Periodically (every 4096 samples) we decrement j_{scale} and try to use fewer halving stages in the FFT. This keeps one data segment with large spectral values from suppressing the dynamic range of everything that follows.

This scheme works fairly well. Processing noise with test signals is usually more than 75 dB below peak magnitudes in the spectrum, which is sufficient for our work. However, this scaling algorithm is probably not optimum. For instance, we require halving at all stages past the first where an overflow occurs. This much scaling might not be needed. A better scheme from the standpoint of dynamic range would be to require halving only in stages where an overflow is actually detected, perhaps associating a bit set in a control word with halving at a particular stage of the FFT. The problem with this is adapting the halving prescription to changing signal characteristics without performing an excessive number of attempts on any given data sequence. (I have also tried halving in the *first* j_{scale} stages. Results were somewhat worse, as might be expected.)

Another approach is to require halving when an overflow occurs but to perform it in that stage of the FFT without repeating the complete transform, perhaps by backing up and halving all values already processed in that stage as well as any remaining ones, or perhaps by saving the results of the previous stage and just beginning the current stage over again but with halving. This is an attractive approach but requires a fair amount of additional bookkeeping.

A final solution is to use floating-point instead of fixed-point (integer) arithmetic. The dynamic range of floating-point numbers is much larger and overflow is not possible. This was not an attractive approach on the Eclipse S/230 computer. Though the floating-point processor is quite fast, loading values into its registers takes longer than with integers. And floating-point numbers take at least twice as much memory storage as integers, something in short supply on this particular machine. On a modern machine floating-point arithmetic might be a good choice.

Calculating the FFT of Real Data. The FFT, as it is usually programmed, takes a sequence $\{x_n\}$ of N complex data points, where N is a power of 2, and generates a transform sequence $\{X_k\}$ of N complex spectral points. Each complex datum $x = a + jb$ is represented as a pair of numbers (a, b), the real and imaginary parts of x. The sequence $\{x_n\}$ might actually be stored as an array $(a_0, b_0, a_1, b_1, \ldots, a_{N-1}, b_{N-1})$, where real and imaginary parts alternate, or perhaps as two arrays

containing the a's and b's separately. Similarly, the sequence $\{X_k\}$ might be returned as an array $(A_0, B_0, A_1, \ldots)$.

Since the waveform data we are dealing with are a sequence of real numbers, we can transform them with a standard FFT routine only if we pass them as an array of complex numbers whose imaginary parts are all zero, that is, as $(x_0, 0, x_1, 0, \ldots, x_{N-1}, 0)$. The output of the routine will be an array of N complex values X_k, for $k = 0, 1, \ldots, N-1$. However, as we saw above, the spectral points $X_{N/2+1}, \ldots, X_{N-1}$ are just the conjugates of $X_{N/2-1}, \ldots, X_1$ and are not needed. Thus we have passed an array half full of zeroes and received back an array the upper half of which is redundant. There must be a more efficient way to process real data.

In fact there is. The following procedure, described by *Cooley et al.* [1970], cuts our work almost exactly in half. Let us arrange the sequence $\{x_n\}$ as an array of close-packed real points, $(x_0, x_1, \ldots, x_{N-1})$, and pass it to the FFT routine which will interpret it as a sequence $\{z_n\}$ of $N/2$ complex data points. That is,

$$z_n = x_{2n} + jx_{2n+1} \qquad \text{for } n = 0, 1, \ldots, N/2 - 1. \tag{2.22}$$

The FFT routine will return as a transform an array of $N/2$ complex spectral points $\{Z_k\}$. We can retrieve the first $N/2$ points of the transform sequence $\{X_k\}$ as follows:

$$\begin{aligned} X_k &= \frac{1}{2}\left[Z_k + Z^*_{N/2-k} - j(Z_k - Z^*_{N/2-k})e^{-j2\pi k/N}\right] \\ X_{N/2-k} &= \frac{1}{2}\left[Z^*_k + Z_{N/2-k} - j(Z^*_k - Z_{N/2-k})e^{+j2\pi k/N}\right] \end{aligned} \tag{2.23}$$

for $k = 0, 1, \ldots, N/4$.

The actual operations performed after the FFT routine to extract the transform of the input close-packed real data are performed in-place on pairs of complex elements in the returned array as follows:

$$\begin{aligned} R'_k &= R_k + R_l - (R_k - R_l)\sin(2\pi k/N) + (I_k + I_l)\cos(2\pi k/N) \\ I'_k &= I_k - I_l - (R_k - R_l)\cos(2\pi k/N) - (I_k + I_l)\sin(2\pi k/N) \\ R'_l &= R_k + R_l + (R_k - R_l)\sin(2\pi k/N) - (I_k + I_l)\cos(2\pi k/N) \\ I'_l &= -I_k + I_l - (R_k - R_l)\cos(2\pi k/N) - (I_k + I_l)\sin(2\pi k/N). \end{aligned} \tag{2.24}$$

for $k = 1, 2, \ldots, N/4 - 1$ and $l = N/2 - k$. We also have the special cases

$$\begin{aligned} R'_0 &= 2(R_0 + I_0) \\ I'_0 &= 0 \\ R'_{N/4} &= 2R_{N/4} \\ I'_{N/4} &= -2I_{N/4}. \end{aligned}$$

We could also calculate the spectral point $X_{N/2} = 2(R_0 - I_0)$, which is always real (like X_0). However, it is not needed.

Note that the operations involving $X_1, \ldots, X_{N/4-1}$ and $X_{N/4+1}, \ldots, X_{N/2-1}$ are very much like a series of FFT butterflies. The extraction process is actually performed in two passes. On the first pass the sums and differences of each *k-th* and *l-th* pair are calculated and replace the original values, and on the second pass butterflies using these sums and differences are performed.

The operations shown here differ from Eq. (2.23) in that we have left out the factors 1/2. We can have arithmetic overflow in these operations, so the extraction routine incorporates provision for overflow detection and data halving just like the basic FFT routine.

TABLE 2.4

FFT Execution Time *vs.* Transform Size, for N Close-Packed Real Data Points

Number of Real Points N	64	128	256	512	1024	2048
FFT ($N/2$ Complex Points)	2.2 ms	5.5 ms	13.3 ms	31 ms	71 ms	160 ms
Extract Real Transform, add	0.7	1.7	3.2	7	12	27
Window Convolution, add						
1*st* Order	0	0	0	0	0	0
2*nd* Order	2.8	5.5	11	21	42	84
3*rd* Order	3.9	7.6	15	30	60	119
4*th* Order	4.8	9.5	19	37	75	149

FFT Execution Time. While I don't mean to go into great detail, programmers may be interested in the following facts. The FFT routine, including extraction of the close-packed real data transform, has been written in Eclipse S/230 micro-code for maximum processing speed. Decimation-in-time is used because its butterfly is faster given the Eclipse accumulator set and arithmetic-logic unit architecture than is the decimation-in-frequency butterfly. Values of $\cos(\theta)$ and $\sin(\theta)$ needed by the FFT and the real-data transform extraction routine are precalculated and stored in an array of alternating cosine/sine entries for $0° \leq \theta \leq 45°$, in increments of 360/2048 degree. Function values needed for angles in the range $45° < \theta < 180°$ are also read from this table with some simple address and sign manipulation. Each of the 514 values in the table is a 16-bit integer from 0 to 32767 representing a cosine or sine value from 0 to 1.00000^{-}. That is, the values stored are $32768\cos(\theta)$ and $32768\sin(\theta)$. The butterflies for angles $\theta = 0°$, $90°$, $45°$, and $135°$ have been written separately and do not refer to the cosine/sine array. The simplifications permitted at these particular angles speeds things up quite a bit since these butterflies are used $N/2$, $N/4$, $N/8$, and $N/8$ times each, respectively.

Table 2.4 shows the measured execution times needed to process real data sequences of various lengths N. We can see the increase in performance that results from using close-packed real data, even though we have to follow the FFT by an additional step to extract the real-data transform. For example, a 1024-point real sequence takes $71 + 12$ or 83 ms to transform, whereas a 1024-point complex sequence (even with all zero imaginary components) would take 160 ms, or just about twice as long.

Table 2.4 also includes the time necessary to convolve the spectrum sequence with a window sequence, for various choices of window order. This process is discussed in detail in the next section. The transformation of 2048 real data points (80 ms of 10.6 kHz data), including 3*rd*-order windowing, takes $160 + 27 + 119$ or 306 ms. Including all other operations, spectrum analysis runs at about $1/10th$ real-time speed for 10 kHz data.

2.5.3 Windowing/Weighting to Specify Filter Shape

The issue of windowing is one that is generally treated rather lightly in texts on signal processing. When mentioned, it is often illustrated by windows that are of more historical than practical interest. The reader is referred to *Harris* [1978] for information about a wide variety of different windows and their effects on the amplitude response of DFT filters. *Nuttall* [1981] corrects some mistakes by *Harris* and gives further details, particularly of windows that are easy to implement by convolution in the frequency domain. No one seems to mention the effect of windows on the phase response of the synthesized DFT filters, probably because the subject has not previously been of much interest, so I will present a few of my own observations.

Why the Synthesized DFT Filters Need Improvement. The $N/2$ spectral points of the discrete Fourier transform generated by the FFT procedure above can be thought of as the instantaneous outputs of $N/2$ bandpass filters, with center frequencies spaced uniformly in frequency from 0 Hz up to (almost) $1/2T$ Hz every $f_D = 1/NT$ Hz. However, the shapes of the filter passbands in the frequency domain are not particularly nice. To discriminate between signal components at different frequencies we wish our filters to have flat tops in the passband, steep skirts in the transition regions between the passband and the stopbands, and low transmission in the stopbands.

To see why the DFT filter responses are not so nice, let's look at the response of a given filter to a particular signal. Consider a signal $x(t)$ defined as

$$x(t) = A\cos[\phi_o + 2\pi f_0(t - t_0)]. \tag{2.25}$$

This signal is a single-frequency tone with peak amplitude A, frequency f_0, and at time t_0 has a phase ϕ_0. We will set $t_0 = NT/2$, measure the signal at intervals of time T, and generate a sequence $\{x_n\}$ of N samples as

$$x_n = x(nT) = A\cos[\phi_0 + 2\pi f_0(n - N/2)T], \qquad \text{for } n = 0, 1, \ldots, N-1. \tag{2.26}$$

This sequence represents a segment of NT seconds of signal, centered about the point $x_{N/2}$ (with the right end point omitted). The phase of the signal at the center of the segment (at sample $x_{N/2}$) is ϕ_0. We assume $0 \le f_0 < 1/2T$ so the signal is not under-sampled.

We will evaluate the DFT sequence $\{X_k\}$ at a particular value of k. That is to say, we will examine the spectrum at a particular frequency $f_k = kf_D = k/NT$, the center frequency of the k-*th* synthesized filter. Let us define the signal frequency f_0 in terms of this filter frequency as

$$f_0 = f_k + qf_D = (k+q)f_D = \frac{k+q}{NT}. \tag{2.27}$$

The offset q is the difference in frequency, in units of the DFT filter spacing f_D, of the signal from the center of the k-*th* filter. Appendix A shows that the spectrum point X_k is given by

$$X_k = \frac{AN}{2}\frac{\sin(\pi q)}{N\sin(\pi q/N)}e^{j[\phi_0 - \pi k]}\left[e^{-j\pi q/N} + \frac{\sin(\pi q/N)}{\sin[\pi(2k+q)/N]}e^{j[-2\phi_0 + 2\pi k/N + \pi q/N]}\right]. \tag{2.28}$$

The most important term is the leading one before the brackets [ ]. The bracketed terms are close to 1 in practice and give only small corrections to the phase and amplitude response of the filter. In particular, the right-hand bracketed term can be thought of as the result of passband aliasing as described above in Sec. 2.5. We will have more to say about these correction terms later.

The amplitude response of the DFT filter has (approximately) the form

$$|X_k| = \left| \frac{\sin(\pi q)}{N \sin(\pi q/N)} \right|. \tag{2.29}$$

For N not too small, X_k is approximately $\sin(\pi q)/\pi q \equiv \mathrm{sinc}(q)$. The upper panel ("1*st* Order") of Fig. 2.6 shows this response in dB plotted against offset frequency in units of f_D. The central peak of the filter has a 3-dB width of about $0.89 f_D$ Hz (0.89 bins) which is just fine, but there are many significant sidelobes on either side of this peak. The first sidelobes at $\pm 1.43 f_D$ are down by only 13.26 dB. These high sidelobes are undesirable since they prevent the filter from discriminating sufficiently between a desired signal in its passband and one just outside. (This problem is sometimes called "spectral leakage.")

Improving the Filter Amplitude Response. As shown in Sec. 2.5, the DFT is effectively the discrete windowed spectrum $S_D(0, k/NT)$ evaluated using the uniform weighting sequence given by Eq. (2.19) above as

$$w_n = \begin{cases} 1, & \text{if } n = 0, 1, \ldots, N-1, \\ 0, & \text{otherwise.} \end{cases} \tag{2.19}$$

The frequency response of a DFT filter is determined by the window function $W_D(f)$, given by Eq. (2.17) as the transform of the weighting sequence. For the uniform weighting sequence we find

$$W_D(f) = \sum_{n=-\infty}^{+\infty} w_n e^{-j2\pi f n T} = \frac{\sin(\pi f/f_D)}{\sin(\pi f/N f_D)} e^{-j\pi f(N-1)/N f_D}. \tag{2.30}$$

The connection with the amplitude response is obvious. The uniform weighting sequence in Eq. (2.19) is merely the result of truncating the stream of input samples to get the data segment $\{x_n\}$ for processing. If we could use some other window we could improve the amplitude response of the DFT filters.

In this system we have provided several alternative windows that can be used. They are the minimum-sidelobe windows given by *Nutall* [1981]. These windows all have weighting sequences of the form

$$w_n = \sum_{l=0}^{L-1} a_l \cos[2\pi l(n/N - 1/2)] \qquad \text{for } n = 0, 1, \ldots, N-1, \tag{2.31}$$

and are zero elsewhere. L is a small integer and a particular window is referred to as the *L-th* order window. These weighting sequences are sums of cosine functions of low order, each symmetrical about the point $w_{N/2}$ and performing a whole number (from 0 to $L-1$) of cycles on the interval [0,N]. The coefficients a_l are positive real numbers and are normalized so

$$\sum_{l=0}^{L-1} a_l = 1. \tag{2.32}$$

Note that the 1*st*-order window is just the uniform weighting sequence of Eq. (2.19).

The reason for choosing this particular form of weighting sequence is that these windows lend themselves especially well to application by convolution in the frequency domain. This works as follows. Assume that we have a weighting sequence $\{w_n\}$. The discrete windowed spectrum that results from using this sequence with a given signal is

$$S_k \equiv S_D(0, k/NT) = \sum_{n=0}^{N-1} x_n w_n e^{-j2\pi nk/N}. \tag{2.33}$$

It is easy to show [*cf. Gold and Rader*, 1969, Sec. 6.2] that we can get the same results by the convolution

$$S_k = \frac{1}{N}\sum_{m=0}^{N-1} X_m W_{k-m} \tag{2.34}$$

where

$$W_k = \sum_{n=0}^{N-1} w_n e^{-j2\pi nk/N} \tag{2.35}$$

is the DFT of $\{w_n\}$. That is, we can calculate a windowed spectrum element S_k by first using the FFT to calculate $\{X_k\}$ and then convolving those results with the window sequence $\{W_k/N\}$. Note that this is a circular convolution. W_k is periodic with period N so $W_{-1} = W_{N-1}$, $W_{-2} = W_{N-2}$, and so on.

The windows given by Eq. (2.31) are useful because each window sequence has the very simple form

$$C_k \equiv \frac{W_k}{N} = \begin{cases} a_0 & \text{if } k = 0, \\ \frac{1}{2}a_{|k|}(-1)^k & \text{if } |k| = 1, 2, \ldots, L-1, \\ 0, & \text{otherwise.} \end{cases} \tag{2.36}$$

An order-L window sequence has only $2L-1$ non-zero elements.

We have two possible approaches to windowing. We can apply the window in the time domain by multiplying each sample x_n by the corresponding weighting sequence element w_n. Both are real numbers so this will take N multiplications. Or we can transform $\{x_n\}$ and then convolve each complex spectrum point X_k with the short sequence $\{C_{L-1}, \ldots, C_1, C_0, C_1, \ldots, C_{L-1}\}$. Since each C_k is a real number, the convolution will take $2L$ multiplications per windowed spectrum point (we add $(X_{k-l} + X_{k+l})$ first before multiplying by C_l), or NL multiplications for all $N/2$ points. If fewer than N/L points are needed in the windowed spectrum, convolution will require fewer multiplications (and thus be faster) than weighting the entire data sequence before the FFT. If memory is in short supply the convolution approach may be preferable because we don't need to take up space storing the array $\{w_n\}$.

Windowing Algorithm. In this system we have chosen to do windowing by convolution, primarily because of limitations on available memory. With low-order windows the penalty in execution time is not serious. Windows of order 1, 2, 3, or 4 can be used, with 3 being the most common choice. Processing proceeds as follows:

1. Before convolving we extend the bottom of the spectrum array with the elements $X_{-3} = X_3^*$, $X_{-2} = X_2^*$, and $X_{-1} = X_1^*$ since these may be needed for the first few products. (If we were calculating the top few elements of the windowed spectrum we would have to extend the DFT spectrum at the top in a similar fashion.)
2. We perform the convolution as

$$S_k = C_0 X_k + C_1(X_{k-1} + X_{k+1}) + \cdots + C_{L-1}(X_{k-L+1} + X_{k+L-1}) \tag{2.37}$$

   for $k = 0, 1, \ldots,$ up to the highest spectral line to be used. Convolution with the window sequence is carried out separately on the real and imaginary components of the spectrum since the coefficients C_k are all real. The coefficients for the various windows are shown in Table 2.5.
3. The normalization of Eq. (2.32) means that the sum of the absolute values of all the coefficients in a given window sequence is exactly 1. This eliminates the possibility of arithmetic overflow

TABLE 2.5

Properties of Analysis Windows

Window Order	1*st*	2*nd*	3*rd*	4*th*
Coefficients:				
C_0	1.00	0.53836	0.4243801	0.3635819
C_1		−0.23082	−0.2486703	−0.24458875
C_2			0.03913965	0.06829975
C_3				−0.00532055
Highest Sidelobe	−13 dB	−43 dB	−72 dB	−98 dB
3-dB Width	0.89 f_D	1.30 f_D	1.61 f_D	1.86 f_D
6-dB Width	1.20 f_D	1.81 f_D	2.25 f_D	2.62 f_D
Processing Loss	0.00 dB	5.38 dB	7.44 dB	8.79 dB
Scalloping Loss	3.92 dB	1.75 dB	1.14 dB	0.85 dB
Overlap Correlation:				
$t_{step} = 0.125NT$	87.5%	91.7%	87.2%	83.0%
$t_{step} = 0.25\ NT$	75.0%	70.5%	57.4%	47.0%
$t_{step} = 0.5\ \ NT$	50.0%	23.1%	9.7%	4.2%
$t_{step} = 0.75\ NT$	25.0%	2.6%	0.3%	0.04%

with integer arithmetic. We actually use the floating-point unit for the convolution, but integer results must still be stored back in the spectrum array.

4. After convolution we correct the phases of odd-numbered spectrum points by negating them:

$$S'_k = -S_k, \qquad \text{for } k = 1, 3, 5, \ldots. \tag{2.38}$$

The reason for this is described below when we discuss the DFT phase response. The resulting array $\{S_k\}$ we will call the windowed spectrum.

The choice between the various window orders involves a tradeoff between the amount of sidelobe suppression on the one hand and passband broadening on the other, as shown in Table 2.5. The width of the filter passband at the 3 and 6-dB points is given in units of the filter spacing f_D, or "bins." The usual choice is the 3*rd*-order window, which suppresses sidelobes by 72 dB and gives each synthesized filter a 3-dB width of $1.6f_D$. The *processing loss* in Table 2.5 is merely the value of windowing coefficient C_0 expressed in dB. This is the amount by which the magnitude of a signal component exactly at the center frequency of a DFT filter is decreased by windowing. This decrease is taken into account later when magnitude data are scaled for plotting and it is invisible to the user. The *scalloping loss* is the additional decrease in magnitude for a signal with frequency offset $q = 1/2$; that is, for a signal exactly midway between two DFT filters. This is the maximum error we might make in estimating the magnitude of an unknown signal if we just chose the output of that filter which had the biggest response. Scalloping loss leads to what is known in the trade as the "picket fence effect," the apparent variation in analyzer gain with signal frequency [*e.g., Bergland,* 1969].

The left half of Fig. 2.6 shows the passband shape imparted to the DFT filters by the various windows. These drawings were made by plotting the magnitude of a single filter output as a frequency ramp was analyzed. (A ramp is a synthetic signal whose frequency increases linearly with time.) We see that windowing has two effects. First, it suppresses the level of the sidelobes so only signal components near the center frequency of a particular spectral filter can pass through it. Second,

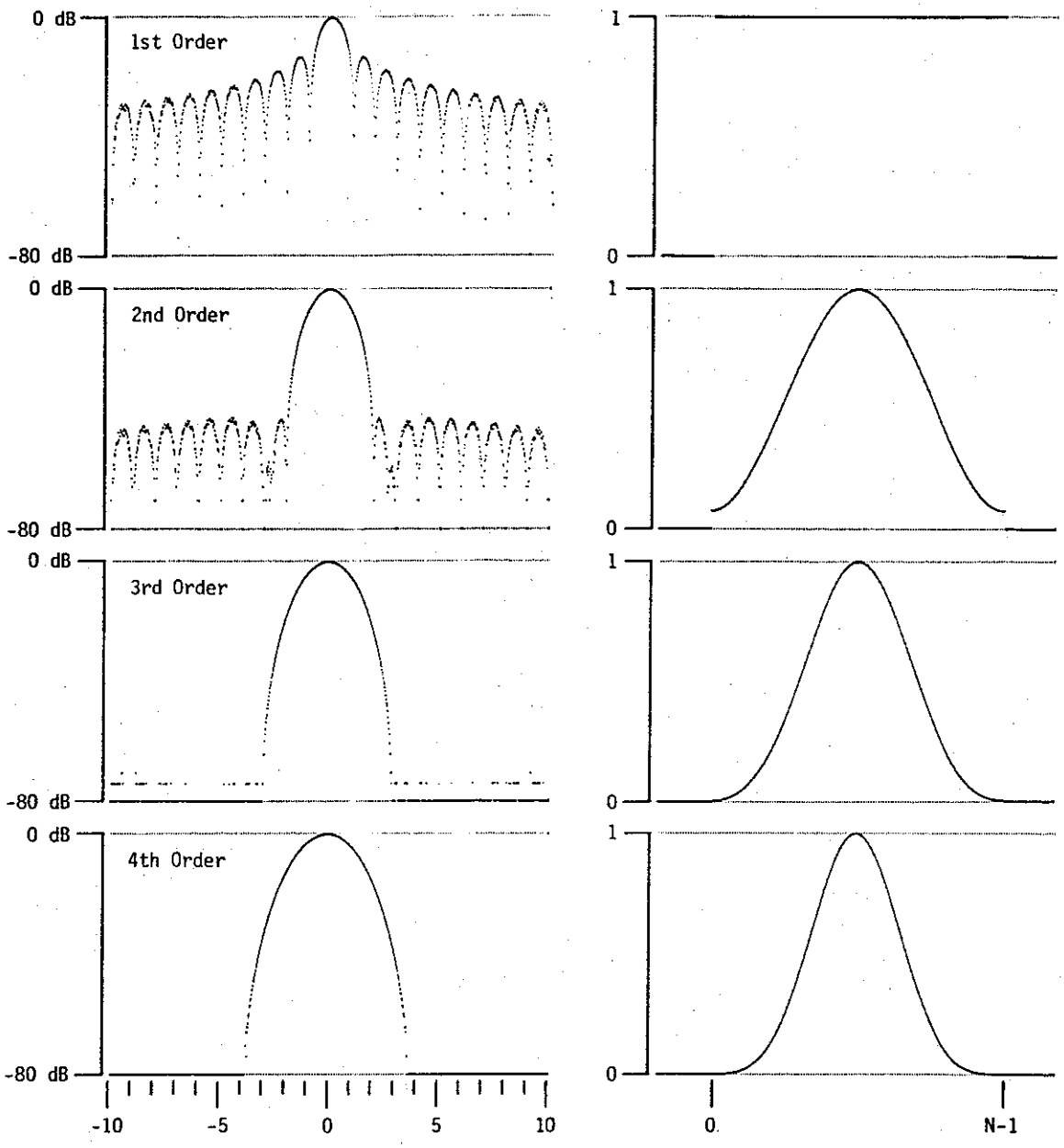

Figure 2.6. DFT filter passband shape and equivalent weighting sequences for different window orders. On the left are plotted the amplitude response in dB of a given DFT filter *versus* signal frequency in units of the filter spacing $f_D = 1/NT$ or "bins." The fuzzy tops to the sidelobes in the $1st$ and $2nd$-order windows are not caused by processing noise but show the filter amplitude response to varying signal phase. See the text for details. On the right are plotted the equivalent weighting sequences in the time domain.

it widens the passband of the filter slightly, flattening the top and giving some overlap between adjacent filters.

There is another interesting feature to note here. The fuzzy tops on the sidelobes in the $1st$ and $2nd$-order windows are not caused by processing noise. (Processing noise can be seen in the quantization of magnitudes to particular values at the bottom of the plots around -72 and -75 dB.) The fuzz is a real effect caused by the changing phase of the analyzed signal. It is the result of the

bracketed correction terms in Eq. (2.28), whose sum depends not only on the frequency of the input signal but (because aliased components are complex-conjugated) on its phase as well. Since the input signal is a ramp, its phase with respect to the analysis window changes with time, and so does the magnitude of the correction terms in Eq. (2.28). We can think of the fuzz as the result of truncating the input signal by the window. Unless the signal frequency f is an exact multiple of $f_D = 1/NT$, a segment of NT seconds of data will not contain an integral number of signal cycles and the power in the segment will depend on the phase of the fractional cycle omitted. This fuzz is one of those details not mentioned in signal-processing texts which can cause the tyro much puzzlement.

The right half of Fig. 2.6 shows the shape of the equivalent weighting sequence in the time domain. These plots were made by moving the data window slowly over an impulse and plotting the magnitude of a single filter output *versus* time. We see that windowing changes the effective weighting of different data points in the data segment, emphasizing points near the center of the segment and reducing the effect of points near the segment ends. (Except for the 1*st*-order window, of course, which uses the uniform sequence.) All of the weighting sequences are discontinuous at the points w_0 and w_{N-1}, dropping abruptly to zero outside, though the effect is not visible on the bottom two plots. This causes the ultimate attenuation of a DFT filter to fall at only 6 dB/octave at frequencies far from the passband. *Nuttall* [1981] lists other windows whose weighting functions are continuous and which have greater ultimate attenuation, but their nearby sidelobes are not as low as the ones we have chosen here.

DFT Phase Response and the Effects of Windowing. Windowing not only gives a good frequency response to each of the synthesized filters but also improves their phase response, a fact not widely known. Without additional windowing (*i.e.*, with a 1*st*-order or rectangular weighting function) the phase of a spectrum point X_k is a complicated function of the input signal phase, frequency, and the filter number k, and equals the signal phase only if the input frequency falls exactly at the center frequency of the synthesized filter.

First of all, let's repeat Eq. (2.28), the DFT response to a signal at frequency $(k+q)f_D$ with segment-center phase ϕ_0:

$$X_k = \frac{AN}{2}\frac{\sin(\pi q)}{N\sin(\pi q/N)}e^{j[\phi_0-\pi k]}\left[e^{-j\pi q/N} + \frac{\sin(\pi q/N)}{\sin[\pi(2k+q)/N]}e^{j[-2\phi_0+2\pi k/N+\pi q/N]}\right]. \qquad (2.28)$$

The phase (in radians) of X_k can be written as

$$\phi_k = \arg(X_k) = \phi_0 + \lfloor|q|\rfloor\pi - k\pi + \epsilon \qquad (2.39)$$

where ϵ is the phase of the bracketed correction terms in Eq. (2.28).

The notation $\lfloor a\rfloor$, read "floor of a," means the largest integer less than or equal to a. The term $\lfloor|q|\rfloor\pi$ accounts for the sign of the $\sin(\pi q)/\sin(\pi q/N)$ factor in Eq. (2.28); the sign of the filter response in Fig. 2.6 alternates with each sidelobe. We will usually be concerned with signals in the main lobe of the filter passband and this term will be zero.

The term $-\pi k$ in Eq. (2.39) means that odd-numbered spectral lines have an additional phase of 180 degrees which must be subtracted when making phase measurements. This arises because we consider the data segment at time t_i to be centered about the point $x_{N/2}$, and measure phase with respect to this point, whereas most FFT implementations assume the segment origin to be the left end point x_0, as given implicitly in the DFT definition in Eq. (2.18). See *Harris* [1978] for a discussion of this problem. We could redesign our FFT procedure but it is easy to correct the phases of odd-numbered lines after windowing by negating both their real and imaginary components as

shown in Eq. (2.38). We could also correct odd-line phases after the FFT but before windowing, if desired, in which case all of the windowing coefficients C_k would have positive values.

For small q (signal frequency close to the filter center frequency) we can expand the bracketed terms in Eq. (2.28) as a power series in $\pi q/N$, and we find that ϵ has a maximum value $\epsilon_{\max}$ given by

$$\epsilon_{\max} = \frac{\pi q}{N}\left[1 + \frac{1}{\sin(2\pi k/N)}\right]. \tag{2.40}$$

That is, the phase ϕ_0 is measured with an error whose maximum value is a linear function of the difference between the signal frequency and the spectral filter frequency. For example, if $N = 64$, $f_D = 1/NT = 400$ Hz (standard 10.6 kHz sampling), and the signal is at 1000 Hz (placing it between the spectral lines X_2 at 800 Hz and X_3 at 1200 Hz), then for X_2 we find $\epsilon_{\max} = 8.61°$, or 24 μs at 1000 Hz. This is not an insignificant error. The problem for us is that this error is proportional to the frequency offset from the center of the filter passband, which will change by different amounts for signals at different frequencies as the rate error changes. This will cause phase errors during analysis similar to those caused by phase distortion in the anti-aliasing filter as discussed in Sec. 2.4.

The error ϵ is due to two causes. Part of it, with maximum value $\pi q/N$, arises because the sequence $x_0, \ldots, x_{N-1}$ is not symmetrical with respect to the point $x_{N/2}$. We are missing the right end point x_N. If we calculate the order $(N+1)$ DFT of the signal in Eq. (2.26) we find this error term is missing. The rest of the error, $\pi q/N \sin(2\pi k/N)$, is due to the aliased response of the filter and is most important at frequencies close to either zero or f_c, *i.e.*, for k near 0 or $N/2$.

Convolving the data transform with the window function tends to cancel the errors between adjacent spectral lines. In the time domain we can think of the equivalent weighting function as diminishing the importance of samples near the ends of the segment and thus decreasing the errors associated with end asymmetry. In the frequency domain windowing reduces the filter sidelobes and thus the aliased responses in those filters close to the edges of the analyzed bandwidth. As an example of the decrease in phase error with windowing I find the following results for $N = 256$, $k = 30$ (the 3 kHz filter with 10.6 kHz data), and $-1 \leq q \leq 1$: With the $1st$-order window ϵ is a linear function of offset q, with maximum value 1.8°. With the $2nd$-order window ϵ is a cubic function of q with a maximum value of 0.10°. With the $3rd$-order window ϵ is a $5th$-order function of q with maximum value roughly 0.02°. With the $4th$-order window ϵ is less than the processing noise. While these particular windows have not been chosen because of their benefits to the DFT filter phase response, their effects are more than satisfactory.

Window Overlap and Output Smoothness. The equivalent weighting function concentrates the effective data near the center of the segment, as shown in Fig. 2.6. Because of this, the correlation between data segments is less than their percentage overlap would suggest. This raises two questions. How much information do we lose by windowing if we don't overlap segments sufficiently, and how does the interpretation of output plots depend on the amount of overlap?

Rather surprisingly, no information is lost by windowing, at least in a formal sense, as long as $t_{step} \leq NT$; that is, as long as we don't skip over any samples as we hop from one data segment to the next. We can always recover the input signal $\{x_n\}$ from the discrete sampled spectrum $\{S_k\}$ given by Eq. (2.33) as follows. First, perform an inverse DFT on $\{S_k\}$ to get the sequence $\{x_n w_n\}$ as

$$x_n w_n = \frac{1}{N}\sum_{k=0}^{N-1} S_k e^{+j2\pi nk/N} \qquad \text{for } n = 0, 1, \ldots, N-1. \tag{2.41}$$

Next, divide each element in $\{x_n w_n\}$ by the corresponding value w_n to recover the sequence $\{x_n\}$.

This works because all the elements in the effective weighting sequences $\{w_n\}$ that we use are non-zero. In fact, *Bastiaans* [1985] shows how an input signal can be recovered from a general sampling of sliding-window spectra in the *f-t* plane, as long as the sampling is carried out on a rectangular lattice whose cells, of size Δf (in Hz) and Δt (in seconds), have area $\Delta f \cdot \Delta t = 1$. If we are examining a spectrogram, where only the magnitudes of filter outputs are plotted, we have indeed lost much of the signal information. The above method does not mean we can recover a windowed signal from a spectrogram—we need phase information as well. In this discussion we have also neglected processing noise, which causes an irreversible loss of information.

Even if no information is lost when windowed data segments just abut (*i.e.*, when $t_{step} = NT$), this doesn't mean that output spectra will be easy to interpret in this case. Weighting does diminish the importance of samples near the ends of the data segment. If a signal transient occurs near a segment end we might well miss it looking at a spectrogram. A good measure of the effect of segment overlap is the overlap correlation coefficient $\gamma(t_{step})$, defined as the normalized autocorrelation function of the weighting function:

$$\gamma(t_{step}) = \int_{-\infty}^{+\infty} w(t)w(t - t_{step})\,dt \Big/ \int_{-\infty}^{+\infty} w(t)^2\,dt \tag{2.42}$$

where $w(t)$ is the continuous-time version of the weighting sequence given by Eq. (2.31) and is zero for t outside the interval $[0, NT]$. This coefficient is the amount of correlation we would expect to see with a random signal between two windowed segments separated by a time t_{step}. Selected values of $\gamma(t_{step})$ are listed in Table 2.5. With the usual choice of the $3rd$-order window, segments that are overlapped by 75% ($t_{step} = NT/4$) are only partially correlated ($\gamma = 57.4\%$). Segments that are overlapped by 50% are essentially uncorrelated ($\gamma = 9.7\%$) and can be regarded as independent. To be sure to show all signal transients on the spectrogram we will want to overlap data segments by at least 50%.

It is very difficult to interpolate by eye between data points from segments which do not overlap sufficiently. A spectrogram made from non-overlapping segments tends to look very harsh and noisy. It is much easier to interpret a spectrogram when the change from pixel to pixel in the plot is not abrupt, in the frequency direction as well as in the time direction. The eye can then separate the smoother underlying signal structure from the higher spatial frequency components due to pixel boundaries. For magnitude and/or phase plots the problem becomes one of interpolating a continuous curve between the discrete points on the plot. I typically find that magnitude-phase plots made with $t_{step} \leq NT/4$ appear reasonably smooth, whereas those with $t_{step} \geq NT/2$ seem too sparse. This subjective finding is in accord with the overlap correlation $\gamma(t_{step})$.

2.5.4 Tracking the Pilot Tone

The pilot tone tracker routine measures the pilot tone magnitude and phase in each transformed data segment. The phase of the pilot tone is compared to a reference phase at frequency f_p, the actual pilot frequency when the tape was recorded. Any difference is used to correct the data segment center time t_i. This time is used when calculating reference phases to correct the tape time error. The phase advance from one segment to the next gives the instantaneous pilot frequency f_{pi}, which is then averaged to form $\bar{f}_{pi}$, the smoothed pilot frequency. $\bar{f}_{pi}$ will be higher or lower than f_p as the analog tape deck was running faster or slower than the deck used in recording the signal. The ratio $\bar{f}_{pi}/f_p$ is the smoothed relative data rate $\bar{r}_i$. This ratio is used when correcting for data frequency shifts (rate error) to generate the interpolated spectrum. The pilot tone magnitude is also displayed at the edge of each data plot to show the occurrence of time ticks and station and time code characters. This time mark trace shows the log magnitude of the pilot tone over a 20 dB range.

Measuring the Data Time from Pilot Phase. The pilot tone tracker routine is called after each data segment has been transformed by the FFT routine and convolved with the appropriate window sequence. The sequence $\{S_k\}$ contains windowed spectral points in (*real, imaginary*) format. The phases of odd-numbered filters have been corrected by Eq. (2.38) so all phases are measured with respect to the center of the data window. The measurement of the pilot tone phase and calculation of the actual data time from this are as follows:

1. First, interpolate between nearest elements of $\{S_k\}$ to find the spectral value S_p for the pilot tone. The rationale for spectrum interpolation is discussed in Sec. 2.5.5. We calculate index j and offset λ as $j = \lfloor \bar{f}_{pi-1}/f_D \rfloor$ and $\lambda = (\bar{f}_{pi-1}/f_D) - j$. We then interpolate $S_p = (1-\lambda)S_j + \lambda S_{j+1}$. ($\lfloor a \rfloor$ or "floor of a" is the largest integer less than or equal to a.) The resulting value S_p is the output of an interpolated filter centered on the frequency $\bar{f}_{pi-1}$ that the pilot tone had in the previous segment.
2. The magnitude $|S_p|$ is calculated, corrected for FFT gain (N), halving, and window processing loss, multiplied by the pilot tone gain factor G_p (a parameter chosen by the operator), and saved for use when plotting the time mark trace at the top of the chart. If the magnitude is below a fixed threshold (10 dB below the smallest magnitude plotted on the time mark trace) we will exit from the tracker routine since presumably there is no pilot tone present. In this case the estimated segment time $\hat{t}_i$ becomes the actual time t_i, and the pilot frequency and data rate retain their previous values.
3. The phase of the pilot tone is calculated as $\phi_p = \arctan(I_p/R_p)/2\pi$ where R_p and I_p are the real and imaginary parts of S_p. The result is the phase in revolutions such that $-1/2 \le \phi_p < 1/2$.
4. We assume that the pilot tone phase was zero at the start of the analysis at time t_{start}. The expected phase (in revolutions) at the the present time is $\hat{\phi}_p = f_p(\hat{t}_i - t_{start})$. This is a large number representing all the cycles of pilot tone that have occurred since time t_{start}. When we measure ϕ_p we cannot know how many whole cycles have already gone by; all we measure is that fractional part of a cycle in progress at the moment, referred to the segment center at sample $x_{N/2}$. When we compare ϕ_p with $\hat{\phi}_p$ to find the time error we ignore any whole cycles of phase. We calculate $\phi_{err} = \phi_p - \hat{\phi}_p$, the difference, and then $\Delta\phi = \phi_{err} - \lfloor \phi_{err} + 0.5 \rfloor$, the difference ignoring whole cycles. Note that $-1/2 \le \Delta\phi < 1/2$. The actual data segment time is now found by adding to the estimated time a correction due to the pilot tone phase error as $t_i = \hat{t}_i + \Delta\phi/f_p$.

The maximum correction to the segment time is limited to $\pm 1/2f_p$ because we only measure the pilot

tone phase to the nearest whole revolution. When tracking the pilot tone in noisy data it is important that the segment step time t_{step} be fairly small compared to the segment length NT. Otherwise, the tracker may skip a whole cycle of pilot tone with resulting timing errors. The allowable step time t_{step} depends on the relative levels of the spherics (the primary noise source) and the pilot tone, and on the variability of the tape speed. I find with typical VLF data that $t_{step} \leq NT$ gives satisfactory tracking performance.

Calculating the Instantaneous Pilot Frequency and Data Rate. Having measured the pilot phase and corrected the data time we can take care of any time error and ensure the accuracy of signal phases when we plot them. Next, we need to measure the rate error so we can correct for frequency shifts and ensure that signal components are properly filtered. The pilot tone frequency is measured as the ratio of the advance in phase from the previous data segment over the advance in sample time as follows:

1. First we check the magnitude of the pilot tone as measured in the previous data segment at t_{i-1}. If below our threshold we will skip the rest of the routine and exit. Presumably the pilot tone just came on and we have no reliable previous value of phase to work with.
2. The pilot tone has advanced $f_p(t_i - t_{i-1})$ revolutions since the previous data segment. Time in the laboratory has advanced by $j_{step}T$ seconds. The pilot tone frequency during this interval was thus $f_{pi} = f_p(t_i - t_{i-1})/j_{step}T$. (If j_{step} is zero, which can happen if t_{step} is very small, we will use the previous value f_{pi-1} because we have analyzed the same segment twice and the pilot frequency cannot have changed.)
3. The instantaneous pilot frequency f_{pi} we have measured will be rather noisy, especially if j_{step} is small. We create a smoothed value by calculating a running average with time constant τ_p as $\bar{f}_{pi} = \alpha \bar{f}_{pi-1} + (1 - \alpha) f_{pi}$ where $\alpha = \exp(-t_{step}/\tau_p)$.
4. The smoothed relative data rate is then calculated as $\bar{r}_i = \bar{f}_{pi}/f_p$.

The tracker time constant τ_p determines how fast the system can respond to a change in data rate. A time constant of 1 s is usually used. The choice of τ_p involves a trade-off between a fast response to tape speed transients and the stability of the tracking loop. If τ_p is large the smoothed data rate will change only slowly, and we may be unable to correct for rapid variations in signal frequency caused by flutter. We saw this as ripple in the pilot tone frequency in Fig. 2.2. On the other hand, if τ_p is too small the tracking routine may lose lock on the pilot tone. This can occur when a noise burst creates a momentarily anomalous value of ϕ_p. If $\bar{r}_i$ is allowed to change very quickly we may find on the next pass that the pilot tone is outside the passband of the interpolated filter S_p. What happens at that point is a matter of chance.

When analyzing data, the coupler "DATA FREQ" display shows the relative data rate $\bar{r}_i$. This can be monitored to check tracker performance. The value for zero tape speed error is 1.000. Analog tapes may show rate errors up to $\pm 2\%$ so the rate display will normally run from 0.980 to 1.020. If the display shows a value below 0.980 or above 1.020 it probably indicates that the tracker has lost lock on the pilot tone.

2.5.5 Interpolation in Frequency to Correct Tape Speed Error

Correcting the Rate Error. When signals are analyzed, signal components are split up into a number of equally-spaced frequency bands, the outputs of a fictitious set of filters. However, if the analog tape deck used to play back the signal when it is digitized is not running at exactly the same speed as the one that originally recorded it, these filtered components will not be at the same frequencies as in the original signal. If the playback deck is running a bit slower than the recording deck, signals reproduced from the analog tape will be a bit lower in frequency than when they were recorded. When the spectrum is calculated we will find that signal components have shifted to lower frequencies, and a given component may even be shifted outside of the passband of the expected filter.

To overcome this problem we must shift the data points in frequency and generate a new spectrum which has the signal components at their original frequencies. The analysis program uses linear interpolation between DFT spectral points to correct the spectrum, interpolating according to the rate error measured from the pilot tone frequency shift. If we have two spectrum points S_k and S_{k+1} found to be at shifted frequencies f_0 and f_1, but we really want a point at some frequency f_a between f_0 and f_1, then we interpolate between S_k and S_{k+1} to find $S_a = [(f_1 - f_a)/(f_1 - f_0)]S_k + [(f_a - f_0)/(f_1 - f_0)]S_{k+1}$. That is, S_a is taken to be a distance between S_k and S_{k+1} in proportion as f_a is located between f_0 and f_1.

Interpolation provides another benefit. It allows us to generate a new spectrum with filters placed at arbitrary frequencies. Filter spacing in the interpolated spectrum does not have to be in increments of f_D as in the DFT spectrum, but can be at some other frequency increment, the interpolated filter spacing f_I. We proceed as follows:

1. Assume we need spectral values at frequencies from f_{lo} to f_{hi}, spaced equally in increments of f_I. These three parameters are selected by the operator. They are related by $f_{hi} = f_{lo} + Jf_I$ where J is an integer.
2. We interpolate between the windowed spectrum points $\{S_k\}$ to get the interpolated spectrum $\{U_j\}$ as follows:

$$U_j = (1 - p)S_k + pS_{k+1} \tag{2.43}$$

where

$$\begin{aligned} j &= 0, 1, \ldots, J, \\ f &= f_{lo} + jf_I = f_{lo}, f_{lo} + f_I, f_{lo} + 2f_I, \ldots, f_{hi}, \\ k &= \lfloor \bar{r}_i f / f_D \rfloor, \\ p &= \bar{r}_i f / f_D - k. \end{aligned}$$

Since the offset p is a real number, the interpolation can be carried out separately on the real and imaginary parts of S_k, as $R'_j = (1 - p)R_k + pR_{k+1}$ and so on.

Using an arbitrary frequency spacing between filters is convenient, particularly when phase measurements are being made. The Siple transmitter often sends signals whose frequencies change in 10 Hz steps. With normal 10.6 kHz data (25.6 kS/s sampling) the minimum value of DFT filter spacing f_D is 12.5 Hz. By interpolating between these points we can generate a new spectrum with filters placed in increments of $f_I = 10$ Hz and more easily measure the relative phases of different transmitter signals. Note, however, that even though we can change the spacing between the filters in the interpolated spectrum, we cannot change their bandwidth, which will remain about $1.6f_D$ for

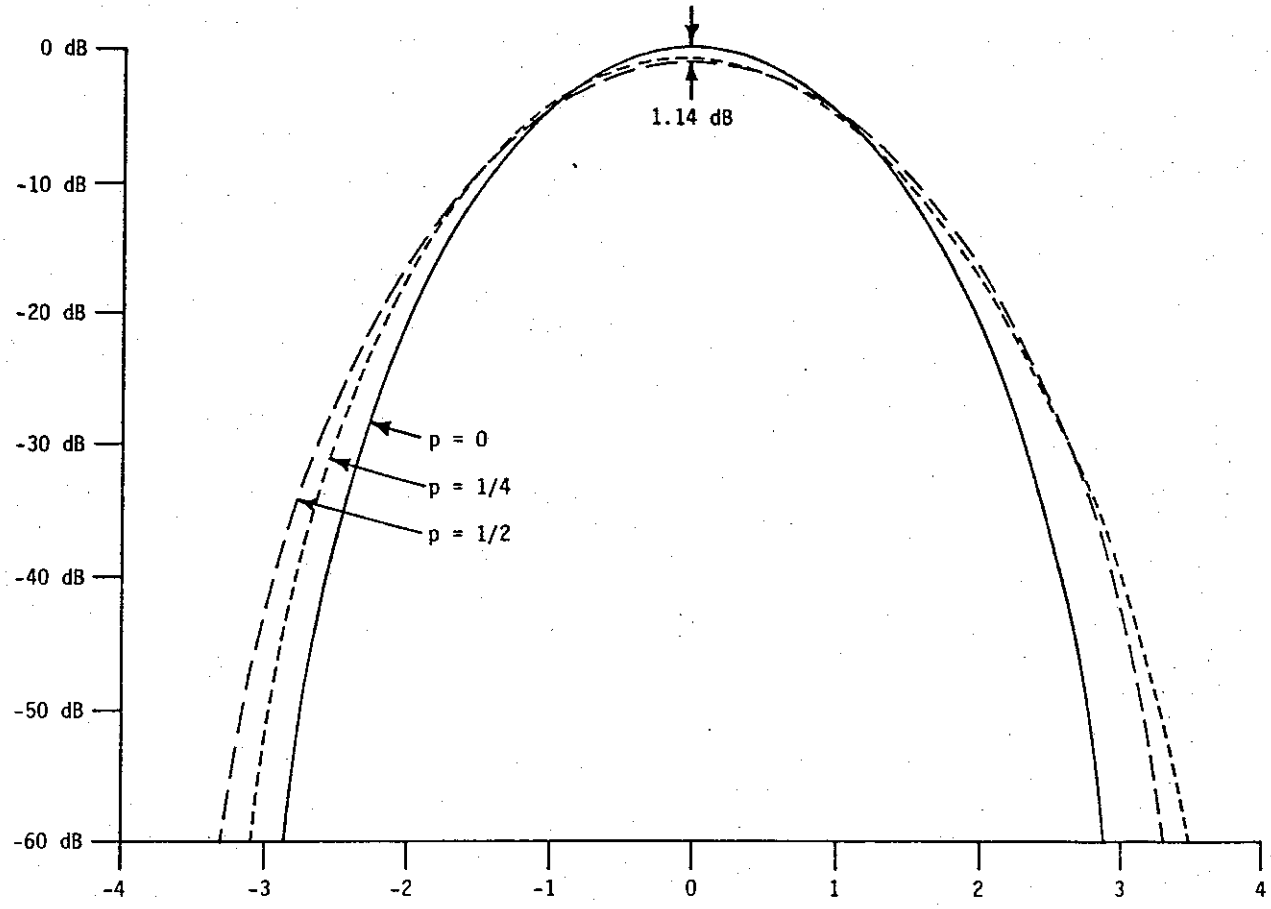

Figure 2.7. Amplitude response of an interpolated filter for various values of interpolation offset p. Frequency across the passband is in units of the DFT filter spacing f_D. The 3*rd*-order window was used. The curve for $p = 0$ is just the 3*rd*-order window passband response from Fig. 2.6. The curve for $p = 1/2$ (interpolation midway between two DFT filters) shows the worst-case change in amplitude response. The curve for $p = 3/4$ would be the same as that for $p = 1/4$ but with the opposite asymmetry. Sidelobes at -72 dB are not shown.

a 3*rd*-order window. Merely placing filters closer together doesn't make them narrower and improve the frequency resolution.

The Effects of Spectral Interpolation. Linear interpolation is a simple method, but it is not immediately clear that it is the best, or even a correct method to use. The test of interpolation is in its effects on the amplitude and phase response of the synthesized filters. The amplitude response of an interpolated filter for various values of interpolation offset p is shown in Fig. 2.7.

If there is sufficient overlap between the passbands of adjacent filters in the windowed spectrum (*i.e.*, if we are using a sufficiently high-order window) then the errors introduced by linear interpolation can be made as small as desired. The worst-case errors introduced by interpolation occur when synthesizing a new filter at a frequency midway between two DFT filters, and, with a 3*rd*-order window function, the change amounts to a 1.14 dB decrease in passband center gain, a negligible change in 3-dB bandwidth, and a slight broadening of the filter skirts. Interpolation has almost no effect on the phase response of a filter.

The decrease in center-frequency gain is unfortunate, but not too serious. Note that the value

1.14 dB is the scalloping loss given in Table 2.4. This decrease is similar in many ways to the picket fence effect encountered with conventional spectrum analyzers when trying to measure the amplitude of a signal that falls between two analysis filters. The difference here is that the decrease in gain depends not only on the signal frequency in comparison to the nearby filter frequencies, but also on the interpolation offsets p for those filters, and thus on the rate error. If we make the pilot tracker time constant τ_p too small in an effort to try and correct for fast changes in data rate, we run the risk of introducing spurious amplitude modulation into the plotted output because of the variation in filter gain with interpolation offset. That is, with τ_p too small, tape recorder flutter becomes amplitude flutter. With the usual choice of $\tau_p = 1$ s, amplitude ripple is not a problem.

The benefits of interpolation and frequency correction outweigh the disadvantages in most cases. If tape speed errors are large, interpolation is absolutely necessary to ensure that the data are filtered properly.

Interpolation Effects versus Window Order. Figure 2.8 illustrates the errors due to interpolation for different choices of window. These plots were made by analyzing a synthetic signal containing two constant-amplitude frequency ramps. One started at 1990 Hz and increased linearly with time at 40 Hz/s. The second one was exactly 4.5 times the frequency of the first, starting at 8955 Hz and increasing at 180 Hz/s. The two ramps were phase-locked, with the phase of the upper one advancing precisely 4.5 times as fast as the lower one.

The signal was analyzed using the upper ramp as a phase-reference pilot tone. The tracker time constant τ_p was set to only 1 ms to allow the smoothed data rate $\bar{r}_i$ to quickly follow the increase in signal frequency. The data segment size was $N = 256$, giving a filter spacing $f_D = 100$ Hz. The curves plotted are the output of an interpolated filter at 2000 Hz. Because the pilot tone was rising in step with the 2000 Hz ramp, interpolation caused this filter to follow the ramp. (The program, of course, thought that the data rate was changing and samples were becoming more and more compressed, and tried to compensate for it. It assumed that the actual signal frequency was constant at 2000 Hz.) The interpolated filter was always centered on the instantaneous frequency of the ramp. What we see, then, is the change in amplitude and phase response with interpolation offset.

Near the beginning of each plot, when the input frequency was 2.0 kHz, we see the output of the windowed filter S_{20} by itself. The interpolation offset p here is zero. A bit later, when the input frequency was 2.05 kHz, the offset p is 0.5 and we are interpolating midway between S_{20} and S_{21}. The gain of the interpolated output is a minimum now, lower than the original output by the scalloping loss. By the middle of the plot the frequency has increased to 2.1 kHz and we again have $p = 0$, but now we are looking at filter S_{21}. This process continues and we begin to interpolate between S_{21} and S_{22} and so on.

The phase plots show the effects of interpolation on the phase response of the system. These plots are only 0.01 rev (3.6° or 5 μs at 2 kHz) full-scale, a much smaller scale than we will normally use. The irregularities in phase are due phase errors in the windowed filters, as described above in reference to Eq. (2.39). Note that the phase errors shown here depend on the phases of *both* the data filter at 2 kHz and the pilot tone filter at 9 kHz, since a phase error in measuring the pilot tone will appear as a time error, and thus as an error in relative phase at 2 kHz. (The amplitude plots do *not* depend on the amplitude of the pilot tone.) In particular, the 1*st*-order window shows phase errors that repeat every 100 Hz in frequency rise (as we pass from S_{20} to S_{21} to S_{22}) and also 4.5 times as fast (as the pilot tracker passes from S_{90} to S_{91} to ... to S_{100}). Using higher-order windows decreases the phase errors of the windowed filters. By the time we get to the 3*rd*- and

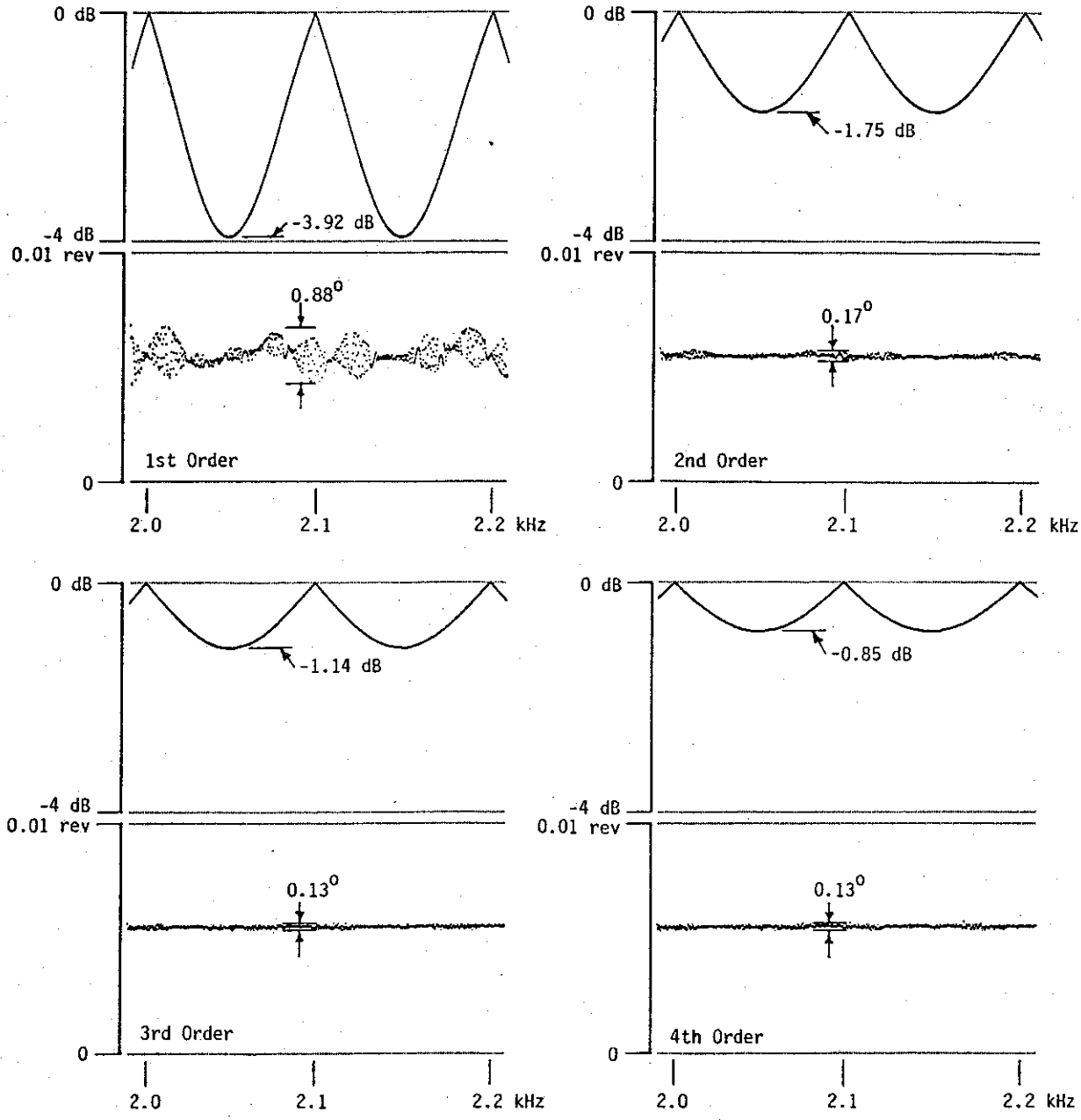

Figure 2.8. Magnitude and phase plots showing errors *versus* window order when tracking a frequency ramp. The input signal was a constant-amplitude tone rising from below 2 kHz to over 2.2 kHz at a rate of 40 Hz/s. The pilot tone at 9 kHz increased proportionally in frequency.

4*th*-order windows the filter phase errors have fallen below the procesing noise, which is about 0.13° peak-to-peak in this case.

2.5.6 Subtraction of Reference Phase to Correct Tape Time Error

Benefits of Relative Phase Measurements. If phase information is being used, the phase of each interpolated spectral filter is converted into a *relative phase* by subtracting from it the phase (modulo 1 rev) of a reference oscillator which has been running at the center frequency of that filter for a time equal to the data segment time. That is, if $\phi(t)$ is the phase of the output of the spectral filter whose center frequency is f_0, then the relative phase of the filtered signal is $\phi_{rel}(t) = \phi(t) - 2\pi f_0 t$, the phase of the signal relative to a reference oscillator at frequency f_0.

Using relative phase measurements does three things for us. First, it allows us to correct for the time error. Until we have measured the pilot tone phase in the current data segment we won't know exactly what the data time is, and we won't know what the absolute signal phase at the desired time is either. Second, it makes it easier to interpret signal phase. A signal of mean frequency f_a has a phase which increases roughly $2\pi f_a$ radians per second. We are often not interested in the mean increase in phase with time but in variations about the mean. By subtracting the average increase we can more easily see the variations. And third, using relative phase makes it easier to measure instantaneous signal frequency. A signal at frequency f_a passed by a filter at f_0 has a relative phase that increases by $(f_a - f_0)$ revolutions per second. If we see a phase advance of one revolution in one second, we know f_a is one hertz above f_0. This is much easier than actually counting f_a cycles over the course of a second. In fact, we don't have to wait for whole cycles of phase difference but can estimate the instantaneous frequency of a signal from the slope of its relative phase. Since we can place our interpolated filters at arbitrary frequencies, we can measure signal components with respect to arbitrary references, a great convenience.

Calculating the Relative Phase. When the phase correction routine is called, $\{U_k\}$ contains interpolated spectral values in rectangular or (*real, imaginary*) form for filters at frequencies from f_{lo} to f_{hi} in steps of f_I. We could convert the interpolated spectrum array to polar or (*magnitude, phase*) form at this point, as it would make subtracting the reference phases a bit easier. However, it would make spectrum averaging, described in the next section, very much harder, so we'll keep everything in (*real, imaginary*) form. Phases in the interpolated spectrum $\{U_k\}$ are converted to relative phases as follows:

1. If we are neither plotting phase values nor going to do spectrum averaging, exit from the routine. Phase information is not needed. Magnitude information is not altered by subtracting a reference phase.
2. For each interpolated filter in $\{U_k\}$ calculate a reference phase (in revolutions) as
$$\begin{aligned}\theta_k &= (t_i - t_{start})f_k \\ &= (t_i - t_{start}(f_{lo} + kf_I).\end{aligned} \tag{2.44}$$
for $k = 0, 1, \ldots, J$. Only the fractional part of θ_k need be saved; whole revolutions are discarded. In this program the resulting values of θ_k are stored in an array as 16-bit integers from 0 to 65535 representing angles from 0 to 360^- degrees, but there are many other ways to approach this.
3. Subtract this reference phase from each line as
$$U'_k = U_k e^{-j\theta_k}. \tag{2.45}$$
The operation is actually carried out on the real and imaginary parts of U_k as
$$\begin{aligned}R'_k &= R_k \cos(\theta_k) + I_k \sin(\theta_k), \\ I'_k &= -R_k \sin(\theta_k) + I_k \cos(\theta_k).\end{aligned}$$

The values $\cos(\theta)$ and $\sin(\theta)$ are interpolated from the table of sines and cosines used for the FFT, described at the end of Sec. 2.5.2. This is much faster than calculating them on the spot.

Necessary Precision Required. The subtraction of the reference phase from each spectral filter output is quite straigntforward. We should note, however, that to achieve the desired accuracy the time variable t_i (and the other variables in Eq. (2.44) if they are not exact values) must be represented with considerable precision. We want to be able to measure signal phase with an accuracy of at least 0.1 μs if not better. (This is a phase of 0.36° at 10 kHz, still bigger than the phase errors of the synthesized filters.) This demands that we be able to specify the time of any data segment to this accuracy. If we handle input signals up to 400 s in duration then we must specify time with a precision of at least 1 part in 4×10^9. In most computer systems this calls for extended-precision arithmetic. In this system we use double-precision (8 byte) floating-point variables for t_i and t_{start}, and use the floating-point unit in its double-precision mode for reference phase calculation.

The need for extended precision can be relieved a little if we can restrict the arbitrary choice of values for f_p, f_{lo}, and f_I. If these three frequencies are all multiples of some common factor, say ν Hz, then it is clear that reference phases for all possible analyzed frequencies will repeat every $1/\nu$ seconds. For example, if f_p, f_{lo}, and f_I are all in whole units of hertz, then the ensemble of reference phases will be in exactly the same state after a lapse of one second. In this case we need only measure time in fractions of a second, and can ignore any whole seconds that have elapsed when it comes to calculating reference phases. This reduces the needed precision in representing time to 1 part in 10^7, which may be a single-precision quantity in some systems. Such a scheme will also be necessary in any system which can analyze signals of unlimited duration to avoid the need for unlimited precision in time variables.

2.5.7 Normalizing, Averaging, and Whole-Revolution Phase Accumulation

The analysis program performs the gain normalization, spectral averaging, and accumulation of whole revolutions of phase in one large loop. In this loop the individual filter outputs represented in the interpolated spectrum $\{U_k\}$ are processed one-by-one to update the output spectrum $\{V_k\}$ and the associated integral-revolution accumulated phase array $\{\Phi_k\}$. For simplicity, we will discuss the steps involved as if they were performed separately.

Normalizing. The magnitudes of spectral components at this point are determined not only by their strengths in the input waveform but by several other factors as well. First, different data segments may have suffered different amounts of halving to prevent overflow in the FFT routine. Second, the gain of the FFT is proportional to the transform size N, and so may vary from one analysis run to another if different DFT filter spacings are used. (The gain is exactly N in this implementation, $N/2$ for the DFT as in Eq. (2.28), and 2 from the real-data transform extraction in Eq. (2.24).) Third, the window convolution introduces a loss (proportional to window coefficient C_0) that depends on the order of the window used. If we are to compare different spectra we will have to take all these factors into account. It is also convenient for the operator to be able to change the overall processing gain to accomodate different input signals, attenuating or amplifying to provide the desired density in the plotted spectrogram. The operator does this via the output gain factor G_{out}.

The task of normalizing is to account for all of these factors and produce an output spectrum referred to the same constant full-scale level for use by the plotting routines. We proceed as follows:

1. Elements in the interpolated spectrum are each normalized as

$$S'_k = G_{norm} S_k \qquad \text{for } k = 0, 1, \ldots, J, \tag{2.46}$$

where

$$G_{norm} = \frac{10000 \cdot 2^{j_{scale}} G_{out}}{2048 N C_0}$$

and where $J = (f_{hi} - f_{lo})/f_I$ is the index of the highest filter in $\{S_k\}$. The factor 10000 is the nominal full-scale level used by the plotting routines. The factor 2048 is the peak waveform value digitized by the 12-bit A/D converter. This is used so a single-frequency signal at peak input level with an output gain of 0 dB ($G_{out} = 1$) will give a full-scale output. The multiplications in Eq. (2.46) are performed separately on the real and imaginary components of S_k since G_{norm} is a real number.

2. Both the real and imaginary components of S_k are limited to ± 32767 to prevent arithmetic overflow. This is necessary since $\{S_k\}$ is an array of 16-bit integers and the normalization factor G_{norm} is very often greater than 1.

There is about 10 dB of headroom between the nominal full-scale level of 10000 and the clipping level, which seems to be adequate. Note, however, that clipping real and imaginary components independently creates phase distortion as well as amplitude distortion. Since both real and imaginary components of clipped signals, such as spherics, tend to ± 32767, the phases of clipped signals tend to 45°, 135°, 225°, or 315° (*i.e.*, those angles whose tangents are ± 1).

A useful and not very difficult improvement would be to use some sort of proportional clipping, clipping both real and imaginary components by the same factor when an overflow occurs in order to preserve the phase angle of the transient. I think this might have a significant effect on the phase noise generated by spherics, especially when averaging.

Averaging. The results in each spectrum are usually processed and plotted independently; each output represents the signals present in the data segment just analyzed and is independent of signals in other data segments. However, spectrum averaging may be used to smooth the output plot by combining the spectrum just generated with previous spectra. The real and imaginary values of each point in the current interpolated spectrum $\{U_k\}$ are weighted and added to the previously averaged values in the output spectrum $\{V_k\}$, and the results are saved as the new output spectrum as follows:

1. If this is the first pass through the data (data time t_0), set all the elements in $\{V_k\}$ to zero so the initial values for averaging are all zero.
2. If not averaging, the interpolated spectrum is just copied directly to the output spectrum as

$$V'_k = U_k \qquad \text{for } k = 0, 1, \ldots, J. \tag{2.47a}$$

3. If averaging, the interpolated spectrum is weighted and combined with previously averaged values as

$$V'_k = e^{t_{step}/\tau_{avg}} V_k + (1 - e^{t_{step}/\tau_{avg}}) U_k \qquad \text{for } k = 0, 1, \ldots, J. \tag{2.47b}$$

This gives an exponentially-weighted running average, where the effect of a given datum decreases as an exponential function of time with time constant τ_{avg}. Note that this is not the same as a block running average. The exponential running average depends on values only in the past and not on future values. The impulse response of a synthesized filter with averaging is not symmetrical in time, unlike the response of the analysis filters up to this point.

Averaging is a useful technique when a smooth plot of otherwise noisy data is desired. However, averaged results must be interpreted with caution. Since the average is one-sided in time, the effect of averaging is to extend transient events forward in time, even though their amplitudes may be

reduced. If the input signal contains impulsive noise, the original unaveraged output will show large excursions in magnitude and phase for all data segments transformed that contain the impulse but will show no effects for data segments before or after the event. Averaging tends to reduce the immediate effects of the impulse, but it will also cause the effects to extend to times after the impulse since the impulsive values decay only exponentially with time rather than being immediately cut off as the data window is moved over the offending samples.

Averaging may also be valuable with certain types of phase-coherent signals where it is desired to have many closely-spaced output lines with narrow passbands. Lines at frequencies offset from a given signal frequency will have relative phases which increase or decrease with time, and averaging will tend to null these out, thus effectively decreasing the width of the synthesized analysis filters. However, the effects of transients are rather complicated, and this technique should be used with caution. It may be helpful to simulate the analyzed signals with a synthetic test signal to be sure that the results are as expected.

Whole-Revolution Phase Accumulation. Phase information in the output spectrum is plotted with a full-scale range of P_{span} revolutions. P_{span} is either an integer n, or a fraction whose value is the reciprocal of an integer as $1/n$. That is, we plot phase as either n revolutions full-scale, or one-nth revolution full-scale. The phase of a complex value in the output spectrum $V_k = R_k + jI_k$ is given by $\phi_k = \arg(V_k) = \arctan(I_k/R_k)$, and can have any value (in revolutions) such that $0 \le \phi_k < 1$. When plotting phase to more than one revolution full-scale we have to keep track of integral revolutions separately from the fractional-revolution phases in $\{V_k\}$.

For example, imagine the relative phase of a signal just a bit above the center frequency of a synthesized filter. Assume we are plotting phase two revolutions full-scale, $P_{span} = 2$. At successive intervals we might see output spectrum phases ϕ_k of, say, 0.8, 0.9, 0.0, 0.1, ... revs. We want to plot the phase as 0.8, 0.9, 1.0, 1.1, ... Where ϕ_k changes from 0.9 to 0.0 revs, we must recognize that the actual signal phase has not decreased by 0.9 revs but rather has increased by only 0.1 rev to 1.0.

We keep track of whole revolutions of phase in the auxilliary array $\{\Phi_k\}$. Each element Φ_k is an integer such that $0 \le \Phi_k < P_{span}$, representing any whole revolutions of phase of the output filter V_k. That is, the total phase of V_k is $\Phi_k + \phi_k$. The phase accumulation algorithm is as follows:

1. If this is the first pass through the data (data time t_0), set all the elements in $\{\Phi_k\}$ to zero. All phase plots start with zero accumulated phase.
2. If we are not plotting phase information, or if $P_{span} \le 1$, exit. Phase accumulation is not needed.
3. Let the previous value of the output spectrum for the k-th filter be $V_k = R_k + jI_k$, with phase $\phi_k = \arctan(I_k/R_k)$ such that $0 \le \phi_k < 1$. Let the values just calculated by Eq.'s (2.47a) or (2.47b) for the current segment be $V'_k = R'_k + jI'_k$ with phase ϕ'_k. We have two cases:
   a. If $\phi_k < 1/2$, we should decrement Φ_k whenever the new angle ϕ'_k is greater than $\phi_k + 1/2$, thus making the change in total accumulated angle less than $1/2$ rev. Equivalently, we should decrement Φ_k when

$$I_k > 0, \quad I'_k < 0, \quad \text{and } \cot(\phi'_k) = R'_k/I'_k < \cot(\phi_k) = R_k/I_k,$$

   or when

$$R'_k \cdot I_k > R_k \cdot I'_k \qquad (\text{since } I'_k < 0). \tag{2.48a}$$

b. If $\phi_k > 1/2$, we should increment Φ_k whenever the new angle ϕ'_k is less than $\phi_k - 1/2$, thus making the change in total accumulated angle less than 1/2 rev. Equivalently, we should increment Φ_k when

$$I_k < 0, \quad I'_k > 0, \quad \text{and } \cot(\phi'_k) = R'_k/I'_k > \cot(\phi_k) = R_k/I_k,$$

or when

$$R'_k \cdot I_k < R_k \cdot I'_k \qquad (\text{since } I_k < 0). \tag{2.48b}$$

4. After decrementing or incrementing Φ_k, the new value is checked to be sure it is still in the range $0 \le \Phi_k < P_{span}$. If not, an amount P_{span} is added or subtracted as necessary.

2.5.8 Rectangular-to-Polar Conversion and Scaling for Plotting

Finally, the points in the output spectrum are either graphed on the terminal, plotted on the plotter, or written to an output file for later use. The program then returns to the first step (Sec. 2.5.1) above and moves to the next segment of input data, continuing until the desired interval of signals has been analyzed. In this section I describe the final operations on the spectral data prior to graphing or plotting.

Rectangular to Polar Conversion. First, the output spectrum $\{V_k\}$ is converted from rectangular (*real, imaginary*) form to polar (*magnitude, phase*) form. If spectrum element V_k is given by $V_k = R_k + jI_k = A_k e^{\phi_k}$, then $A_k = (R_k^2 + I_k^2)^{1/2}$ is its magnitude, and $\phi_k = \arctan(I_k/R_k)$ is its phase. The following algorithm is adapted from procedures presented by *Hart et al.* [1978]:

1. R_k and I_k, the real and imaginary parts of V_k, are signed 16-bit integers ranging from -32767 to $+32767$. Call the one with the larger absolute value u and the other one v. Calculate the ratio $x = v/u$, and the product x^2. Note that $|x| \le 1$.
2. The magnitude A_k is given by $A_k = u(1+x^2)^{1/2}$. We calculate $(1+x^2)^{1/2}$ by Newton's method. Let $a = 1 + x^2/2$. Find $b = [a + (1 + x^2)/a]/2$. Find $c = [b + (1 + x^2)/b]/2$. Then c is a very good approximation to $(1 + x^2)^{1/2}$, and we assign $A_k = uc$. The result is an unsigned 16-bit integer ranging from 0 to 46340 $(= 32767 \cdot 2^{1/2})$.
3. Approximate ϕ_k as a power series as $\phi_k = (20846x^7 - 6071x^5 + 3044x^3 - 806x)/2$. If $u = I_k$ then assign $\phi_k = 2^{14} - \phi_k$. If $u < 0$ then assign $\phi_k = \phi_k + 2^{15}$. All operations are modulo 2^{16}. The resulting value of ϕ_k is a 16-bit unsigned integer ranging from 0 to 65535, representing angles from 0 to 1^- revolution (0 to 360^- degrees).

All of the operations above are performed with integer arithmetic. That is, x is represented by $32768x$, and so on. The entire procedure requires only seven 16-bit by 16-bit integer multiplications, and three 32-bit by 16-bit integer divisions. The magnitude calculation has a maximum error of $+2/-0.5$ over the full range of magnitudes, with an error of only ± 0.5 for small A_k. The phase calculation has a maximum error of ± 2 ($\pm 0.011°$), and an rms error of about 0.7, for all values of ϕ_k. The errors in the rectangular-to-polar conversion are small compared to processing noise in earlier steps like the FFT.

Scaling. The resulting magnitudes and phases are next scaled to the number of levels required by the output plot. Lines on the output plot are not continuous, of course, but are drawn by turning on or off various numbers of the 1600 writing nibs that cross the paper as it is pulled through the plotter. The 200 dot/inch resolution of the plotter is fine enough to make lines that look acceptably

continuous, but they are ultimately discrete. The actual number of magnitude or phase levels that can be plotted depends on the format of the plot, and the pixel size or trace deflection specified.

Assume that Λ magnitude and phase levels are to be plotted. Λ is an integer typically from 10 to 500. The task in scaling is to map the magnitude and/or phase values into integers from 0 to Λ. There are three different cases:

1. *Linear Magnitude.* This is the simplest case. Here the darkness of the spectrogram or the deflection of a trace is directly proportional to the magnitude of a spectral component. The scaled magnitude is calculated as $\lambda_k = A_k\Lambda/10000$, where the divisor 10000 is the normalized full-scale magnitude level (see Sec. 2.5.7 above). If $\lambda_k > \Lambda$, then $\lambda_k = \Lambda$ is used.
2. *Logarithmic Magnitude.* In this case magnitudes are to plotted over a range of M_{span} decibels, where M_{span} is between 1 and 80 dB. The scaled magnitude is calculated as $\lambda_k = \Lambda + 20\Lambda\log(A_k/10000)/M_{span}$, if $A_k > 0$, else $\lambda_k = 0$. The operation is actually performed as $\lambda_k = a\log(A_k)+b$, where $a = 20\Lambda/M_{span}$ and $b = \Lambda - a\log(10000)$ are pre-calculated constants. The resulting value of λ_k is then clipped to the range 0 to Λ.
3. *Phase.* Phase is plotted P_{span} revolutions full-scale, where P_{span} is either an integer n, or the reciprocal of an integer as $1/n$; in either case n is from 1 to 100. If $P_{span} > 1$ then $\{\Phi_k\}$ contains any accumulated whole revolutions of phase, otherwise it is all zeroes. Phase is scaled as follows: First, find $\Psi = 2^{16}\Phi_k + \phi_k$, the total phase. Next, let $\psi = (\Psi/P_{span}) \pmod{2^{16}}$, the phase P_{span} revolutions full-scale. Finally, find the scaled phase as $\lambda_k = \psi\Lambda/2^{16}$. In this case λ_k is an integer from 0 to $\Lambda - 1$.

2.5.9 Plot Formats

Figure 2.9 shows examples of the four different types of output formats that can be used to plot analyzed signals. All examples use the same input data, a synthetic signal containing a constant-frequency tone at 4000 Hz, and a second tone specified by $f = 4800 - 1000t + 250t^2$, where f is in hertz and t in seconds. This second tone has a parabolic frequency-time shape, and reaches a minimum frequency of 3800 Hz after 2 seconds. Both tones have the same amplitude. Both tones begin abruptly at 0 s, and cease abruptly at 4 s. However, even though all four plots present the same signal, they show very different aspects of it. In this section we will examine some of the characteristics of each format.

F-T Spectrogram Format. The top plot in Fig. 2.9 is an *f-t* spectrogram. The density of the plot at a given position is proportional to signal magnitude at the corresponding frequency and time. The plot is actually divided up into rectangular pixels, each h dots high and w dots wide. Every h dots in the vertical direction represents an increase in frequency of f_I Hz, and every w dots in the horizontal direction is an increase in time of t_{step} seconds. Up to $\Lambda = h \cdot w$ different signal levels are shown by turning on different numbers of dots in each pixel. Linear magnitude scaling is usually used.

The *f-t* spectrogram is an effective format for showing the gross characteristics of signals. Its biggest advantage is that it can give complete coverage of the *f-t* plane in the most compact form. In the spectrogram in Fig. 2.9 the *f-t* plane is well oversampled since the interpolated filters (f_I = 12.5 Hz) are spaced at about one-third their 3-dB width ($1.6f_D$ = 40 Hz), and their outputs are sampled (t_{step} = 11 ms) almost four times as often as minimally necessary ($1/f_D$ = 40 ms). We can see some fairly small signal features, such as the transients at the beginning and end of the tones, and the ripples when the tones cross, though what these features mean may not be entirely clear.

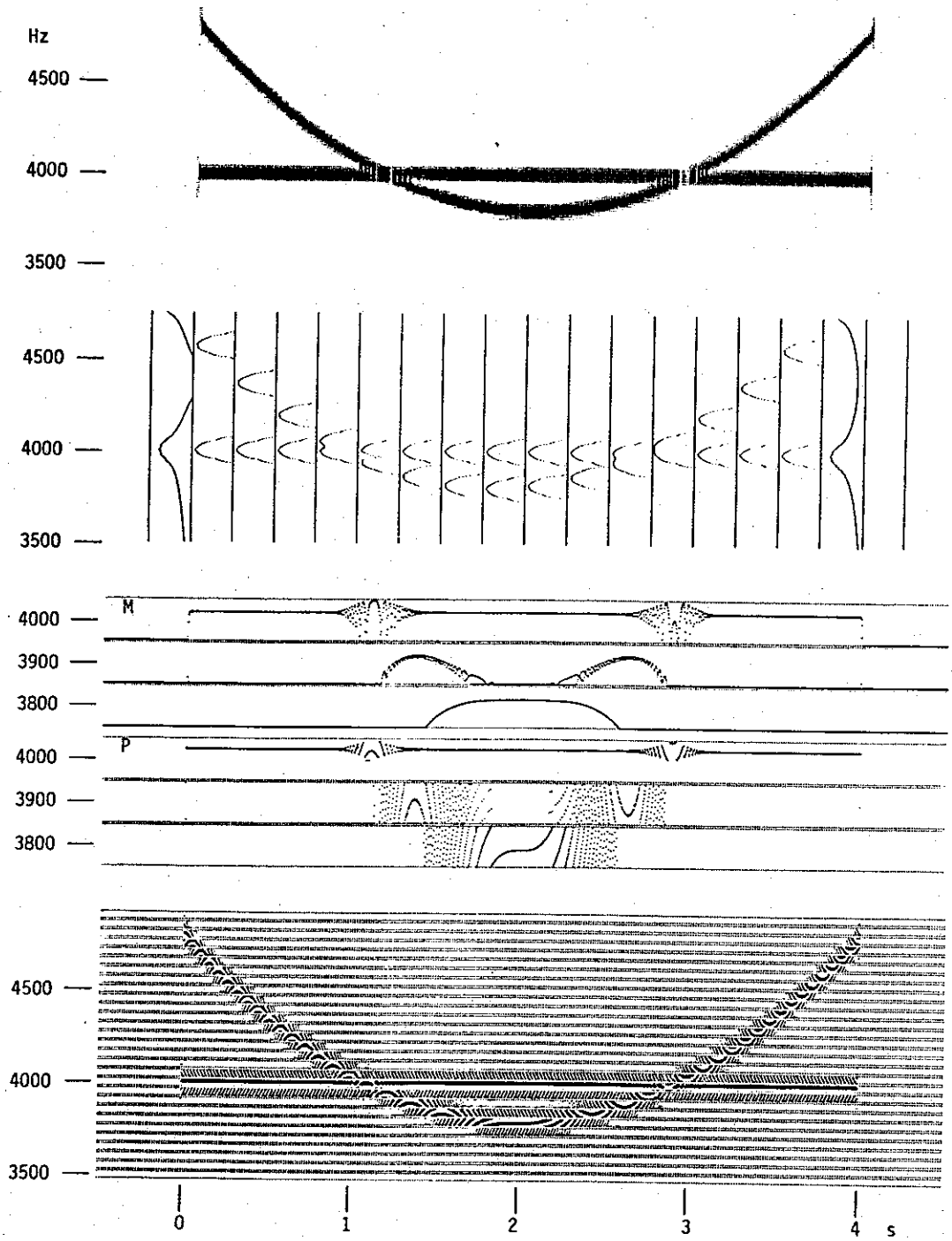

Figure 2.9. Sample output formats. All plots show the same signal containing a constant-frequency tone and a second tone with a parabolic dependence of frequency on time. The plots are, from top to bottom, an *f-t* spectrogram, an A-scan plot, a magnitude-phase plot, and a gray-scale phase plot. Analysis parameters are as follows:

	t_{step}	f_D	f_I	M_{span}	P_{span}	
F-T	11 ms	25 Hz	12.5 Hz	linear	—	pixel 5×4 nibs
A-scan	250 ms	25 Hz	2.5 Hz	40 dB	—	max deflection 90 nibs
Mag-Phase	2.75 ms	50 Hz	100 Hz	20 dB	1 rev	max deflection 89 nibs
Gray-scale Phase	2.75 ms	50 Hz	50 Hz	linear	1 rev	max deflection 15 nibs

The main shortcomings of an *f-t* spectrogram are the limited dynamic range that can be plotted, and the lack of phase information. Even though there may be Λ dots available in a given pixel (20 in this case), it is unlikely that Λ different signal levels can be discerned in black-and-white. The problem here is that the difference between a density of, say, 0 and 1 dot per pixel is much greater to the eye than the difference between $\Lambda - 1$ and Λ dots per pixel. As far as lack of phase information goes, note that the two places where the tones cross over look different. This is because the tones meet each other with different phases in the two cases, something not visible in the spectrogram. Also, we cannot tell the minimum frequency at the bottom of the parabola, at least closer than about f_D Hz, without phase information.

A-Scan Format. The second plot in Fig. 2.9 is an example of the A-scan or amplitude-scan format. In this format the magnitudes of successive filter outputs at a given time are plotted as horizontal deflections of a vertical trace. Additional traces are plotted at intervals of t_{step} seconds. If the synthesized filters are spaced closely enough, the traces will appear to be continuous in frequency, representing the magnitude of the windowed spectrum at discrete intervals of time. The vertical scale assigns h dots to each successive filter at increments of f_I Hz. The horizontal scale depends both on the peak deflection Λ of each trace and the time t_{step} between traces. Either linear or logarithmic magnitude scaling can be used.

The A-scan format overcomes the dynamic range limitation of the *f-t* spectrogram, and makes it possible to measure the actual magnitude of a signal component. In particular, note that the A-scan plot in Fig. 2.9 has captured the transients at the beginning and end of the tones. Signal magnitudes are shown here with a range of 40 dB, much greater than in the spectrogram, and the transients are seen to extend to much wider frequencies. The capture of these transients also points up a major limitation of the A-scan, sparse coverage in time. To provide complete sampling along the time axis requires an enormous amount of plotting in the A-scan format. In Fig. 2.9 t_{step} was 250 ms, more than six times the data segment length. 84% of the input samples were skipped over. If the windows that captured the end transients were moved as little as 20 ms, the transients would not be seen.

The effects of linear interpolation between DFT filters can be seen in the A-scan in Fig. 2.9. Ten interpolated filters are plotted for each windowed filter. Note that the passband-like shapes are actually made up of straight line segments. These are the lines along which the output filters were interpolated.

The A-scan format used here shows signal spectra as a series of separate little graphs. A similar format which is more efficient in its use of paper is the overlapping-trace format where the horizontal distance between successive traces is less than the deflection permitted on any given trace. A strong signal component may produce a deflection that moves its trace beyond the baseline of one or more previous traces. When this format is plotted with frequency across the paper and time increasing downward it is called a "waterfall" display. Unfortunately, plotting an overlapping-trace display requires, at any given moment, saving a few previous spectra, and this takes more memory space than we have available. It will be a desirable format in any future system.

Magnitude-Phase Format. The magnitude-phase format, the third plot in Fig. 2.9, shows the magnitude and/or phase of a set of filter outputs as functions of time. The vertical scale in each trace shows Λ different values of magnitude or phase. Successive traces represent filters at successive increments of f_I Hz in frequency. In the horizontal direction we plot w dots for each increment t_{step} in time. If w and t_{step} are small enough, each trace will appear to be continuous in time, representing the actual output of a fictitious filter. Either linear or logarithmic magnitudes can be plotted, and

to wow and flutter. Timing error correction is necessary if phase measurements are to be made. The frequency of the reference tone is not important as long as it falls within the sampling bandwidth. All recordings made by Stanford since 1973 include a pilot tone at either 1, 9, or 10 kHz. However, any contant-frequency signal in the data may be used. If the signal is amplitude modulated with time ticks, these may also be displayed.

5. *Analyze Data with Various Filter Bandwidths.* The FFT algorithm is used during analysis to calculate the discrete Fourier transform (DFT) of the sampled signal. The DFT represents the signal as passed through a bank of bandpass filters spaced in frequency at multiples of $f_D = 1/NT$, the DFT filter spacing. By varying the transform size N we can change the spacing of these synthesized filters. N can be any power of 2 from 64 to 2048, so with 10.6 kHz sampled data ($T = 1/25600$ s) the DFT will synthesize filters spaced every 12.5, 25, 50, 100, 200, or 400 Hz. The bandwidth of each filter is equal to the filter spacing f_D multiplied by a factor that depends on the window function used with the FFT. With a 3*rd*-order window (the usual choice) the factor is 1.6, giving filter bandwidths of 20, 40, 80, 160, 320, or 640 Hz.

6. *Interpolate Between DFT Spectral Points for Arbitrary Filter Spacing.* Linear interpolation between points in the DFT spectrum can be used to generate an output spectrum with filters spaced at an arbitrary frequency increment f_I. However, the bandwidth of these interpolated filters is still approximately the bandwidth of the DFT filters. Interpolation is also used to correct for frequency shifts due to tape speed errors.

7. *Make Accurate Magnitude Measurements.* The modulus of a given spectral filter output at a given time is the magnitude of the signal component at the filtered frequency at that time. Signal magnitudes can be displayed either linearly or logarithmically (over a 1 to 80 dB range) and can be measured directly from the output plot. The relative magnitudes of signal components at different frequencies are easily determined. If a signal of known level is measured (say, a receiver calibration tone), then absolute measurements of signal intensities can also be made.

8. *Measure Relative Signal Phase.* The argument (angle) of a spectral filter output is the phase of the signal component at that frequency and time. The phase of a constant-frequency component at frequency f is, of course, just ft revolutions (or $2\pi ft$ radians) at time t, and increases f revolutions every second. The analysis system calculates the phase at the output of each of the spectral filters, subtracts an amount equal to $f_0 t$ (where f_0 is the center frequency of the particular filter), and can display the results in a variety of formats. The resulting values are relative phase measurements, representing the difference in phase between a signal component at frequency f and a reference oscillator at the filter center frequency f_0. These relative phases may be of interest in themselves, or their slope may be measured from the output plot to find the frequency difference between the signal and the reference and thus make precise instantaneous signal frequency measurements. However, to generate meaningful phase plots, tape timing errors must be corrected, which means that a reference pilot tone must be available in the data.

9. *Average Spectra to Measure Weak Signals.* The complex spectra may be averaged with time to help bring weak coherent signals out of the background noise. If the signal is coherent (and exactly at the center frequency of a spectral filter) its relative phase will be constant with time, whereas noise in the filter bandwidth may be expected to have a random phase from one moment to the next. By averaging, the noise tends to disappear, and the signal-to-noise ratio improves directly with the averaging time. When averaging is used, the complex output of each spectral line is combined with previous values to form a running average with an exponential decay in time. The time constant can

be chosen depending on the strength of the desired signal and its coherence time. The magnitude and/or phase of the resulting average may be plotted in a variety of ways. For best results, tape timing errors must be corrected, which means that a pilot tone must be present in the data.

10. *Produce Frequency-Time Spectrograms.* The analysis system can produce frequency-time spectrograms similar to the film and paper records produced by analog spectrum analyzers. The system plotter has sufficient resolution (200 dots/inch) to make gray-scale plots much like lithographic half-tone prints. These plots are used to help locate interesting signal features.

11. *Produce A-Scan Plots.* An A-scan (amplitude-scan) plot shows signal magnitude versus frequency at a given time (i.e., for a given data segment) as the deflection of a horizontal trace. As the analysis proceeds, a separate trace is plotted for each step in time t_{step}. The resulting plots can be used to measure signal magnitudes at different frequencies. Analog spectrum analyzers can produce A-scan plots with difficulty, but the digital system does so with ease.

12. *Produce Magnitude-Phase Plots.* These are the most useful plots for the analysis of small-scale signal structure. In a magnitude-phase plot the magnitude and/or phase of one or more spectral filters is plotted as a function of time. Either linear or logarithmic magnitudes can be plotted, and phase plots can be made from 0.01 to 100 revs full-scale. One filter output at a given frequency may be shown to large scale, enabling accurate magnitude and phase measurements to be made from the resulting plot, or the outputs of several filters may be plotted next to one another to show the relationship between different signal components.

13. *Produce Gray-Scale Phase Plots.* In the gray-scale or width-modulated phase plot the magnitude and phase of a spectral filter are combined and plotted together as a function of time. Trace deflection is proportional to signal phase and trace width to signal magnitude. A single filter may be plotted to large scale, showing the correlation between signal behavior in magnitude and phase, or the outputs of many filters may be plotted to a smaller scale, in which case the plot is similar to an f-t spectrogram but also showing the phase behavior of coherent signals.

2.7 What This System Cannot Do

1. *Digitize Signals from 1/2-Inch Tapes, Quarter-Track Tapes, or FM Recordings.* The analog tape recorder used in the digital analysis system can only reproduce 1/4-inch half-track direct (AM) recorded tapes. All other tapes must first be dubbed onto a 1/4-inch tape in this format. This means, for instance, that NASA satellite tapes (1/2-inch FM) and many Palmer Station tapes (quarter-track) cannot be directly analyzed.

2. *Translate Data When Sampling.* All data are normally sampled at 25600 samples/second, preserving signals from 0 to 10.6 kHz. While the effective data bandwidth with respect to recorded signals may be varied by changing the playback speed, all sampled signals start at zero frequency. The system has no provision to select a small passband from the recorded signal and translate it down before sampling, as can be done with some of the analog analyzers.

3. *Analyze Data with Arbitrary Filter Bandwidths.* The transform size during analysis may be changed from 64 to 2048 points by factors of 2. In practice, this means that each transformed spectrum may contain 27, 54, 107, 213, 425, or 849 unaliased spectral lines evenly spaced from 0 Hz up to the sampled data bandwidth. For example, with 10.6 kHz data, a 2048-point transform provides 849 lines spaced every 12.5 Hz from 0 to 10.6 kHz. With a 3*rd*-order window, the actual 3-dB width of each equivalent spectral filter is 1.6×12.5 or 20 Hz. Interpolation may be used to generate additional spectral lines (say every 1 Hz), which are often useful when making phase measurements, but the bandwidths of these lines are still about 20 Hz. There is no way to analyze 10.6 kHz data with filter bandwidths less than 20 Hz. (Spectrum averaging can help in certain cases.)

4. *Correct Tape Wow and Flutter Without a Pilot Tone.* A constant-frequency reference or pilot tone is needed to make tape timing error corrections, necessary for phase plots or averaging. During analysis, the phase of the pilot tone is measured and used to determine the actual data time of each data segment. The data time determines the reference phase which is subtracted from each spectral line phase to give the relative signal phase. Without tape timing error correction, the phase of a signal component will depend not only on the signal itself but also on the accumulated speed errors of both the recording and playback decks at that moment. Tape speed errors are such as to often completely mask any changes in signal phase that are to be measured. All Stanford recordings made since 1973 have a pilot tone that can be used for this purpose (though some operators have inserted the tone at a low level, which can cause problems). Pre-1973 data are in general not usable for phase analysis.

5. *Correct Wow and Flutter of Multi-Channel Recordings.* Multi-channel data are digitized by interleaving samples from the different channels. During analysis, the samples of the desired channel are extracted and processed while samples from other channels are discarded. Tape timing errors can only be corrected on those channels that contain a constant-frequency pilot tone. The system does not allow error corrections derived for one channel to be applied to any other channel. Tape slew across the heads (azimuth error) and head scatter (tape head gaps for different channels may not be exactly in line) mean that tape timing errors will differ slightly from one track to the next. Even if the analysis program allowed us to measure a pilot tone on one track and apply the data times measured to samples from another track, we would find that the residual timing errors from slew and scatter would render phase measurements very noisy. See Sec. 2.3 for a discussion of this problem.

6. *Make Useful Phase Measurements of Changing-Frequency Signals.* Phase measurements can only be easily interpreted for signals that are relatively constant in frequency, or whose phases can be referred to a constant-frequency signal. For instance, when studying wave growth in the magnetosphere we might look at the phase of a received signal amplified from or triggered by a constant-frequency input pulse. By measuring the relative phase of the received signal we can determine the instantaneous frequency of the amplified signal or the emission since we know the frequency of the input wave. However, if the input signal is not constant in frequency, its changing relative phase behavior will be added to that of the interaction and the resulting output phase plot will be very complicated. This means that the phase behavior of transmitted frequency ramps and other more complicated signals is very difficult to interpret. Calculating dispersion from the phase behavior of ramps (or whistlers) is beyond the capabilities of the present system.

3. SIGNALS WITHOUT GROWTH

3.1 Identification of VLF Transmitter Modulations

In this section we will show how the relative phase of the signal from a VLF transmitter can be used to determine its frequency and modulation. From these we may infer the purpose of the signal and, occasionally, the state of development (or at least maintenence) of the transmitting equipment.

VLF Transmitters. The spectrum from 10 to 30 kHz is dominated by the sub-ionospheric signals from various VLF transmitters. The two main uses of these transmitters are navigation and communications. In addition, some stations also periodically send time signals, such as GBR, Rugby, UK, at 16.0 kHz; and UTR3 and UQC3, Gorki and Khabarovsk, USSR, at frequencies from 20.5 to 25.5 kHz.*

As VLF researchers we are not particularly interested in the message traffic these stations transmit, but the signals themselves can be very useful as probes to study geophysical phenomena. For instance, whistler-mode signals at 15.5 kHz from station NSS (Annapolis, MD) were heard at Cape Horn in November, 1956, supporting Storey's explanation of the magnetospheric propagation of whistlers which had been published only three years earlier [*Helliwell and Gehrels*, 1958]. The triggering of emissions by a man-made signal was first observed on whistler-mode signals at 18.6 kHz from NPG (NLK, Jim Creek, WA) received in Wellington, New Zealand [*Helliwell et al.*, 1964].

The frequencies of many transmitters are controlled by very accurate frequency standards, and the phases of received signals reflect this accuracy, at least after any message modulation has been removed. Sub-ionospheric signals from these transmitters have been used for many years as references against which the accuracy of local frequency standards can be measured. However, the phase of the signal at the receiver also depends on conditions along the path of propagation, and can change as ionospheric conditions change. *Chilton et al.* [1963] report phase advances of up to 30 μs lasting some tens of minutes due to ionization in the D region caused by solar flares. These signals are also subject to Trimpi events, the smaller phase and amplitude perturbations caused by ionization from magnetospheric particle precipitation.

Figure 3.1 shows an f-t spectrogram and a gray-scale phase plot of the spectrum from 10 to 22 kHz as recorded at Palmer Station, Antarctica, illustrating the types of sub-ionospheric signals that can be seen at any VLF receiving site. Table 3.1 lists the frequencies and modulations of the various signals in Fig. 3.1, and identifies their transmitters when known. The analog tape containing these signals was played back at about 48% of its recorded speed so the analyzable bandwidth of the digitized signal would be 22.26 kHz instead of the standard 10.6 kHz (*cf.* Sec. 2.4, Item 2). The spectrogram on the left shows the presence of at least a dozen signals. Those above the 10 kHz pilot tone up to 14 kHz are pulses from various Omega navigation transmitters, those 15 kHz and above are from communications stations using various modulations. (The signal at 14001 Hz is believed to be due to local interference at Palmer.) However, while we can tell the rough frequency of each signal from the spectrogram we cannot say very much about its modulation, except that some signals seem to have wider bandwidths than others.

The gray-scale phase plot on the right of Fig. 3.1 tells us more about the signal frequencies and modulations. Several signals are constant in frequency, and some even fall at the center frequencies

* The latter two stations may be part of a navigation system [*Klawitter*, 1983].

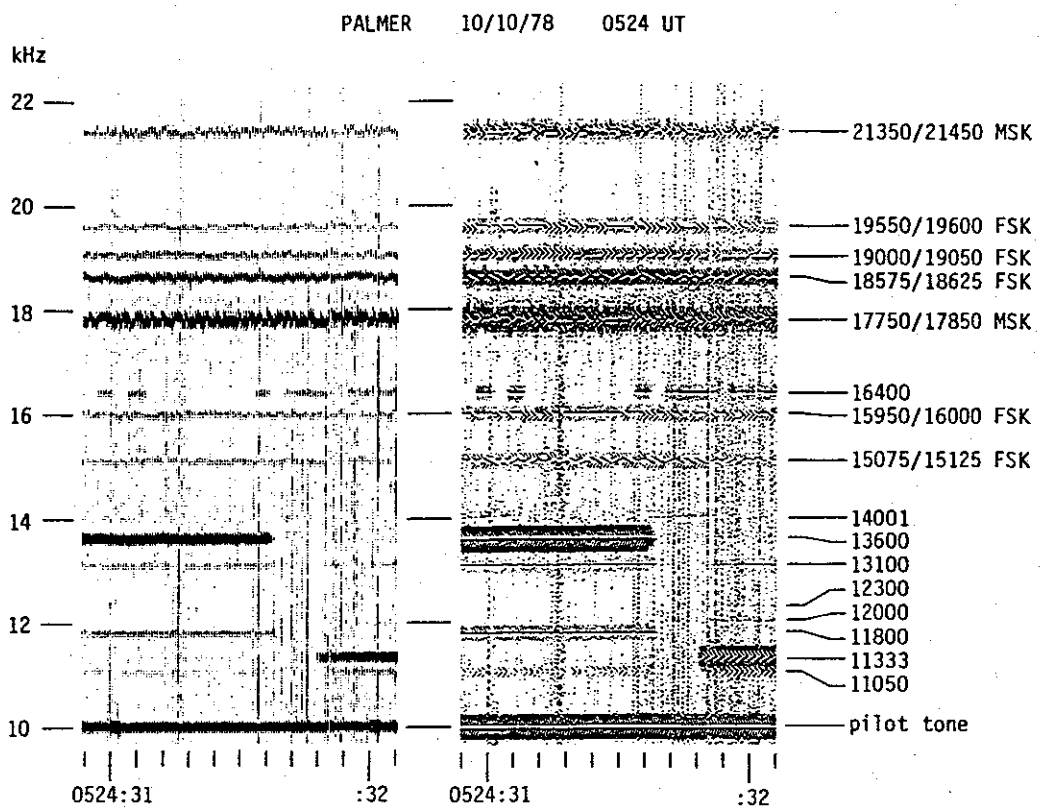

Figure 3.1. Signals from various VLF navigation and communications transmitters. On the left is a conventional *f-t* spectrogram, filter spacing $f_I = 50$ Hz, BW = 84 Hz. On the right is a gray-scale phase plot, $f_I = 100$ Hz, BW = 168 Hz, $P_{span} = 1$ rev. The phase characteristics of the signals reveal their exact frequencies and modulations.

of the analysis filters (spaced every 100 Hz) so their relative phases appear as straight lines. (The 10 kHz pilot tone phase is constant as well; but this is only as expected since it is the reference phase for everything shown). However, many of the signals are quite complicated, though their phase structures still show underlying regularities. The rest of this section will be spent examining all of these signals in more detail.

Omega. The frequency band from 10 to 15 kHz is used primarily for navigational purposes. There are two VLF navigation systems in use, the western Omega system and the Soviet Union's Alpha system. The two systems use different transmitters, different frequencies, and different pulse formats, but are similar in operation. The Omega system [*Swanson*, 1983] has eight VLF transmitters at locations around the world to provide global signal coverage. By measuring the phase difference between signals received from two transmitters a user can tell how much closer he is to one station than to the other, placing himself on a hyperbolic line of position whose foci are the two transmitters. By measuring phase differences from a third station he can calculate a second such line of position, and fix his location with a typical accuracy of about one nautical mile. In practice, signals at a given frequency are not transmitted simultaneously by the eight stations in the network, since it would be impossible to separate them at the receiver. Instead, stations transmit at a given frequency one at a time, and phase differences between pulses from different transmitters are measured with the aid of a local frequency standard at the receiver. Also, phases at four different frequencies are measured in order to resolve ambiguities in position due to whole-cycle uncertainties in phase measurement.

TABLE 3.1

VLF Signals in Figure 3.1

Frequency, Hz	Modulation	Station	
21350/21450	MSK, 200 baud	NSS, Annapolis, MD	
19550/19600	FSK, 50 baud	Soviet Union ?	
19000/19050	FSK, 50 baud	Soviet Union ?	
18575/18625	FSK, 50 baud	NLK, Jim Creek, WA	
17550/17850	MSK, 200 baud	NAA, Cutler, ME	
16400	keyed CW, 16 baud	JXZ, Oslo, Norway ?	
15950/16000	FSK, 50 baud	GBR, Rugby, UK	
15075/15125	FSK, 50 baud	Soviet Union ?	
14001	CW	local interference	
13600	Omega	Ω Argentina	Ω Australia
13100	Omega	Ω N. Dakota	Ω N. Dakota
12300	Omega		Ω La Reunion
12000	Omega		Ω Liberia
11800	Omega	Ω Hawaii	
11333.$\overline{3}$	Omega		Ω Argentina
11050	Omega	Ω Liberia	Ω Hawaii

at a single frequency. The transmission format is such that each of the eight stations sends a pulse at each of the four navigation frequencies once every ten seconds, as well as four pulses at a unique frequency peculiar to that station.

From our perspective, the important characteristics of the Omega signals are that they are constant-frequency pulses about 1 second long (the actual length varies among the eight segments in each ten-second cycle) followed by a 200 ms dead time, and that the frequencies and phases of the signals are accurately determined by a bank of cesium frequency standards at each station. Figure 3.1 shows signals from six of the eight Omega transmitters. The only stations not seen are Omega Japan at 12.8 kHz (left segment in Fig. 3.1), and Omega Norway at 12.1 kHz (both segments). Also, neither Omega Australia (left segment) nor Omega Japan (right segment) can be seen at 10.2 kHz because of the proximity of the much stronger pilot tone. The navigation frequencies are 10.2 (not seen), 11.05, 11.33$\overline{3}$, and 13.6 kHz; the other frequencies are the unique frequencies assigned to particular stations.

Because of the phase stability of the transmitters, Omega signals can be used to detect Trimpi events. *Inan et al.* [1985] report phase changes of up to 1 μs on the Omega Argentina signal at 12.9 kHz as received at Palmer Station. The Omega Argentina signal is quite strong at Palmer, and its phase is easy to measure. The disadvantage of using this signal is that, even at the 12.9 kHz unique frequency, the duty cycle is less than 50% and one cannot obtain complete time coverage of an event.

The Omega system has given us another benefit. The present VLF transmitter at Siple Station is a modified Omega transmitter that we obtained surplus when the original models were replaced in the mid 1970's.

Communications Transmitters. Frequencies above 15 kHz are used primarily for communications. Figure 3.1 shows eight communications stations between 15 and 22 kHz. VLF signals have been used in this way for decades, particularly by the world's navies. The advantage of VLF for communications is reliablility; VLF signals are much less subject to the vagaries of signal propaga-

tion that affect shortwave circuits. An additional advantage due to the long wavelengths at VLF frequencies is that attenuation in water is low (about 3 dB/m at 10 kHz) and communication with submarines is possible, at least to a depth of ten or twenty meters. The disadvantages of VLF signals are the need for high transmitter power to overcome atmospheric noise (station NAA *radiates* 1 MW), large transmitting antennas, and the limited bandwidth available (usually about 100 Hz). Most VLF communications transmitters are at fixed, land-based locations, but both the US Navy and Air Force have transmitters on aircraft using long trailing-wire antennas.

In order to transfer information, the VLF signal from a transmitter must be modulated in some way. There are two types of modulation commonly used, keyed CW (Continuous Wave), where the transmitter is turned on and off, and FSK (Frequency Shift Keying) where transmitter power is constant but the signal is shifted between two frequencies. Some US transmitters also use MSK (Minimum Shift Keying), which is really just a particular type of FSK.

Keyed CW is the simplest form of modulation. However, since the transmitter is off roughly half the time, average signal power is less than with FSK; and turning the transmitter on and off repeatedly may stress it more than constant-power transmission. Keyed CW is usually used at low bit rates, often to send Morse code messages intended for manual reception and decoding. Figure 3.1 shows a station at 16.4 kHz transmitting keyed CW at a speed of 16 baud (1 baud = 1 bit/s), or 19.2 words/minute. An hour earlier this station was heard sending "W33DG W33DG W33DG VVV VVV VVV DE JXZ JXZ JXZ" repeatedly (*i.e.*, "Calling W33DG, test test test, this is JXZ"). There is a Norwegian Army HF radioteletype station JXZ in Oslo (Kolsaas), Norway [*Ferrell,* 1983], and our VLF station may be associated with it. I have no idea who W33DG is.

We used the Siple transmitter when it first started operation in 1973 to send a keyed CW message in Morse code to Roberval over a whistler-mode path. Because of distortion in the magnetosphere (a problem not faced by sub-ionospheric users) the message had to be sent very slowly, at a rate of one baud or only a little over one word per minute. Manual keying at this rate is very tedious.

Frequency Shift Keying. FSK is the modulation most commonly used for communications. In this technique the message, which may be in Morse code but is more likely the output of a teletype machine or some more complicated encoding device, is represented as a stream of bits, 1's and 0's, or *marks* and *spaces* as they are called. The message is sent by transmitting at one of two possible frequencies, the mark frequency f_1 or the space frequency f_0, changing frequency according to successive message bits. There are two parameters in FSK modulation, the difference between the mark and space frequencies, or *frequency shift* $f_s = |f_1 - f_0|$; and the speed at which message bits are transmitted, the *keying rate* or *bit rate* $f_k = 1/T$, where T is the time taken to send an individual bit. The ratio $h = f_s/f_k$ is called the *modulation index* or *frequency deviation ratio.*

What does an FSK signal look like? Let's consider a string of message bits $\{u_n\}$, a sequence of 1's and 0's. Let the modulating waveform $u(t)$ that controls the transmitter be a function whose value, either 0 or 1, is the value of the message bit being sent at the given time t. That is, $u(t) = u_n$ during the interval $nT \le t < (n+1)T$. We will assume that $f_1 > f_0$ (mark frequency > space frequency), though this choice is arbitrary. Given these definitions, the frequency of the signal will be

$$\begin{aligned} f(t) &= \begin{cases} f_0, & \text{if } u(t) = 0, \\ f_1 = f_0 + f_s, & \text{if } u(t) = 1 \end{cases} \\ &= f_0 + u(t) \cdot f_s, \end{aligned} \tag{3.1}$$

and we can write the waveform of the transmitter as

$$\begin{aligned} s(t) &= \cos[2\pi \textstyle\int_0^t f(\tau)\, d\tau + \phi_0] \\ &= \cos[2\pi f_0 t + \phi(t) + \phi_0], \end{aligned} \tag{3.2}$$

where

$$\phi(t) = 2\pi f_s \int_0^t u(\tau)\, d\tau \tag{3.3}$$

carries the message information, and where ϕ_0 is some initial phase at time $t = 0$. Note that the sum $\phi(t) + \phi_0$ is the relative phase of the signal at the space frequency f_0 as we might measure it during spectrum analysis.

Now, the integral of $u(\tau)$ is just the sum of all the message bits already sent times the duration T of each bit, plus a little for the current bit in progress. If we are currently in the interval $nT \le t < (n+1)T$, sending bit u_n, then we can write

$$\begin{aligned} \phi(t) &= 2\pi f_s \left[T \sum_{i=0}^{n-1} u_i + \int_{nT}^{t} u_n\, d\tau \right] \\ &= 2\pi f_s T U_{n-1} + 2\pi f_s u_n \cdot (t - nT) \\ &= 2\pi h U_{n-1} + 2\pi f_s u_n \cdot (t - nT), \end{aligned} \tag{3.4}$$

where U_{n-1}, the sum of all previous message bits (*i.e.*, the number of marks), is an integer. What this says is that the relative phase at the space frequency at a given time depends on the modulation index h and all previous message bits, and is constant at the moment if we are sending a space ($u_n = 0$), or is increasing at f_s revolutions/second if we are sending a mark ($u_n = 1$). We can write a similar expression for the relative phase at the mark frequency f_1.

Phase Coherent FSK and MSK. Now we can examine two particular types of FSK. These are the modulations used by the majority of VLF communications stations. In the first type the frequency shift is exactly equal to the bit rate, so the modulation index h is exactly 1. In this case hU_{n-1} is always an integer, and the first term in Eq. (3.4) always represents some number of whole revolutions and can be ignored. Every time we send a space the relative phase at the spacing frequency has the same value (ϕ_0); every time we send a mark, the relative phase increases by exactly one revolution. We can make a similar statement about the relative phase at the mark frequency f_1: every time we send a mark the relative phase at f_1 is constant; for every space it decreases by one revolution. This special case is called *phase coherent* FSK. (Actually, any integral value of h will give phase coherence, but $h = 1$ is used to minimize the transmitted signal bandwidth.)

In the second particular type of FSK the frequency shift is exactly half the bit rate, or $h = 1/2$. In this case hU_{n-1} may be an integer, or it may be an integer plus one-half, and the first term in Eq. (3.4) is always an integral number of half-revolutions. Now there are *two* possible values of relative phase at the space frequency when we send a space. Which value occurs depends on the previous message bits. If we send a mark the relative phase at f_0 increases by one-half revolution. Similarly, at the mark frequency f_1 there are two values of relative phase for marks, and each space decreases the phase by one-half revolution. The case $h = 1/2$ is called *minimum shift keying* or MSK [*Doelz and Heald,* 1961].

The gray-scale phase plot in Fig. 3.1 shows five coherent FSK signals and two MSK signals. For three of the FSK signals (15950/16000, 19000/19050, and 19550/19600 Hz) one of the transmitted frequencies is at an analysis filter frequency and the relative phase is constant at that frequency

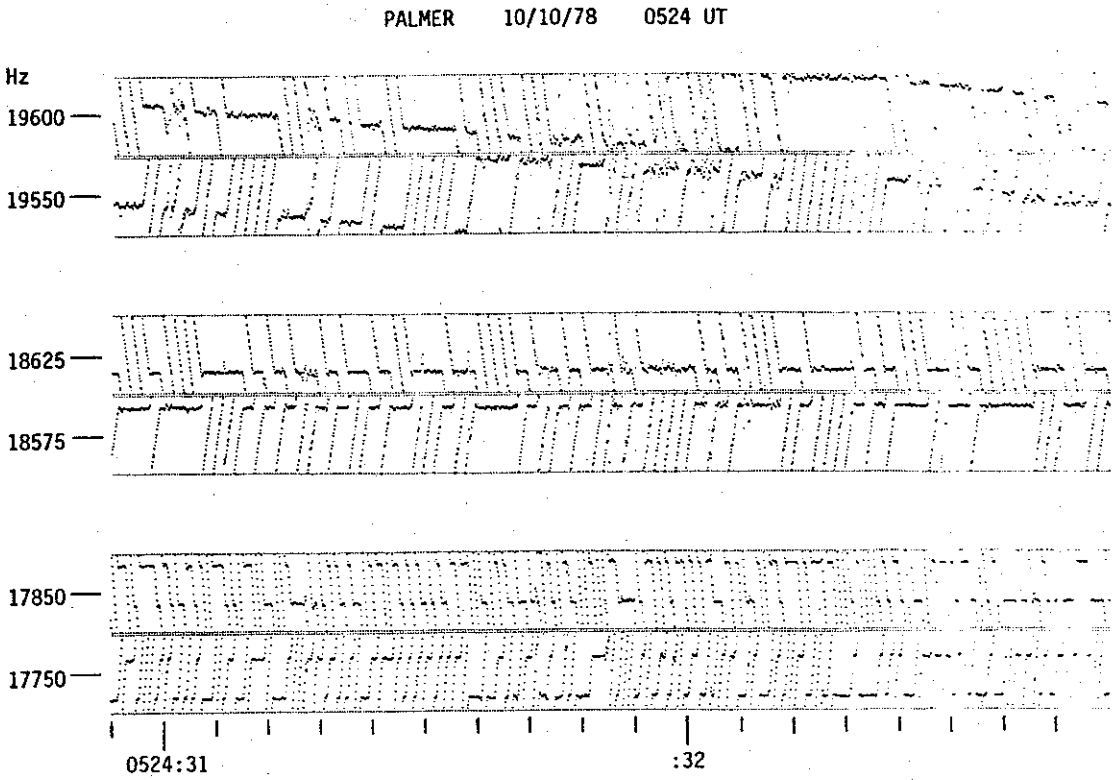

Figure 3.2. Phase plots of three selected VLF transmitters showing details of their modulation (BW = 336 Hz, P_{span} = 1 rev). The upper plot is a presumed Soviet transmitter sending FSK with 50 Hz shift and a keying rate of 50 baud. The average transmitter frequency is 0.6 Hz lower than expected. The center plot is NLK sending FSK with 50 Hz shift at 50 baud. The lower plot is NAA sending MSK at 200 baud; note there are two possible phase values at each frequency.

(though there may be a slow drift in the 19550/19600 Hz signal). The other two FSK signals (15075/15125 and 18575/18625 Hz) straddle a filter and their phase coherence is harder to see. The two MSK signals are quite complicated and the scale of the gray-scale phase plot is too small to say much about them.

Phase Plots of FSK and MSK Signals. With the aid of Figure 3.2 we can examine some of these signals in more detail. This figure shows phase plots for two FSK signals ($h = 1$) and one MSK signal ($h = 1/2$). In each plot two analysis filters have been synthesized, one at the mark frequency and one at the space frequency of that particular signal.

The simplest signal is shown in the center plot. This is station NLK sending 50 Hz-shift FSK at 50 baud. The space frequency is 18575, and the mark frequency is 18625. (Or *vice versa*. We cannot really tell mark from space without knowing how the message is encoded. The distinction is immaterial here and we will assume that the mark frequency is the higher one. However, all of the message information *is* displayed in the phase plot. If we knew the code we could read the message from this plot.) Note that the signal is phase coherent at the space frequency. The relative phase at 18575 is exactly the same for each space. Each mark sent lasts $T = 20$ ms and causes the phase at 18575 to increase by exactly one revolution; n marks increase it by n revolutions. Behavior at the mark frequency is similarly coherent—stationary at a constant value for marks and decreasing by one revolution for each space.

A phase-coherent FSK signal like this one is ideal for use in checking the accuracy of a local frequency standard (by comparing long-term phase changes), or for monitoring propagation effects due to flares or Trimpi's (by looking for phase changes over time scales of minutes or seconds, respectively). To extract phase information from the signal it is only necessary to filter it at either the mark or the space frequency using a filter with a suitably narrow passband. An FSK signal can be thought of as the sum of two keyed CW signals, one at the mark frequency sending the marks, and a second one at the space frequency sending the inverse message. Each of these sub-signals can be thought of as an amplitude-modulated signal with half of its power in information-carrying sidebands and half in a constant-power carrier. Thus a coherent FSK signal actually broadcasts one-quarter of its power as a constant signal at the mark frequency, and another quarter at the space frequency. These are the signals we can filter out for phase measurements.

Phase coherence is actually quite important when it comes to transmitting information, and its convenience for our uses is merely accidental. The importance of coherence stems from the improvement in receiver signal-to-noise ratio that it allows. This works as follows: The simplest way to demodulate an FSK signal is to have two bandpass filters, one at the space frequency and one at the mark frequency, each followed by an envelope detector (*i.e.*, a rectifier and a low-pass filter). Whether a given bit is a mark or a space is decided in the receiver according to whether the output of the mark filter envelope detector is bigger than that of the space filter detector, or *vice versa*. This *asynchronous* detection scheme will work with any FSK signal, coherent or not. However, if the signal is phase coherent we can use *synchronous* detection. Instead of an envelope detector, the outputs of the two filters are detected by mixing them with in-phase reference signals at the mark and space frequencies (either from a local oscillator which is phase-locked to the received signal or perhaps just narrow-band filtered signals from the CW sidebands as discussed above) and low-pass filtering the results. This eliminates that half of the received noise which is in phase quadrature with the mark and space sidebands, and improves the signal-to-noise ratio of the receiver by 3 dB over the asynchronous case.

The top plot in Fig. 3.2 shows an FSK signal at approximately 19550/19600 Hz. This is presumed to be from a transmitter in the Soviet Union. Note the slow drift in the relative phase at each frequency. The phases of successive marks and spaces seem to be decreasing at the same rate, about 0.6 revolution per second. A careful count of the keying rate shows it to be within 0.1 baud of 50.0 baud. Both the mark and space phases change together with a differential drift less than 0.1 rev/s, so the transmitter frequency shift is within 0.1 Hz of 50.0 Hz. Thus the modulation index h is equal to 1.000 as close as we can tell, consistent with phase coherent FSK. What is happening is that the average frequencies of both the marks and spaces are 0.6 Hz below the expected frequencies; that is, the transmitter is 0.6 Hz low. (Curiously, note that many of the individual pulses seem pretty close to 19550 or 19600 Hz, as their slopes look flatter than the overall drift in phase.) Of course, there is no reason why the transmitter should not be operating on purpose at 19549.4 and 19599.4 Hz, but it does seem at variance with standard practice. The other presumed Soviet stations, at 15075/15125 and at 19000/19050 Hz, are within 0.01 Hz of the numbers given here, all exact multiples of 25 Hz.

It seems likely that this 0.6 Hz offset is not intentional but instead is due to a slight misadjustment of the transmitter. But if it is not intentional, it says something puzzling about the nature of the transmitting equipment. A frequency offset of 0.6 Hz at 19000 Hz represents a relative frequency error of 3.2×10^{-5}. This is roughly the error that might be expected from a well-designed LC oscillator. However, using an LC oscillator as the master oscillator for a VLF transmitter would be a very crude approach by today's (or even yesterday's) standards. The most likely master oscillator

not recorded well unless fast tape speeds (15 ips) are used. Unless the level of the recorded signal is high, phase and amplitude measurements will be noisy, especially if high time resolution is desired. Because of their short wavelengths, signals at the upper end of the recorder frequency range are subject to increased amplitude variations from tape skew, an effect which adds additional noise.

3. Notch filters have been used at some field sites to eliminate intermodulation caused by particularly strong VLF stations. These stations would otherwise be excellent beacons for Trimpi measurements, but notching them out destroys their usefulness. The notch filter attenuates them, making them more difficult to measure, and the steep skirts of the notch change the frequency modulation of stations sending message traffic into amplitude modulation.
4. Tape speed errors limit the minimum bandwidth of filters that can be used when making analog amplitude measurements. A tape flutter of 0.2% becomes a frequency variation of 40 Hz on a 20 kHz signal. Unless the analysis filter is very flat across the signal bandwidth, tape flutter will be translated into a corresponding amplitude modulation of the signal.
5. Typical Trimpi phase changes, around 1 μs, are about the same size as the noise in phase measurements made with the present digital analysis system, as we will see below.

Figure 3.3 shows magnitude-phase plots of a signal from the Siple Station VLF transmitter as received at South Pole Station. These data were used by *Carpenter et al.* [1985] to evaluate the Siple transmitter as a potential tool for Trimpi studies. The top half of Fig. 3.3 shows magnitude and phase data without complex spectral averaging. The bottom half shows the same data with 100 ms averaging. There are at least five major Trimpi events shown in Fig. 3.3. In this particular case, *f-t* spectrograms show that each event is preceded (and caused) by a whistler.

Figure 3.3 certainly does demonstrate the usefulness of the Siple Station VLF transmitter as a beacon for Trimpi studies. The events seen here show phase advances (decreases in phase path length) of up to 15 μs and amplitude dips of up to 3 dB. Each event develops uniformly during an interval of 2 s, and decays more or less exponentially with a time constant of about 7 s. These events are particularly large and well defined. Whether their size is due to the low frequency of the transmitter (at 3790 Hz less than one-fifth that of the typical VLF station), the relatively simple mode structure of the sub-ionospheric path from Siple to South Pole, or just a lucky circumstance, is unknown.

Current Limitations. However, Fig. 3.3 also points up the limitations of the current digital analysis system for making Trimpi measurements from broadband field recordings. Remember, this is especially good data. The amplitude and phase perturbations are large (3 dB *vs.* the usual 1 dB, and 15 μs *vs.* 1 μs). This is a continuous and not a synoptic recording, so signals of long duration are available. The transmitter frequency, 3970 Hz, is below the stopband frequency of any interference-elimination filters in the receiver at South Pole Station (which is not bothered much by interference anyway). This frequency is also well within the bandwidth of the tape recorder so fidelity is not a problem. The signal is strong. The level of spherics and local interference is low so the signal-to-noise ratio is good. The pilot tone is recorded at its proper level and quite clean. Even so, the noise in the unaveraged output is high. The magnitude trace shows a noise level of about 0.4 dB rms, and the phase trace noise of about 1 μs rms.

Averaging, as shown in the bottom of Fig. 3.3, decreases the noise of the phase trace considerably, to about 0.3 μs rms. Spherics affect the phase randomly, and averaging smooths the phase trace quite nicely. However, it does little for the magnitude trace. Spherics seem to cause blocking

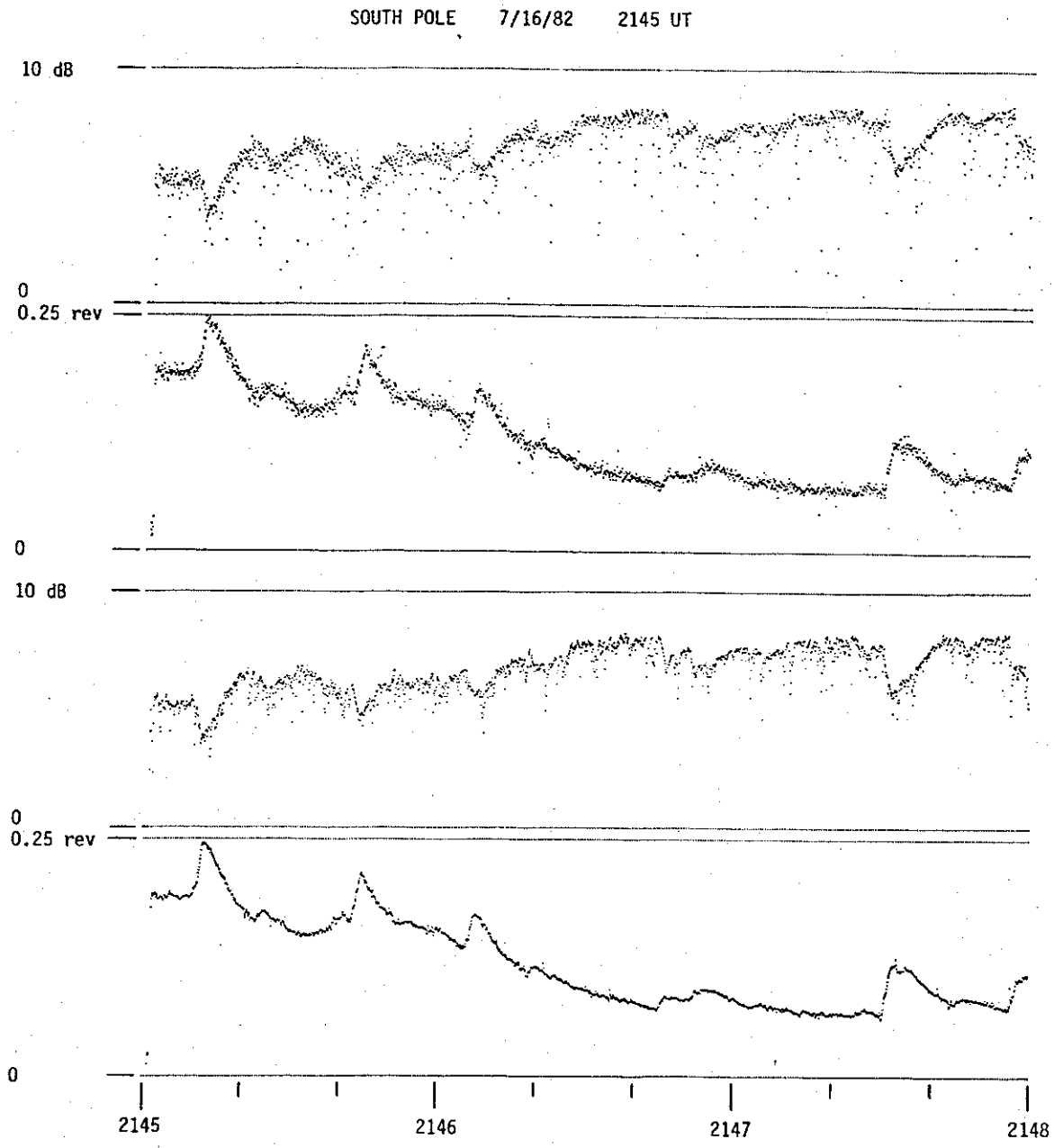

Figure 3.3. Magnitude-phase plots of a 3790 Hz transmission from Siple as received at South Pole showing Trimpi events (BW = 40 Hz, P_{span} = 1/4 rev = 66 μs). Top half: no spectral averaging. Bottom half: averaging time constant τ_{avg} = 100 ms. Whistler-induced electron precipitation into the lower ionosphere causes the sudden amplitude and phase perturbations of at least five Trimpi events. Spectral averaging improves the phase plot but not the magnitude plot. Bottom half after *Carpenter et al.* [1985].

of the signal with the result that the magnitude trace decreases when a spheric occurs, and averaging biases the trace toward these dips and leaves it still rather noisy. The reason for the blocking by spherics in this case is not entirely clear. The analog tape was played back and digitized at a level somewhat lower than usual to eliminate any clipping during the sampling process. Even so, an *f-t* spectrogram shows some intermodulation from spherics on both the 9 kHz pilot tone and the Siple signal, possibly indicating tape saturation.

The record shown in Fig. 3.3 contains the best examples of Trimpi events I have studied with the digital analysis system, primarily because of a strong sub-ionospheric signal at a frequency well within the bandwidth of the analog tape deck recorded at a site with low noise. Yet still the events stand out only because they are so large. I have never had much success trying to measure Trimpi phase changes on recorded signals above 10 kHz, mostly because of poor signal-to-noise ratio. For instance, I have analyzed recordings of Omega Argentina and NAA (at 17.8 kHz) made at Palmer Station in 1978. Noise in the phase plots is such that the minimum observable phase change is about 2 μs. This, of course, is bigger than the typical Trimpi event at Palmer. Even with strong recorded signals, residual timing errors when correcting data for tape wow and flutter seem to be a few tenths of a microsecond. Thus even in the best of cases the current digital analysis system can be only marginally useful. I predict that the measurement of Trimpi events, particularly the measurement of phase, will continue to be an area of VLF research where special-purpose equipment in the field outperforms general-purpose analysis equipment back in the laboratory.

3.3 Duct Motion from Slow Whistler-Mode Phase Changes

Linear Whistler-Mode Propagation. Signals often propagate on whistler-mode paths with little apparent distortion. These signals have received little attention at Stanford compared to those showing amplitude growth, emission triggering, and other non-linear effects, examples of which we will see in Chapter 4. I doubt that we even know what fraction of the time undistorted signals can be heard. Yet these most basic of whistler-mode signals have their own allure and carry some surprising information about the paths they traverse.

Figure 3.4 shows signals from the Siple Station transmitter as received at Roberval during a time of linear propagation. The signals include constant-frequency pulses and ramps with slopes of +500, ±1000, and ±2000 Hz/s. Signals were transmitted from 2030 to 3030 Hz, but those below about 2300 Hz are too weak to be seen. This is due mostly to a fall-off in transmitter power and antenna efficiency at lower frequencies. (These signals were from the older "Zeus" transmitter whose power dropped rapidly with a decrease in frequency.) However, while these signals propagate without distortion through the magnetosphere, there may still be some linear amplification due to wave-particle interactions; our knowledge of the input field strength and our models of ionosphere-magnetosphere signal coupling are not accurate enough to rule this out. If linear amplification is occurring, some of the fall-off below 2300 Hz may be due to decreased magnetospheric gain at lower frequencies. The local time at Roberval is just after dawn and sub-ionospheric propagation is good—seven of the eight Omega transmitters can be heard at 10.2 kHz. Local interference from power lines includes odd-numbered harmonics of 60 Hz up as high as 6540 Hz, the $109th$ harmonic. A high-pass filter was used at the receiver to attenuate power-line interference below 1.5 kHz but some is still present. The pilot tone is at 1 kHz.

Figure 3.5 shows magnitude-phase plots of some two-second constant-frequency pulses received at this time, including the last pulse in Fig. 3.4. Judging by the plateaus on the rising edge of the magnitude plots, there are at least three paths of propagation with different group delays. Each pulse was sent when the clock read xxxx:x0 seconds, so pulse group delay can be read directly from

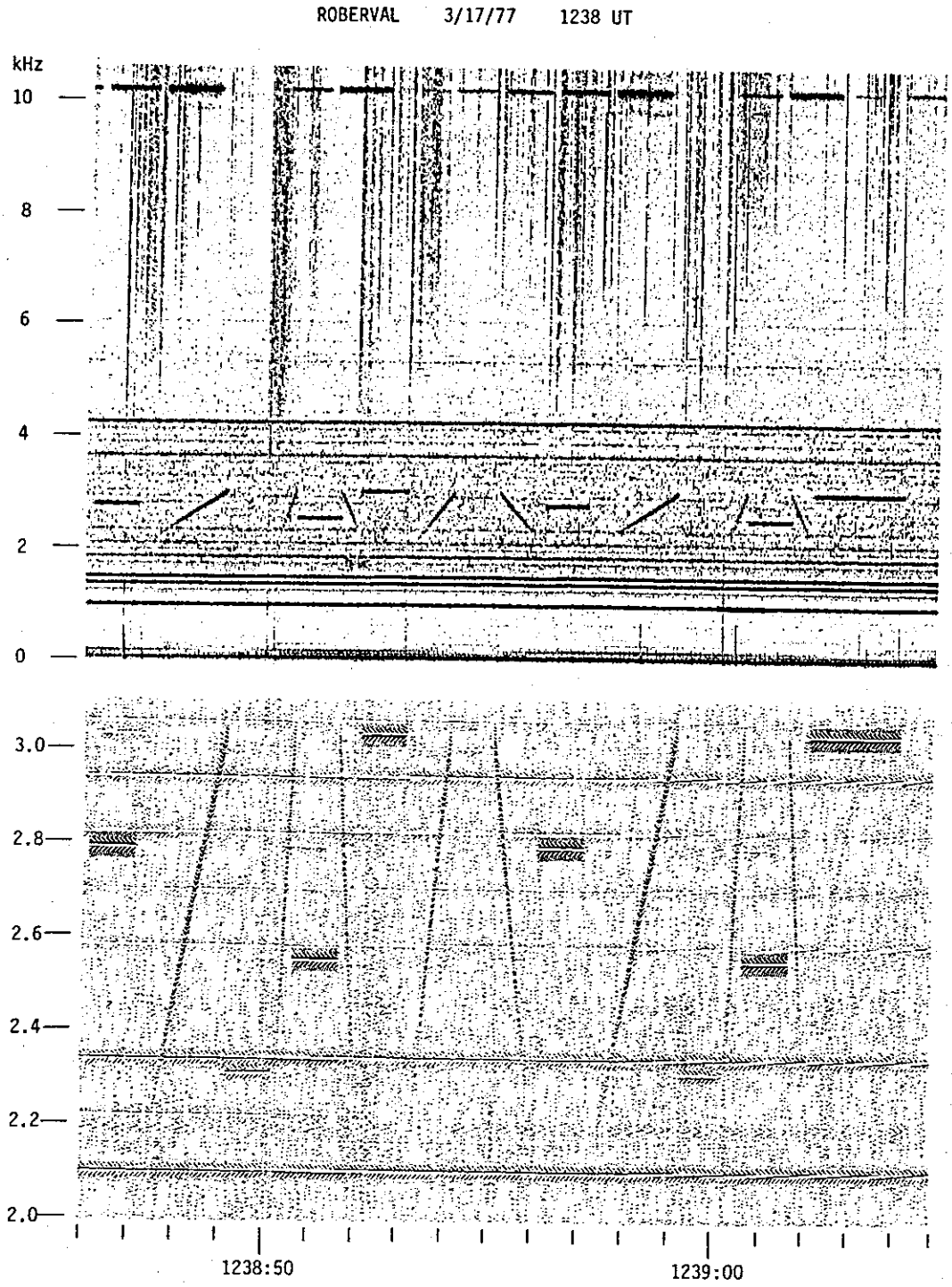

Figure 3.4. Spectrogram and gray-scale phase plot (f_I = 10 Hz, BW = 20 Hz, P_{span} = 1 rev) of signals from the Siple Station transmitter received at Roberval with linear propagation. Constant-frequency pulses and frequency ramps are received without apparent distortion. Local power-line interference extends up to 6540 Hz. Seven of the eight Omega stations are seen at 10.2 kHz.

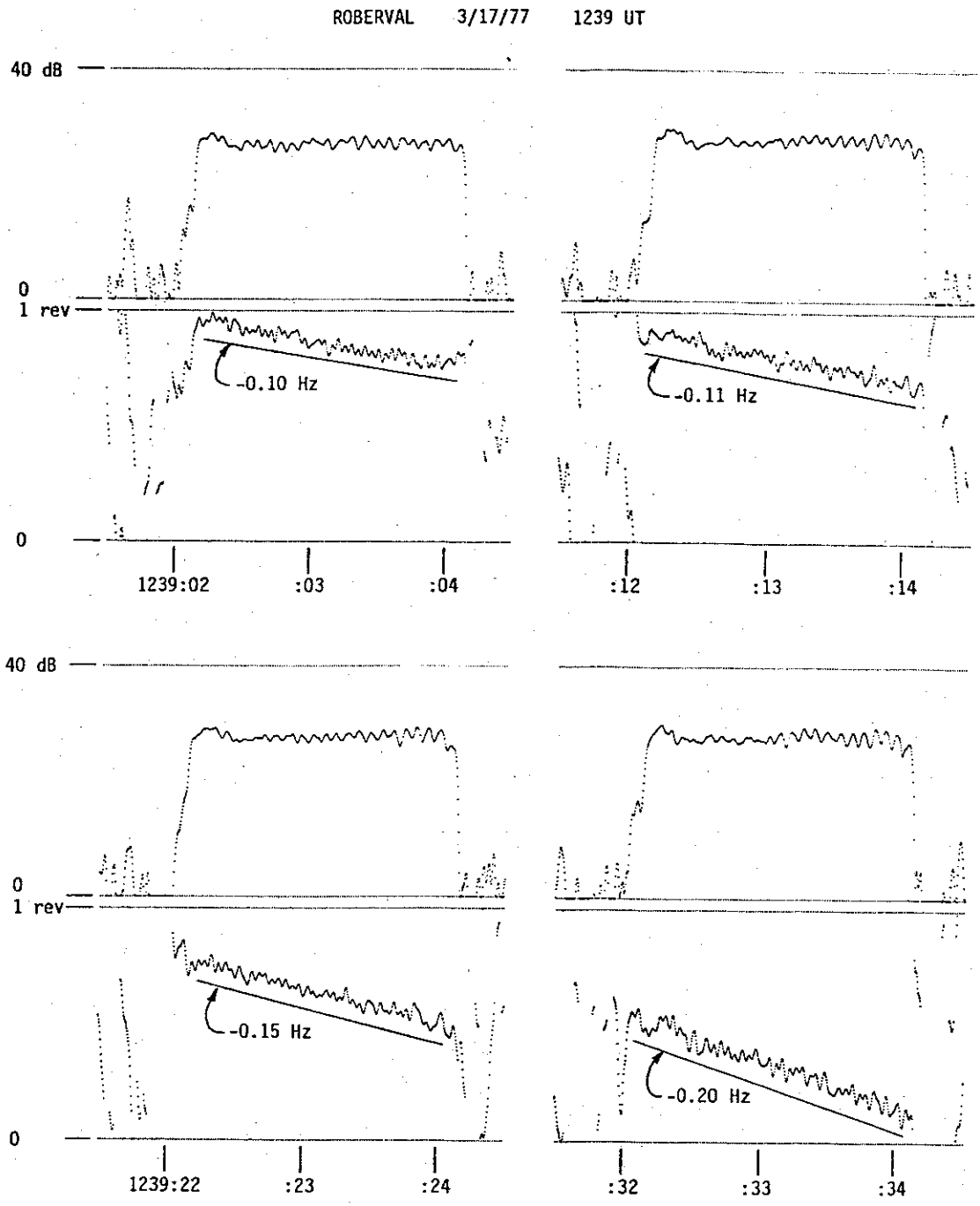

Figure 3.5. Magnitude and phase plots (BW = 20 Hz) of 2-second pulses at 3030 Hz from the Siple transmitter, including the last pulse from Fig. 3.4. Pulses show a Doppler shift Δf from -0.10 to -0.20 Hz mostly due to a gradual increase in the length of the magnetospheric path. Changes in Δf from pulse to pulse are due to changes in the electric field E that causes path drift. Pulse at 1239:22 after *Paschal and Helliwell* [1984].

the time scale. (The error in time between the clocks at Siple and Roberval was at most a few milliseconds.) Paths show group delays of 2.05, 2.10, and 2.15 seconds. The strongest signal is the last one to arrive; it dominates all the others.

Before we discuss Doppler shifts and duct drift, let me mention two other features of the pulses in Fig. 3.5—ripples in magnitude and phase. First, the magnitude plots all show a ripple in amplitude at a frequency near 11 Hz reaching about 2.5 dB p-p by the end of the pulse. As mentioned in *Paschal and Helliwell* [1984], this ripple is similar to (though smaller and more rapid than) the pulsation phenomenon seen by *Bell and Helliwell* [1971] on whistler-mode signals from station NAA. The pulsations cannot be caused by multipath beating since this would require one signal to be offset by 11 Hz from the transmitted frequency, a value much larger than is ever seen with non-growing signals. It is possible that all the signals seen in Fig.s 3.4 and 3.5 are linearly amplified by wave-particle interactions, though we see no evidence of the temporal growth associated with non-linear amplification (cf. Sec. 4.1). A possible cause of the amplitude ripple might be inherent oscillations in the linear amplification mechanism.

Second, the phase plots in Fig. 3.5 also show ripple. There is a fairly steady ripple at 20 Hz and a somewhat irregular ripple at a lower frequency. The 20 Hz ripple in not a natural feature but an artifact of the analysis. It is caused by contamination of the 1 kHz pilot tone with a weak power-line signal at 1020 Hz (the $17th$ harmonic of 60 Hz.) Beating between the two signals causes a fluctuating timing error when the pilot tone phase is measured to calculate the data time and the reference phase at 3030 Hz. (This affects only the phase plots; it has no effect on the magnitude plots.)

An interesting feature of the 11 Hz magnitude ripples is the type of sideband structure in the frequency domain associated with them. That is, are there symmetrical sidebands both 11 Hz above and below the transmitted frequency, or perhaps only an upper sideband, or some combination of the two. To determine this we need to know the size (and timing) of any 11 Hz phase ripples (see Sec. 4.4). Unfortunately, any 11 Hz phase ripple present is masked by the 20 Hz ripple from the contaminated pilot tone, and not much can be said about the nature of the 11 Hz sidebands from Fig. 3.5. This question might be answered directly by a narrow-band *f-t* spectrogram made with the standard analog spectrum analyzer, provided that there isn't too much tape flutter. (The current digital analysis system does not have filters sufficiently narrow to resolve 11 Hz sidebands in a spectrogram.)

Doppler Frequency Shift and Changes in the Phase Path. Note that all of the pulses in Fig. 3.5 show a small decrease in relative phase at a rate from −0.10 to −0.20 rev/s indicating that the received frequency of the pulse was 0.10 to 0.20 Hz below the transmitted frequency of 3030 Hz. This is not an error in the transmitting or recording equipment but a real effect, a Doppler shift in frequency. It indicates that the phase length of the path of propagation is gradually increasing. In the rest of this section we will examine some of the possible causes of this effect.

For ducted propagation in the magnetosphere the refractive index μ is given by [*e.g., Helliwell*, 1965, Ch. 3]

$$\mu = \left[1 + \frac{f_N^2}{f(f_H - f)}\right]^{1/2} \approx \frac{f_N}{f^{1/2}(f_H - f)^{1/2}}, \tag{3.5}$$

where

$$f_N = \frac{e}{2\pi}\left[\frac{N}{\epsilon_0 m_e}\right]^{1/2} = 8.98 N^{1/2}\ [\text{Hz-m}^{3/2}] \tag{3.6}$$

is the *plasma frequency*, which depends only on the local number density of electrons N, and where

$$f_H = \frac{Be}{2\pi m_e} = 2.80 \times 10^{10} B \text{ [Hz/T]} \tag{3.7}$$

is the *electron gyrofrequency* and depends only on the local magnetic field flux density B. In the above, e and m_e are the charge and mass of the electron, and ϵ_0 is the permittivity of free space. For a 3 kHz wave at the top of a duct on a field line at $L = 4$, typical values are $f_N = 150$ kHz and $f_H = 13.7$ kHz, giving a refractive index $\mu = 26.5$. Note that the refractive index at the top of the path is large and the approximation in Eq. (3.5) is quite accurate.

The *phase delay* t_p over a path S is given by

$$t_p = \int_S \frac{1}{v_p} ds = \frac{1}{c} \int_S \mu \, ds \approx \frac{1}{c} \int_S \frac{f_N}{f^{1/2}(f_H - f)^{1/2}} ds \tag{3.8}$$

where v_p is the *phase velocity* of the wave and c the velocity of light. For the sub-ionospheric part of the path μ is close to 1, and the small contribution to the phase delay from this part depends only on the distance travelled. In the ionosphere and magnetosphere μ is generally large, and the contribution to the phase delay here depends not only on the length of the path but on the amount and distribution of plasma N and the strength of the magnetic flux density B (the latter a known function if we assume a dipole magnetic field). As it happens, the total phase delay is determined primarily by conditions in the magnetosphere near the top of the path where the gyrofrequency is lowest and μ is largest.

Phase Path Changes from Duct Drift. *McNeill* [1967] studied whistler-mode signals from VLF station NLK (Jim Creek, WA) at 18.6 kHz as received in Wellington, NZ, and found frequency shifts of about 0.1 Hz, similar to those seen in Fig. 3.5. In this case signals were only received at night, and showed a positive frequency shift indicating a decrease in phase path length. *McNeill* [1967] considered possible changes in the sub-ionospheric path, the path through the ionosphere from 70 to 1000 km altitude, and the magnetospheric path. He concluded that the effect must be occurring predominately in the magnetosphere, presumably from motion of the duct.

Thomson [1976a] studied further the problem of signals from NLK heard in New Zealand. As he points out, there are three factors that could affect the phase delay. First, the magnetic field B along the path could change, affecting the gyrofrequency f_H and thus the refractive index μ. However, this is not the case, at least in general, because it would require magnetic activity much higher than is usually observed on the ground. Second, the electron density N could change, affecting the plasma frequency f_N and thus μ. Finally, the path S could move, affecting the domain of integration of Eq. (3.8). By comparing the change in phase delay with the change in group delay (signal travel time, which has a different dependence on N and B) he was able to show that the slow decrease in path length was due to a duct drifting inward to lower magnetic latitude while the total electron content along the path (tube content) remained constant. This was true, at least, for nighttime paths during magnetically quiet conditions.

The inward motion of the whistler-mode duct described above is caused by azimuthal electric fields in the magnetosphere which are mapped up from the ionosphere. These electric fields induce $\mathbf{E} \times \mathbf{B}$ drift of the plasma in the duct. An east/west azimuthal electric field causes an outward/inward duct motion. (Radial electric fields would cause a duct to drift in longitude, but this is much harder to measure.) *Park and Carpenter* [1978] review the subject of drift and electric fields, particularly fields revealed by whistler measurements. In this technique, the nose frequency of whistlers in a

particular duct are monitored. The nose frequency is proportional to the gyrofrequency f_H at the top of the path, and thus reveals B and the latitude of the duct. Changes in nose frequency over time show the radial motion of the duct and from this is inferred the azimuthal electric field.

Instead of whistlers, whistler-mode signals from VLF transmitters can be used to measure magnetospheric electric fields. Frequency shift can be measured directly from a phase plot such as in Fig. 3.5, giving an instantaneous value of azimuthal electric field. *Thomson* [1976b] has used the nighttime NLK data and measured westward azimuthal fields around 0.2 mV/m on paths at $L = 2.3$, similar to those found by other techniques. The Siple transmitter can be used in a similar fashion for ducts at higher latitudes ($L \approx 4$), and on daytime paths as well. Besides giving instantaneous measurements of electric field, man-made signals may be usable at times when, because of low lightning activity, there are no whistlers.

However, the phase of a constant-frequency signal as in Fig. 3.5 cannot be used to monitor more than one dominant path (at least if the paths have similar drifts), unlike whistlers where several ducts can be monitored because of their different group delays. If multiple paths are to be watched at once, a more complicated scheme of signal modulation and detection is needed. For instance, *Thomson* [1981] describes a technique that uses cross-correlation and filtering of MSK-modulated signals from NLK and which can measure both the group delays and Doppler shifts (not phase delays) of whistler-mode components on multiple paths. Unfortunately, the technique averages data over 15-minute intervals and cannot see more rapid changes. The problem seems to be in getting sufficient signal-to-noise ratio for components on individual paths and sufficient rejection of the strong sub-ionospheric signal. A similar technique should be possible with the Siple transmitter. In fact, it might be easier in some respects since there is little (if any) sub-ionospheric signal to reject at the northern end of the path, and group delay correlations are simpler when we have complete knowledge of the timing of the transmitted signal. Still, the achievable time resolution will be limited by the strengths of individual path components and is unlikely in multipath cases to be as good as that in the predominately single-path examples shown in Fig. 3.5 and below in Sec. 3.4.

Phase Path Changes from Plasma Flux. The studies of Doppler-shifted whistler-mode signals mentioned above concentrated on measuring duct motion and the magnetospheric electric fields that cause it. However, we see from Eq. (3.8) that a change in electron density N can also change the length of the phase path through its effect on the plasma frequency f_N. We might expect that N will change with time, even in the absence of duct drift due to electric fields and the convective motions caused by magnetic storms, because of the flow of plasma between the ionosphere and magnetosphere. During the day, solar irradiation ionizes plasma in the upper atmosphere which then diffuses up along magnetic field lines into the magnetosphere. During the night, plasma diffuses back from the magnetosphere to sustain the the ionosphere which would otherwise be depleted through the recombination of ions and electrons. We thus expect, in the absence of other factors, that the electron density N at a given location in the magnetosphere will increase during daylight hours and decrease at night. The question is to estimate how much this will affect the phase delay compared to, say, duct drift. In the following discussion we will find that plasma flux into the magnetosphere might account for at most 0.043 Hz of the 0.10–0.20 Hz Doppler shift seen in Fig. 3.5.

It is convenient here to introduce the concept of *tube content* N_T, the total magnetospheric electron content of a tube extending along the magnetic field from the top edge of the ionosphere to the equatorial plane. The tube is conventionally taken to start at an altitude of 1000 km and to have there a cross-section of 1 cm^3. Above 1000 km altitude the cross-section of the tube increases

in inverse proportion to the magnetic flux density, and so we can express the tube content as

$$N_T = \int_{1000\text{ km}}^{\text{equator}} \frac{B_{1000}}{B} N\, ds \tag{3.9}$$

where B_{1000} is the field at the 1000 km point.*

The importance of N_T is as follows. From Eq. (3.6) we know that the plasma frequency f_N is proportional to $N^{1/2}$. If the refractive index μ is large so the approximation of Eq. (3.5) is valid, then μ is also proportional to $N^{1/2}$. For a given path, we see from Eq. (3.8) that the phase delay t_p is proportional to $\int N^{1/2} ds$. Finally, if the relative density of electrons along the path retains the same form (as a function of s) as plasma diffuses between the ionosphere and magnetosphere, then we see that $t_p \propto N_T^{1/2}$. In particular, as *Andrews et al.* [1978] have noted, from this we can calculate the rate of change of phase delay for a given plasma flux as

$$\frac{dt_p}{dt} = \frac{t_p}{2N_T} \frac{dN_T}{dt}. \tag{3.10}$$

In order to apply Eq. (3.10) to the pulses shown in Fig. 3.5 we need to know the total phase delay t_p at 3030 Hz. When we examine the relative phase in Fig. 3.5 we can see *changes* in t_p, but there is no direct way to measure its total value.† However, we can measure the signal travel time or *group delay* t_g. This is the time it takes a signal (a pulse, say) to travel from the transmitter to the receiver, and, as mentioned above, is 2.15 s for the pulses in Fig. 3.5. In Appendix B it is shown that the ratio t_p/t_g depends, to a close approximation, only on the ratio of the signal frequency to the equatorial gyrofrequency f_{Heq} for the given path, and not on the path latitude or tube electron content. Thus, knowing the group delay t_g and either the nose frequency (which is about $0.37 f_{Heq}$) or the path latitude (which also determines f_{Heq}) we can find the total phase delay t_p. In the case of the signals in Fig. 3.5 we will assume the path to be at $L = 4.2$, typical of Siple signals, which gives $f_{Heq} = 11791$ Hz. The signals at 3030 Hz are then at $0.26 f_{Heq}$, and from Appendix B we find $t_p/t_g = 1.69$, giving a phase delay of $t_p = 1.69 \times 2.15 = 3.63$ s.

The tube content N_T can be estimated from the group delay at the nose frequency using the techniques given by *Park* [1972]. The nose frequency for a duct at $L = 4.2$ is 4363 Hz. Calculations show that if the group delay at 3030 Hz is 2.15 s, then the group delay at the nose frequency (t_n) is 2.06 s. From *Park* [1972] (ignoring corrections for ionospheric dispersion) we find the tube content to be $N_T = 3.8 \times 10^{13}$ electrons/cm^2.

Finally, we need an estimate of the rate of change of tube content dN_T/dt. *Park* [1970], using whistler group delays to monitor tube content over a period of several days in magnetically quiet conditions, determined that the upward flux across the 1000-km level for paths in the range $L = 3.5$–5 was about 3×10^8 electrons/cm^2-s in the daytime (and about 1.5×10^8 electrons/cm^2-s downward at night). At the time of the pulses shown in Fig. 3.5 both ends of the path are in morning sunlight.

---

* There is some confusion in the literature over the length of the tube. *Andrews et al.* [1978] and *Andrews* [1980] consider the tube to extend from one hemisphere to the other, which gives a value of N_T twice that in Eq. (3.9). I have adopted the convention used by *Angerami* [1966] and *Park* [1972].

† We can imagine sending a signal whose frequency starts at zero and slowly increases until it reaches the frequency of interest. The difference at that time in signal cycles sent from the transmitter and those counted at the receiver is the total phase delay (in revolutions) at that frequency. Unfortunately, this is not a practical method.

However, the solar elevation angles are not especially large, 15.5° at Roberval (48.4° N, 72.3° W) and 4.5° at Siple (75.9° S, 84.2° W), and the rate of photoionization (proportional to the sine of the solar elevation) is probably less than its average daytime value. Still, we will estimate the flux of plasma into the magnetosphere from each ionosphere as 3×10^8 electrons/cm^2-s, realizing that this probably overestimates the true state of affairs and will give us an upper bound on the rate of change of t_p.

Using the above values in Eq. (3.10), we estimate the rate of change of phase delay t_p due to the flow of plasma into the magnetosphere to be $dt_g/dt = (t_g/2N_T)dN_T/dt = 3.63 \times 3 \times 10^8/(2 \times 3.8 \times 10^{13}) = 1.43 \times 10^{-5}$ s/s. For a 3030 Hz signal this gives a frequency offset of -0.043 Hz. As noted above, the actual plasma flow may be less and the Doppler shift it causes may be smaller. However, even comparing the upper bound of -0.043 Hz to the -0.10 to -0.20 Hz shifts seen in Fig. 3.5, we must conclude that changes in electron density can account for only a small part of the changes in phase delay that are actually observed.

With knowledge only of the change in phase delay with time we cannot conclusively separate those changes due to duct motion from those due to plasma flux, though the latter are probably of minor importance as we have just seen. If we have knowledge of the change in group delay with time as well, then we might hope to separate the two because of their different effects on t_p and t_g.* *Andrews et al.* [1978] and *Andrews* [1980] used the New Zealand NLK data to try to measure the flux into the nighttime ionosphere, and found downward flows on the order of 1–2×10^{-8} electrons/cm^2-s. While the results are consistent with those of *Park* [1970] (who used the whistler method), the uncertainties are nearly as large as the measured values, and the technique is of marginal value. The errors they encountered were primarily errors in measuring changes in t_g when averaging over intervals of less than 90 minutes, and errors integrating Δf to get changes in t_p over time intervals longer than 90 minutes. Using relative phase ϕ instead of Doppler shift Δf, as we did in Fig. 3.5 above, may alleviate the latter error (though the data processing would have to be greatly streamlined, perhaps by measuring phase directly at the receiver). However, the problem of accurately measuring changes in group delay remains. *Andrews'* [1980] error in measuring changes in t_g was about ± 1.5 ms. For comparison, the error in estimating (relative) t_p from Fig. 3.5 is 150 times better, about 10 μs (0.03 rev at 3030 Hz). This suggests that given a better signal we might be able to improve the measurement of t_g as well. Perhaps some scheme can be invented that uses the phase of a broad-band signal, say a series of frequency ramps spanning a range of 1 kHz, to directly measure changes in t_g.

* As mentioned in Appendix B, a change in tube content N_T affects the phase and group delays in such a way that the ratio t_p/t_g remains constant (at least for paths inside the plasmapause, and assuming the relative distribution of N along the path remains the same). A change in duct position due to radial motion, on the other hand, changes the equatorial gyrofrequency f_{Heq} and, as Fig. B.2 shows, changes the ratio t_p/t_g.

3.4 Correlation of Phase Changes with Magnetic Micropulsations

In this section we will compare equatorial duct motion, as deduced from phase path measurements, with variations in the magnetic field at the base of the duct. To do this we need to monitor the phase of a whistler-mode signal for a relatively long time, for many minutes at least, instead of just catching a glimpse here and there as from the short pulses in the previous section. The obvious approach is to transmit a constant-frequency whistler-mode signal and measure its relative phase (and the local magnetic field) at the other end of the path. Unfortunately, a single-frequency signal is very likely to show amplitude growth caused by wave-particle interactions. The phase advance associated with this growth (described in Section 4.1) will completely mask any smaller phase changes due to duct motion. However, if we transmit two signals simultaneously about 20 to 30 Hz apart, then we find that each will tend to suppress the growth of the other. The result in this case is often a stable signal without growth-related phase advance, suitable for duct motion studies. We may still see other effects of wave-particle interactions, such as sideband generation, but these do not appear to affect the phases of the two transmitted components.

Two-Tone LICO1 Format. The spectrogram in the top half of Fig. 3.6 shows the start of a LICO1 transmission from the Siple transmitter as received at Roberval. The LICO1 (LIne COupling, version 1) format contains 10.5 minutes of two-tone transmission with 30 Hz separation, followed by some 2.5 minutes of LICO-type format (10-second segments with single tones and tone-pairs with various separations). The important section for us is that first 10.5 minutes. The two-tone signal starts at 1606:03.1 (1 second late?) and continues for 10m29s until the LICO format begins at 1616:32.1. The two-tone signal fades in and out with a 10 to 30-second period, but it is always visible. Though the signal is strong, it does not show any spectral broadening or trigger any emissions, characteristics associated with particle-induced growth and phase instability. There are frequent whistlers (somewhat compressed in Fig. 3.6) with one-hop delay times from 2.1 s up (also weak ones from 2.0–2.1 s) and nose frequencies near 3.5 to 4 kHz. Some of these are seen to generate precursors (see Sec. 4.5.3) on the LICO1 signal (for example, at 1607:34), but these cause only a momentary perturbation. (The spectrogram in Fig. 3.6 also shows a weak 1.5 Hz pulsation; this is an artifact of the analysis due to aliasing of the 30 Hz beat in the signal by the output plotting time $t_{step} = 0.1906$ s.) We will assume that the Siple signal is propagating on the same path as the first strong component of the nose whistlers. Using a nose frequency of $f_n = 4000$ Hz and a nose group delay of $t_g = 2.1$ s, we find from *Park* [1972] a path L-value of $L = 4.32$ and a tube content of $N_T = 3.5 \times 10^{13}$ electrons/cm^2 (or equatorial density $N_{eq} = 280$ electrons/cm^3, a typical value for magnetically quiet conditions).

The gray-scale phase plot in the bottom half of Fig. 3.6 shows the phase behavior at the beginning of the LICO1 signal. Both the 3950 and 3980 Hz components in the transmitted signal show the slow phase changes characteristic of duct motion. The maximum Doppler frequency is about ± 0.5 Hz. Both transmitted components show similar, though not identical, phase behavior. Each signal shows periods of fading, sometimes with 1/2-rev jumps in phase. A good example is at 1607:37 on the 3980 Hz component. The fading and phase jumps are probably caused by beating between signals propagating in different ducts. Note that the amplitude nulls and phase jumps occur at different times at the two frequencies. We will use this fact later on to remove the phase effects of fading. There is a small blip in the phase of the 3980 Hz tone due to the whistler precursor at 1607:34, but its effect is only momentary.

There are weak sidebands 30 Hz below and above the transmitted signals (at 3920 and 4010 Hz) present through most of the 10m29s two-tone transmission. Weaker sidebands 60 Hz above and below

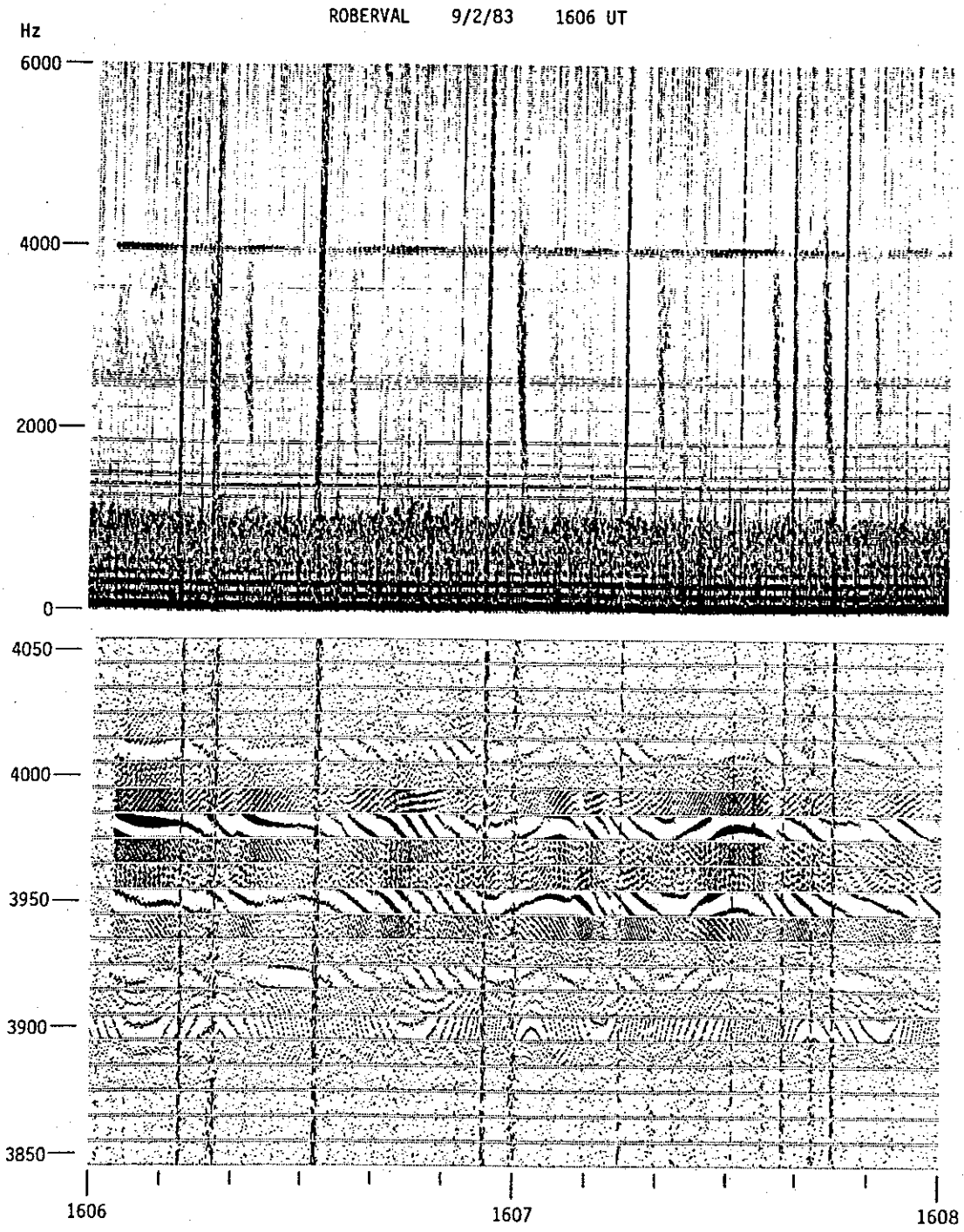

Figure 3.6. Spectrogram and gray-scale phase plot (f_I = 10 Hz, BW = 20 Hz, P_{span} = 1 rev) of a two-tone LICO1 transmission from Siple Station as received at Roberval. The LICO1 tones at 3950 and 3980 Hz begin at 1606:03 UT. The transmitted tones show phase changes due to motion of the magnetospheric path. Also present are sidebands at 3920 and 4010 Hz, and local power-line interference near 3900 and 4020 Hz.

(3890 and 4040 Hz) are sometimes also seen (though not in Fig. 3.6). The phases of the sidebands track those of the main signals. Sidebands indicate that some non-linear process is occurring, and are evidence of wave-particle interactions. There are also signals at 3900 and 4020 Hz. These are power-line signals at the $65th$ and $67th$ harmonics of 60 Hz picked up locally at the receiver, and will not interfere with the analysis. As expected, their phase variations have no correlation with the Siple signal phases.

Evidence Against Particle-Induced Phase Effects. Before we proceed, we must be sure that the phase changes we see in Fig. 3.6 are due only to changes in the effective length of the path of propagation (the effect under study) and are not corrupted by spurious effects due to wave-particle interactions. What would such spurious effects look like? The most common evidence of wave-particle interactions, as we will see in Section 4.1, is the growth in amplitude of an input signal accompanied by a simultaneous advance in relative phase. Growth and phase advance almost invariably occur together. Amplitude decline also seems to be associated with phase retardation, but the evidence here is not as complete. (Examples of the latter are less common because growing signals tend to terminate by generating emissions rather than by slowly dying away.)

Do we see any correlation between amplitude and phase in Fig. 3.6 that might indicate growth-related phase effects? The answer is no. Changes in amplitude seem to occur independently of changes in phase (except for the rapid phase changes due to fading that always occur at amplitude nulls). We see amplitude growth when the relative phase is retarding, as at 1606:20 and 1606:40, as well as when the relative phase is advancing, as at 1607:25. Those overall changes in phase that occur simultaneously on both transmitted signal components seem to be unrelated to signal amplitude and therefore are likely to be due only to changes in the path of propagation.

However, the phase variations of the two input signal components are not identical. While the common large-scale phase changes may not be due to particle effects, the possibility remains that the differences in behavior at the two frequencies are still due to wave-particle interactions. This wouldn't be as serious, but it might still color our interpretation of phase as a measure of path length. To test this possibility I analyzed the record by plotting the amplitude and phase of the 3950 Hz component while tracking on the 3980 Hz one instead of the pilot tone (plot not shown). Tracking on the upper signal removes those phase variations common to both components so the plot at 3950 Hz shows only the differential phase behavior. If there were any residual particle-induced phase effects we might expect this differential phase to be correlated with the difference in signal amplitudes. In fact, what I found is as follows: The transmitted signals fade slowly, with a period of 5 s or more, often independently on each frequency, sometimes with deep nulls. The differential phase (*i.e.*, $\phi_{3950} - \phi_{3980}$) shows changes around 1 rev, sometimes correlated with the difference in amplitudes at the two transmitted frequencies (consistent with phase advance caused by the growth of one signal with respect to the other) but just as often not correlated or even anti-correlated. I conclude that the differential phase changes are not due to differential growth but most likely are just the result of multipath interference.

Removing Fading and Other Artifacts. The next step in the analysis is to extract the phase information from the LICO1 signal and put it in a form suitable for further processing. To do this, the Eclipse spectrum analysis program was run in its "XOUT" mode. In this mode the normalized and averaged spectral data ($\{V_k\}$ from Eq. (2.47)) are written to an output file as a series of 16-bit integers instead of being scaled and plotted. Two filters were synthesized, at 3950 and 3980 Hz, each with a $3rd$-order window response and a 3-dB bandwidth of 20 Hz. One output point (a complex number in (x, y) format) was written for each frequency every 250 ms. (The data segment length was

$NT = 80$ ms. A data step time $t_{step} = 41.66\bar{6}$ ms was actually used to overlap segments by almost 50% and ensure accurate tracking of the pilot tone. Only every sixth analyzed segment was output.) No averaging was used. The magnitude $A = (x^2 + y^2)^{1/2}$ of each filter output at a given time is the amplitude of the corresponding component, and the phase $\phi = \arctan(y/x)$ is the relative phase of that component, revolutions at the center frequency of the filter having already been subtracted. At this point, the spectral data were transferred to a Hewlett-Packard Vectra computer for further processing.

Figure 3.7 shows the relative phases, ϕ_{3950} and ϕ_{3980}, of the LICO1 signal as filtered at 3950 and 3980 Hz, as well as the combined phase ϕ_c (labeled "Weighted Sum") derived from them, calculated as follows:

1. Let (x_1, y_1) and (x_2, y_2) be the outputs of the filters at 3950 and 3980 Hz, respectively, at a given time. Let $A_1 = (x_1^2 + y_1^2)^{1/2}$ and $\phi_1 = \arctan(y_1/x_1)/2\pi$ be the magnitude and phase (in revolutions) of the first filter, and similarly with A_2 and ϕ_2. Let ϕ_1' and ϕ_2' be the previous phase values (250 ms earlier).
2. Calculate the finite differences $d\phi_1 = \phi_1 - \phi_1'$ and $d\phi_2 = \phi_2 - \phi_2'$, representing the advance in relative phase (in revs) from the previous output sample for each filter. These differences will normally be quite small, since the signal phase does not change much over an interval of 250 ms. We accomodate the movement of signal phase across the ends of the range of our arctan function by incrementing or decrementing the difference $d\phi_i$ by exactly one rev to keep it in the range $-0.5 \leq d\phi_i < 0.5$. That is, as ϕ_i increases 0.8, 0.9, 0.0, 0.1, we calculate $d\phi_i$ as 0.1, 0.1, 0.1, and not 0.1, −0.9, 0.1.
3. Calculate the filter relative phases including whole revolutions by integrating the differences as

$$\begin{aligned} \phi_{3950} &= \phi'_{3950} + d\phi_1 \\ \phi_{3980} &= \phi'_{3980} + d\phi_2, \end{aligned} \tag{3.11}$$

   where ϕ'_{3950} and ϕ'_{3980} are their previous values.
4. Calculate the combined phase ϕ_c as the advance of phase at each frequency weighted by the amplitude of that component. That is, let

$$\phi_c = \phi'_c + \frac{A_1 d\phi_1 + A_2 d\phi_2}{A_1 + A_2}, \tag{3.12}$$

   where ϕ'_c is its previous value.

The calculation of the relative phases ϕ_{3950} and ϕ_{3980} in Eq. (3.11) above is similar to the whole-revolution phase accumulation during spectrum analysis described in Sec. 2.5.7, though in the present case we will have some insignificant round-off errors whereas the algorithm in Sec. 2.5.7 was exact (whole and fractional revolutions being processed separately). The results are a series of samples every 250 ms running from an initial value of 0 in both cases to around −90 revs at the upper frequency and −100 revs at the lower at the end of the two-tone transmission. These results are plotted modulo 20 revs in Fig. 3.7.

The ϕ_{3950} and ϕ_{3980} plots in Fig. 3.7 contain artifacts of several kinds. Those marked "F" are due to fading, and sometimes include a 1/2-rev jump in phase, at other times just a momentary change. Fading jumps occur as the amplitude of a signal component goes through a null, and happen at different times at the two frequencies. Now we can see the utility of the combined phase ϕ_c given by Eq. (3.12). Since fading jumps occur at amplitude nulls, the combined phase, which integrates

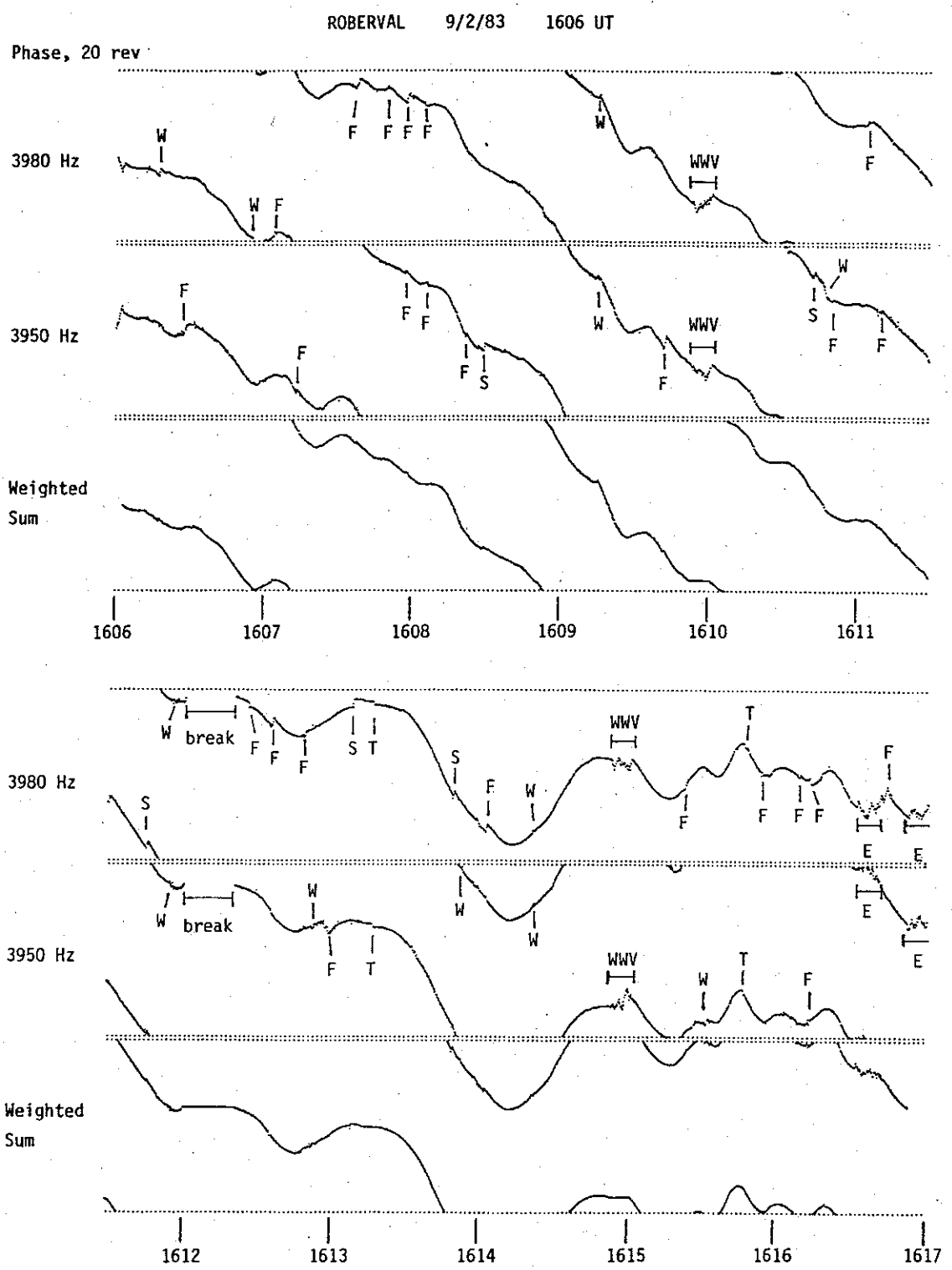

Figure 3.7. Phase of the received LICO1 signal. The plots at 3950 and 3980 Hz are the actual signal components (BW = 16 Hz, $P_{span} = 20$ rev). The plot "weighted sum" is the combined phase ϕ_c corrected for fading, data breaks, and tracking errors. Data features: E emissions; F fading; S spheric; T pilot tracker error; W whistler and/or whistler precursor; WWV time signal break; *break* 20 s break in data.

phase differentials weighted by amplitude, will tend to follow the phase of the stronger component and not show the effects of fading. The combined phase ϕ_c should give us a clearer, less noisy picture of actual changes occurring in the magnetosphere. We can, for practical purposes, regard ϕ_c as the relative phase of a single-frequency signal, at the mean frequency 3965 Hz, which has propagated on a single whistler-mode path without distortion by wave-particle interactions.

The features marked "S" and "W" are phase jumps due to spherics and whistlers or whistler precursors. To the extent that these are more detrimental to weaker signals their effects will also be reduced in the combined phase. The effect of one spheric, at 1611:44.5, was removed by hand from ϕ_c by forcing the phase to remain constant for one sample (250 ms) at that time. Two pilot tone tracker errors, labeled "T" at 1613:16.5 and 1615:46.5, were similarly explicitly removed.

The features marked "WWV" are eight-second breaks in the data during which the audio signal from a shortwave receiver is recorded. There is no LICO1 signal on the tape during these intervals and the phase behavior observed is spurious. There is also a mysterious 20-second interval of missing data labeled "break." The cause of this omission is unknown, but it may be due to an equipment malfunction or an operator error that momentarily stopped the field tape recorder. Both WWV breaks and the missing data interval were removed from ϕ_c by holding its phase constant for the appropriate time. This is a simple way to fill in missing data, of course, but it can also introduce artifacts. The reader may judge whether a straight line of some different slope, or perhaps a spline curve, might have better bridged the gaps.

Finally, the two features near the end marked "E" are emissions from the first two single-tone segments of the following LICO transmission format. The 10-second segment between then was a two-tone LICO segment, also at 3950/3980 Hz, and I desired to include it with the preceding data. I ignored the first set of emissions since they didn't seem too noisy, and terminated ϕ_c at the start of the second set. The total sequence of phase samples ϕ_c runs from 1606:03 through 1616:52.5.

Magnetometer and Phase Data Compared. Figure 3.8 shows the H and D components of the earth's magnetic field as measured by the Bell Laboratories magnetometer at La Tuque, Quebec (about 130 km south-southwest of Roberval), and the LICO1 combined phase ϕ_c from Fig. 3.7. H is the horizontal component in the magnetic north-south direction and D is the component in the east-west direction. The Z (vertical) component was not available at this time. The magnetometer data are a series of samples taken at two-second intervals. Values at quarter-second intervals were interpolated to match the phase data by assuming that each magnetic field component was constant during the two-second interval centered on the actual sample. The initial value of H in Fig. 3.8 (including the static part of the earth's field) is 687.8 nT, and the initial value of D is 10.8 nT. The timing error in the magnetometer data is not known but is unlikely to exceed a few seconds. Timing errors in the VLF phase data are at most a few milliseconds and can be ignored.

There are two different types of behavior seen in Fig. 3.8—slow drifting with a scale time of many minutes, and smaller ripples with periods in the range of 20 to 40 seconds. All three signals show somewhat similar long-term behavior. The magnetic field changes slowly, H increasing and D decreasing, over a range of about 10 nT, and the VLF phase decreases (phase path becomes longer) by about 100 rev (25.2 ms), until 1614 UT. At that time, H reverses its behavior and starts to decrease, and D and ϕ_c stop decreasing and remain more or less constant.

Long-term changes in magnetic field strength and VLF phase may or may not have much connection with each other. We know there are some mechanisms which can cause a change in phase path length without causing any change in the earth's magnetic field. For instance, the change in phase path may represent $\mathbf{E} \times \mathbf{B}$ drift of the duct due to magnetospheric electric fields

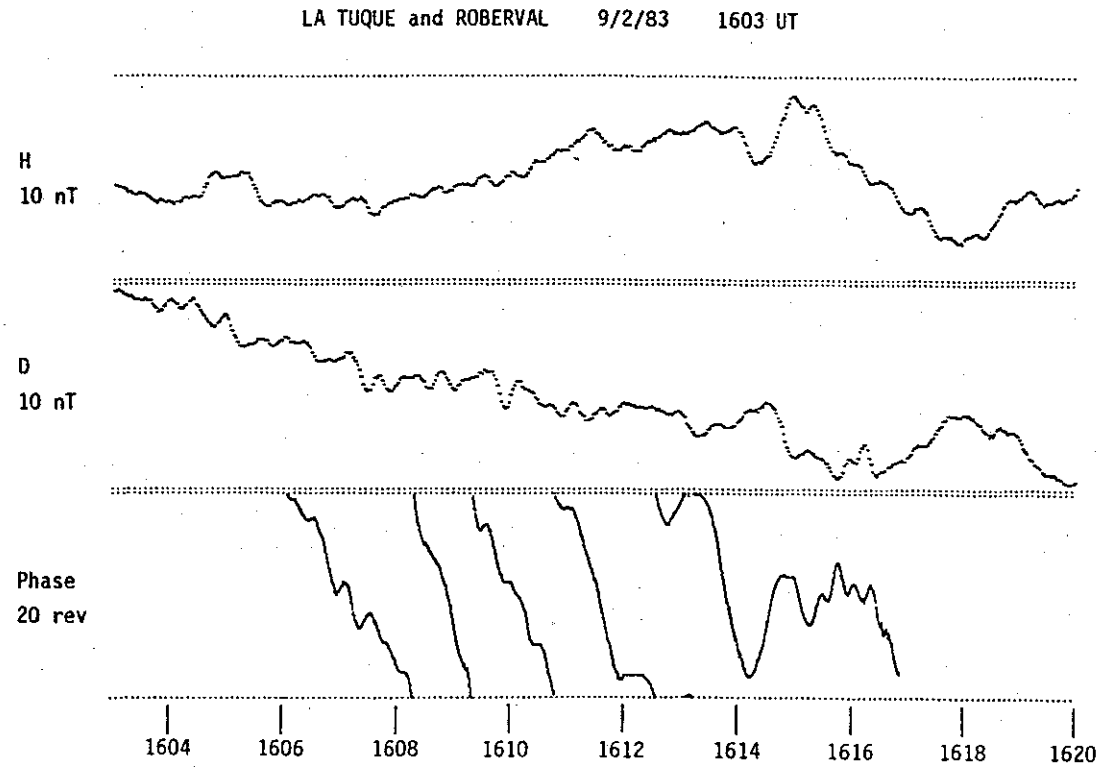

Figure 3.8. La Tuque magnetometer components H and D (10 nT = 10 γ full-scale), and Roberval VLF phase ϕ_c (20 rev full-scale). All three signals show small fluctuations with periods of 20–40 s on top of larger changes occurring over many minutes.

as described in Section 3.3, and need not imply any corresponding change in magnetic field. On the other hand, the local time at both ends of the magnetospheric path is shortly before noon, and both ionospheres are in sunlight (solar elevation only 5° at Siple). The lengthening of the phase path (decrease in ϕ_c) may merely reflect increasing plasma density in the magnetosphere due to diffusion up from the sunlit ionosphere, again not connected with magnetic field changes. From Appendix B we find that a signal on a path at $L = 4.32$, at a frequency of 3965 Hz and with a group delay of $t_g = 2.1$ s, has a total phase delay of $t_p = 3.23$ s. As noted above, the tube content is $N_T = 3.5 \times 10^{13}$ electrons/cm^2. Using an upward flux $dN_T/dt = 3 \times 10^8$ electrons/cm^2-s (*Park's* [1970] typical value) we find from Eq. (3.10) a Doppler shift of -0.055 Hz. This would account for one-third of the -100 revs/11 min $= -0.152$ Hz average shift in Fig. 3.8. In any case, we need more data than that given in Fig. 3.8 before drawing any conclusions about long-term correlations between magnetic field and path length.

The faster variations seen in Fig. 3.8 are oscillations with periods of 20 to 40 seconds and peak values of one or two nT/revs. The magnetic variations are classed as Pc 3 micropulsations.* Over

* Regular pulsations are classed according to their periods as follows [*Jacobs et al.*, 1964]:

Pc 1	0.2–5 seconds
Pc 2	5–10 seconds
Pc 3	10–45 seconds
Pc 4	45–150 seconds
Pc 5	150–600 seconds

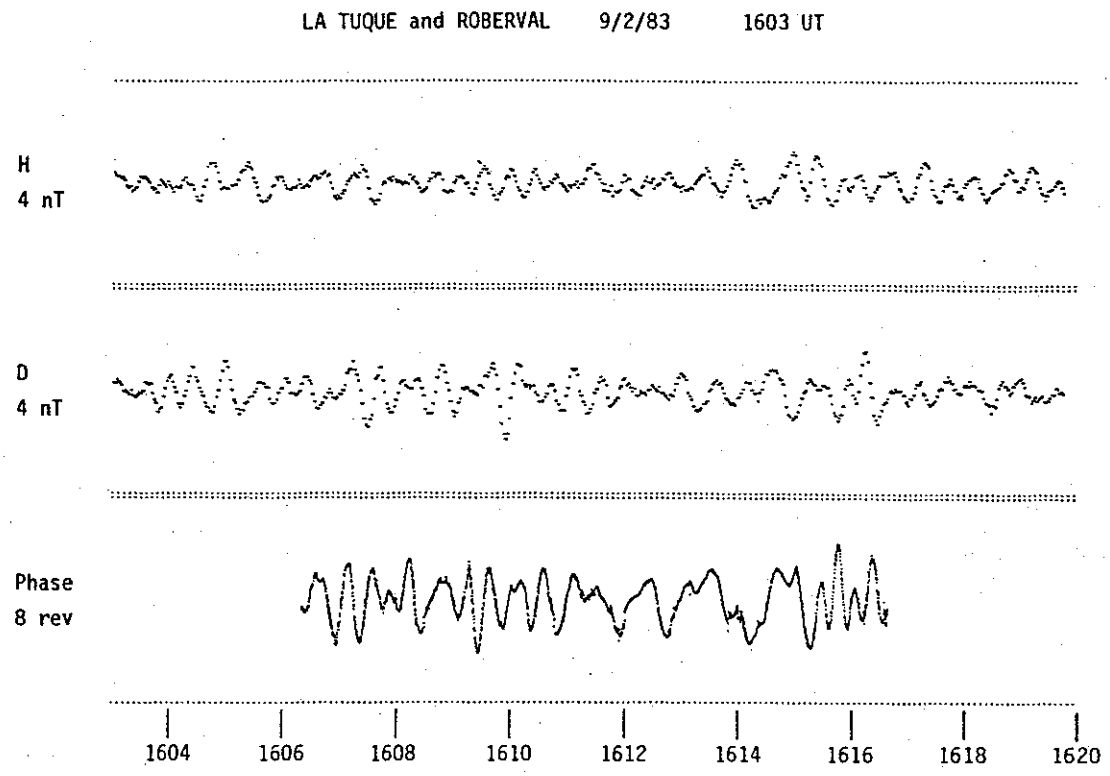

Figure 3.9. La Tuque magnetometer components H and D, and Roberval VLF phase ϕ_c from Fig. 3.8, high-pass filtered by subtracting a running average with a 32 s time constant. Note the change in scale to 4 nT/8 rev full-scale. The long-period variations have been filtered out and the Pc 3 micropulsations are more obvious.

the short period of the micropulsations, plasma flow between the ionosphere and the magnetosphere is negligible and phase variations on this time scale cannot be due to changes in tube content. However, micropulsations seen on the ground correspond to changes in magnetic field extending up into the magnetosphere. We would expect these to affect the phase delay.* *Andrews* [1977] has studied this question and finds that phase length will be affected by radial motions of the duct near the equator caused by micropulsations, but will be unaffected by azimuthal motions. An inward displacement of a duct will give an increase in the north-south magnetic field at the base of the duct and a decrease in phase path length, that is, an increase in relative phase. We thus expect to see a direct correlation between the north component of field at the bottom of the duct and ϕ_c.

To eliminate the large, long-term variations that obscure the micropulsation activity, all three data sets were high-pass filtered. The filter used was simple but crude: a block average in a 32-second interval centered at each sample point was subtracted from that sample. Since the block average contains mostly low-frequency components, the remainder after subtraction has most of those components removed. In fact, if the input sequence $\{x_n\}$ with sampling time T has a discrete

* This may not be quite as obvious as it seems. For VLF frequencies well below the gyrofrequency f_H, the refractive index μ in Eq. (3.5) is approximately $f_N/(ff_H)^{1/2}$, which is proportional to the ratio $(N/B)^{1/2}$. This ratio remains constant since plasma remains trapped on field lines during rapid changes in magnetic field. Changes in phase path are due to changes in duct position rather than just changes in B.

Fourier transform $X_D(f)$ (Eq. (2.13)), then the filtered sequence $\{y_n\}$ calculated by

$$y_n = x_n - \frac{1}{2Q+1}\sum_{i=-Q}^{Q} x_{n+i} \tag{3.13}$$

has a spectrum given by

$$Y_D(f) = X_D(f)\left[1 - \frac{\sin[\pi f(2Q+1)T]}{(2Q+1)\sin(\pi fT)}\right]. \tag{3.14}$$

For $(2Q+1)T = 32$ s, we find that components with 10-minute periods or longer are attenuated by 46 dB or more, whereas components with periods of 40 s or less are attenuated by at most 2 dB. (The filter has about 3 dB of passband ripple, not very important in this particular case.)

Figure 3.9 shows the filtered H, D, and ϕ_c signals. The long-period variations have been eliminated, leaving only the Pc 3 micropulsations. The signals look fairly complicated, and contain energy over at least an octave in frequency judging by the various periods involved. There may be some correlation between the waveforms, particularly between D and ϕ_c, but how much is not very clear. The rms values of the signals (from 1606:19 to 1616:36.5) are $H = 0.200$ nT, $D = 0.282$ nT, and $\phi_c = 0.837$ rev (211 μs). (Note that the duration of the phase signal has been shrunk by 16 seconds at each end from 1606:03–1616:52.5 in Fig. 3.8 to 1606:19–1616:36.5. Without extending the original data set there is no way to filter it at the ends since averaging would involve nonexistent data points.)

In order to better visualize the structure of the filtered signals, the filtered sampled waveforms were transferred back to the Eclipse system and *f-t* spectrograms were made. Figure 3.10 shows the results. Each spectrogram is heavily oversampled, of course. Data segments are overlapped by about 94%, and almost eight filters are synthesized for each DFT filter. We are trying to display all possible information from a limited amount of data.*

The remarkable result is that all three spectrograms show similar features. The features extending below 0.02 Hz down to zero frequency from 1614 to 1616 are remnants of the long-period variations that have not been completely filtered out. These features are strongest in the H and phase plots, though also present in the D plot. As mentioned above, there may or may not necessarily be a connection between magnetic field and path length behavior at these lowest frequencies, though there does seem to be some similarity here.

More interesting are the Pc 3 signals above 0.025 Hz (period $\leq$ 40 s). Note the similarity between those features in the first half of the phase plot and corresponding features in D and, to a lesser extent, H. In particular, note that the phase features seem to precede those in D by 20 to 30 seconds. (This is most easily seen by tilting the page and viewing it from the bottom. Successive rasters in the plot are spaced by $t_{step} = 8$ s.) There is a similar correlation between phase and H, though it is not as strong.

---

* Note that with the phase spectrogram we are in the mathematically exquisite position of calculating the Fourier transform of the phase of the Fourier transform of a signal. If only there were a coherent signal here so we could examine *its* relative phase....

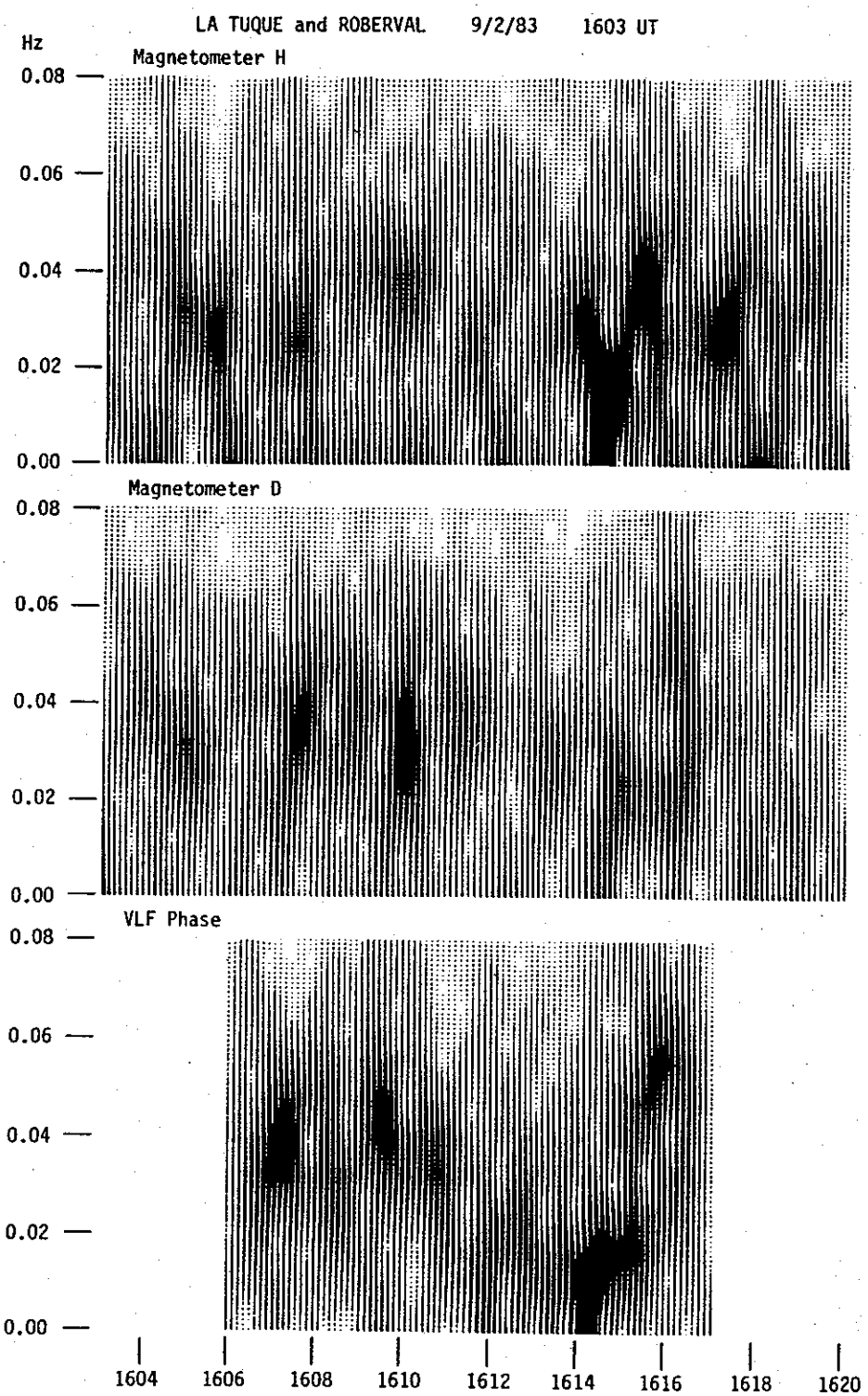

Figure 3.10. Spectrograms of the filtered magnetometer and VLF phase data in Fig. 3.9 ($N = 512$, $f_I = 0.001$ Hz, BW = 0.0145 Hz, $t_{step} = 8$ s). Similar features can be seen in all three plots. In particular, features from 0.025–0.05 Hz at 1606–1612 UT in the phase plot seem to precede similar features in D by 20 to 30 seconds.

Figure 3.11 shows plots of the cross-correlation* of the various filtered signals *vs.* lag time. These were made in order to quantify the delay from VLF phase to magnetic field seen in Fig. 3.10. When the cross-correlations were made, care was taken to use the full length of the sequence of phase samples ($N = 2471$) and to lag over the longer sequences of H and D. To facilitate comparison, the H *vs.* D plot was made using 2471 samples of D covering the same time interval as ϕ_c, even though more magnetometer data were available.

Figure 3.11 shows rather little correlation between H and D. I found this a bit surprising. However, hodograms of H *vs.* D (of both unfiltered and filtered data) show no preferential polarization; indeed, they look quite random. There is some correlation (0.43) between H and ϕ_c, with phase preceding magnetic field by 18 seconds. There is a stronger correlation (-0.55) between D and ϕ_c, again with phase leading by 18 s. (The negative sign in -0.55 is not a lack of correlation, of course. It means that an *increase* in ϕ_c is correlated with a *decrease* in D, or ϕ_c is correlated with $-D$.) There is also a correlation almost as large in the opposite sense (0.52) between $+D$ and ϕ_c with phase leading by 30 s. I want to emphasize here that the best evidence for the existence of a correlation between magnetic field and VLF phase is in the spectrograms of Fig. 3.10. The correlation plots merely quantify this.

Discussion. The spectrograms of the high-pass filtered data in Fig. 3.10 show similar spectral features in both the magnetic micropulsations and the VLF phase data. The cross-correlation plots in Fig. 3.11 show that phase changes precede magnetic field changes by 18–30 s. The VLF phase is most sensitive to conditions (electron density and magnetic field strength) at the top of the whistler-mode path where the wave spends most of its time. The magnetometer data, of course, measures conditions at ground level at one end of the path. At the time of the events studied here the magnetospheric path was on the dayside of the earth, approaching local noon.

A simple interpretation is that a disturbance of some kind (possibly a solar wind disturbance travelling inwards from the magnetopause) reaches the equatorial region of the Siple-Roberval whistler mode path; there it affects the propagation of the LICO1 signal from Siple, and then

* Given a sequence of N samples $\{x_n\}$ of one signal, sampled at times nT, and a (somewhat longer) sequence of samples $\{y_n\}$ of a second signal, the correlation coefficient of y *vs.* x at some lag $\tau = kT$ is

$$\rho(\tau) = \frac{\sum_{n=0}^{N-1}(x_n - \bar{x})(y_{n+k} - \bar{y})}{\left[\sum_{n=0}^{N-1}(x_n - \bar{x})^2 \sum_{n=0}^{N-1}(y_{n+k} - \bar{y})^2\right]^{1/2}}, \tag{3.15}$$

where $\bar{x}$ is the mean of the N samples of $x_0, \ldots, x_{N-1}$ and $\bar{y}$, which depends on the lag k, is the mean of $y_k, \ldots, y_{k+N-1}$. The means do not have to be calculated separately. Since $\sum(x_n - \bar{x})^2 = (\sum x_n^2) - N\bar{x}^2$, the correlation coefficient is easier to calculate as

$$\rho(\tau) = \frac{\sum_{n=0}^{N-1} x_n y_{n+k} - \frac{1}{N}\sum_{n=0}^{N-1} x_n \sum_{n=0}^{N-1} y_{n+k}}{\left[\sum_{n=0}^{N-1} x_n^2 - \frac{1}{N}\left(\sum_{n=0}^{N-1} x_n\right)^2\right]^{1/2}\left[\sum_{n=0}^{N-1} y_{n+k}^2 - \frac{1}{N}\left(\sum_{n=0}^{N-1} y_{n+k}\right)^2\right]^{1/2}} \tag{3.16}$$

since this can be implemented in a program as a single loop.

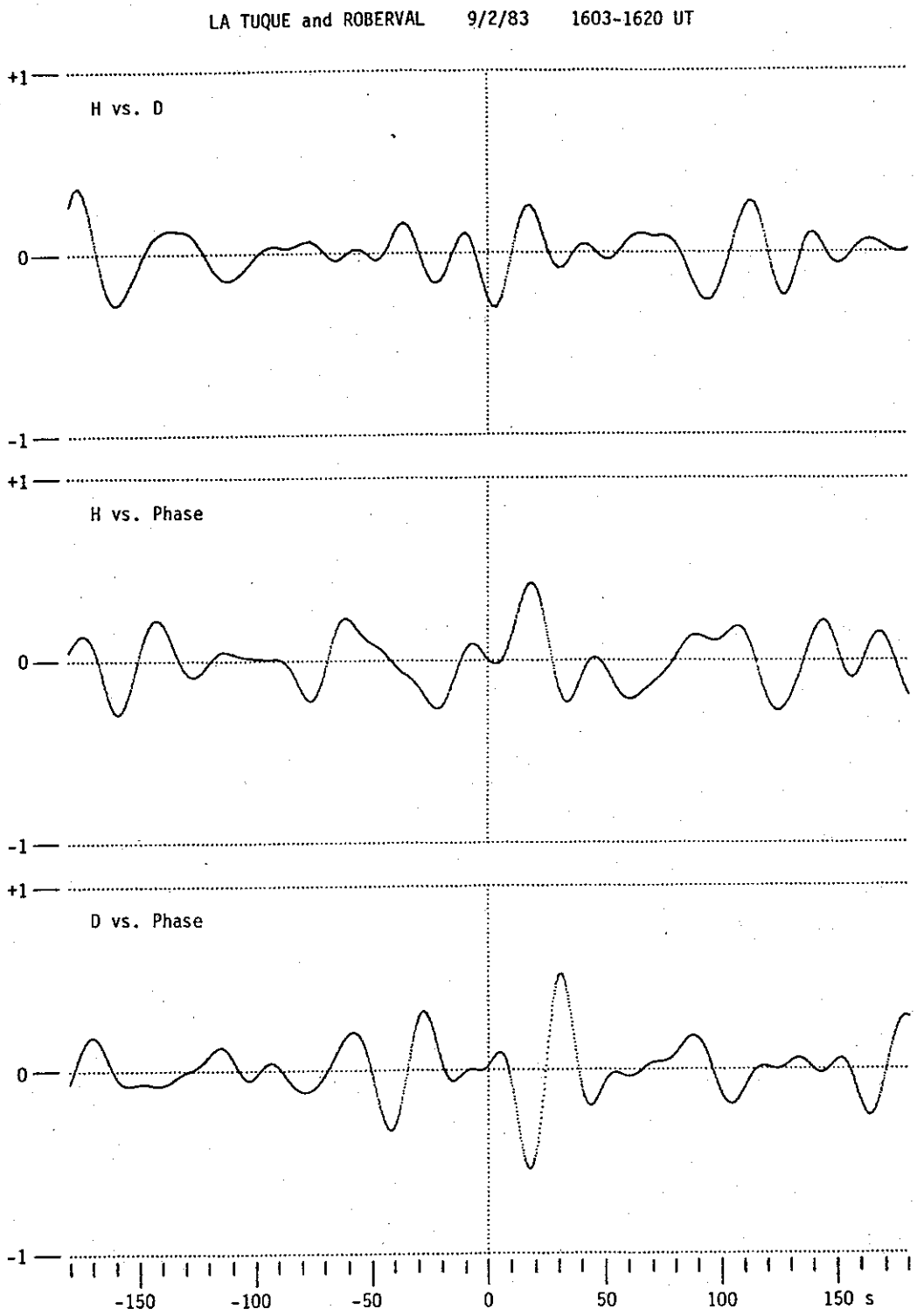

Figure 3.11. Cross-correlation plots of the filtered magnetometer and VLF phase data in Fig. 3.9. In each plot, the first variable leads the second for negative offset times, and lags it for positive times. There is little correlation between H and D. Both H and D show some correlation with VLF phase. The correlation between D and phase is bipolar, -0.55 with phase leading magnetic field by 18 seconds and $+0.52$ with phase leading by 30 s.

or about 12 s. Given a one-hop travel time for the micropulsation of $2 \times 25 = 50$ s, the average disturbance occupies only about one-quarter of the length of the field line.

Other Studies of VLF Phase and Micropulsations. *Rietveld et al.* [1978] report correlations between whistler-mode signals and micropulsations similar to those I have just described. They used signals at 6.6 kHz from a transportable VLF transmitter in Alaska as received in Dunedin, New Zealand ($L = 2.7$). Signals were transmitted in a variety of single-tone formats [*Dowden et al.*, 1978]. The Doppler frequency shift Δf of the signals was measured with the "phasogram" technique during two different events (intervals of 14 and 10 minutes duration), and correlated with Pc 4 micropulsations (with 94 and 60 s periods, respectively). The whistler-mode signals in these two events travelled in ducts at $L = 3.77$ and 2.85, respectively. Micropulsations were measured in Dunedin with a horizontal loop as dZ/dt.

There are two main differences between the approach of *Rietveld et al.* [1978] and the one I have followed here. First, signals they received showed non-linear amplification (*i.e.*, growth with phase advance) just before the first interval they studied, though not during the interval. Two-tone transmission was not used—its growth-suppression effect had not yet been discovered. Second, the transmission formats, consisting in some cases of pulses, did not allow them to use the relative phase of the signal as such; instead the instantaneous frequency of the signal was estimated by measuring the slope of the phase *vs.* time plot. Note that the two quantities they studied, Δf and dZ/dt, are essentially the derivatives of the two we used above, ϕ_c and either H or D. Their frequency measurements Δf tend to be noisier than our phase measurements ϕ_c, and their measurements do not give complete coverage in time during their two events. On the other hand, using instantaneous frequency instead of phase and dZ/dt instead of H or D, they were not troubled by the need to filter out long-term signal variations. Despite these differences, both our approaches are comparable.

Reitveld, et al. [1978] plot cross-correlations of Doppler shift Δf and dZ/dt *vs.* lag time for their two events (their Fig. 3). For the first event, they measure a time delay of 24 s with Δf lagging dZ/dt (correlation $\rho = 0.4$), quite close to one-quarter of the 94 s period of the micropulsation. For the second event, they find a time delay, again with Δf lagging dZ/dt, of 10 s ($\rho = 0.95$), just a bit small to be a quarter-period of the 60 s micropulsation. They claim that these measurements show Δf and dZ/dt to be in phase quadrature. They interpret their data as being consistent with a lowest-order resonant oscillation of a field line—a standing wave with an antinode (and maximum radial motion) near the equator. Resonant field-line oscillations are the standard model for Pc 2–Pc 5 micropulsations [*Jacobs*, 1970, Ch. 5].

There is one other study of the relationship between micropulsations and VLF phase path length of which I am aware. *Andrews* [1977] studied the correlation between micropulsations in the Pc 4–Pc 5 range (having periods from 2.5 to 6 min) and the Doppler shift Δf of whistler-mode signals from NLK (Jim Creek, WA) at 18.6 kHz as received in Wellington, NZ. The whistler-mode ducts were at relatively low latitudes ($L = 2.4$–2.6) compared to the Siple-Roberval case above. Whistler-mode signals were seen only at night. Doppler shift was measured with a spectrographic receiver containing a bank of filters spaced every 0.05 Hz centered at 18.6 kHz. The receiver was made for long-term studies of duct behavior, and had barely enough time resolution to resolve pulsations with periods as short as 2.5 min.

Andrews' [1977] development the theory of the effect of resonant micropulsations on the whistler-mode phase path (mentioned earlier) seems complete. His experimental data support the idea that phase path changes are due to radial motion of the duct in the equatorial region. Unfortunately, the time resolution of his measurements was not sufficient to determine if there was any delay between

changes in the VLF phase path and changes in the magnetic field.

Conclusions. All the studies mentioned above support the idea that magnetic micropulsations affect VLF propagation in whistler-mode ducts. The radial motion induced by the micropulsation changes the length of the phase path of the VLF signal. Inward displacement of a duct is associated with an outward meridional micropulsation field b_x north of the displacement and an inward b_x south of it, and with a shortening of the phase path or an increase in relative phase. Because of rotation of the plane of polarization caused by the ionosphere, b_x above the ionosphere becomes $-D$ at the ground in the northern hemisphere; thus we see a correlation between VLF relative phase ϕ_c and $-D$. This correlation has a bipolar shape for short, non-echoing disturbances. The Siple-Roberval measurements also show a time delay between phase changes occuring near the equator and magnetic field changes at the base of the duct consistent with the propagation of an Alfvén wave from the equator down the field line to the ground. The length of the average duct displacement event in the Siple-Roberval case seems to be only a small portion of the total length of the field line, and is thus different from the resonant field-line micropulsations studied previously.

The advantages of the continuous two-tone LICO1 signal as used in the Siple-Roberval experiment are several. First, a two-tone signal tends to suppress the temporal growth and associated phase shift that can otherwise mask path-related phase changes, and allows phase errors due to multipath fading to be corrected. Second, a continuous transmission allows the phase of the received signal to be measured and used directly. Phase may be a less-noisy measure of path conditions than its time derivative, the Doppler shift Δf. Third, compared to Navy transmitters, the lower frequency of the Siple transmitter allows propagation on higher-latitude paths where micropulsation activity may be more interesting, and on daytime as well as nighttime paths. Finally, if accurate data timing is available, delays between changes in the VLF phase path and changes in the magnetic field can be measured and it may be possible to study the shape of transient field-line disturbances. The percentage of time when a two-tone transmission from Siple can be received in usable form in the northern hemisphere is not known and should be the subject of further research.

4. SIGNALS WITH GROWTH

4.1 Phase Behavior During Growth

Characteristics of Growth. In Section 3.3 we saw examples of whistler-mode signals that propagated from source to receiver essentially unaltered except for Doppler frequency shifts caused by duct motion and plasma flux into the magnetosphere. With the exception of small distortions in phase delay, the magnetospheric medium was linear—what went in at one end of a duct was what came out at the other. This is not always the case. Whistler-mode signals are often greatly altered in their journey through the magnetosphere. Pulses from a VLF transmitter at one end of a duct may be found at the other end to have amplitudes, frequencies, and spectral structures which are not only different from what went in but which change rapidly. A common behavior is *temporal growth,* where the amplitude of a constant input signal is seen at the output to increase exponentially with time. Temporal growth is just one of a group of associated characteristics that are believed to be caused by cyclotron interactions that take place near the top of the path between the whistler-mode wave and energetic electrons trapped by the earth's magnetic field. Other growth-related features (summarized in Section 4.7) include an advance in relative phase, a slowing of growth as *saturation* is reached, magnitude and phase ripples and *sidebands* at saturation, a *band-limited impulse* or BLI at the end of a pulse, and a *triggered emission* which continues after the end of the transmitted signal.

Some of these features, such as emission triggering, are easy to pick out in spectrograms and have been observed for a long time. Risers (rising-frequency emissions) triggered by man-made signals were apparently first found in recordings of signals from NPG (NLK) made in Wellington, New Zealand, during the IGY [*Helliwell et al.,* 1964]. Others, the phase features in particular, have only been seen relatively recently [*Dowden et al.,* 1978; *Paschal and Helliwell,* 1984]. In this and the following sections we will examine some of these features, especially those about which phase analysis has something to say.

A Suite of Growing Pulses. Figures 4.1 and 4.2 show *f*-*t* spectrograms and magnitude-phase plots of four one-half second pulses sent from the Siple Station VLF transmitter as received at Roberval, Quebec. When transmitted, all pulses were at the same constant amplitude and a frequency of exactly 4500 Hz. As received, the pulses show an increase in amplitude with time, an advance in relative phase at an increasing rate, and other growth-related features. The pulses shown were part of the transmission format "ULF75," which consisted of alternating 0.5 s pulses at 4500 Hz and 0.5 s segments of idler signal. The idler signal, 50 ms pulses alternating between 4100 and 4000 Hz, keeps a constant load on the station generators without triggering emission activity in the magnetosphere. ULF75 was an unsuccessful attempt to create 1 Hz micropulsations.

This particular record was selected because the behavior of individual pulses is quite typical, and because there seems to be little if any multipath propagation. Multipath propagation causes confusion for two reasons. First, signals with different group delays propagating on different paths may show widely differing levels of activity, yet be overlapped at the receiver. Second, differential phase changes between different paths may create fading with amplitude and phase anomalies unrelated to the growth process. The signals seen here propagated predominately on only one path, and we can interpret them more easily because of this. The level of growth activity at this time

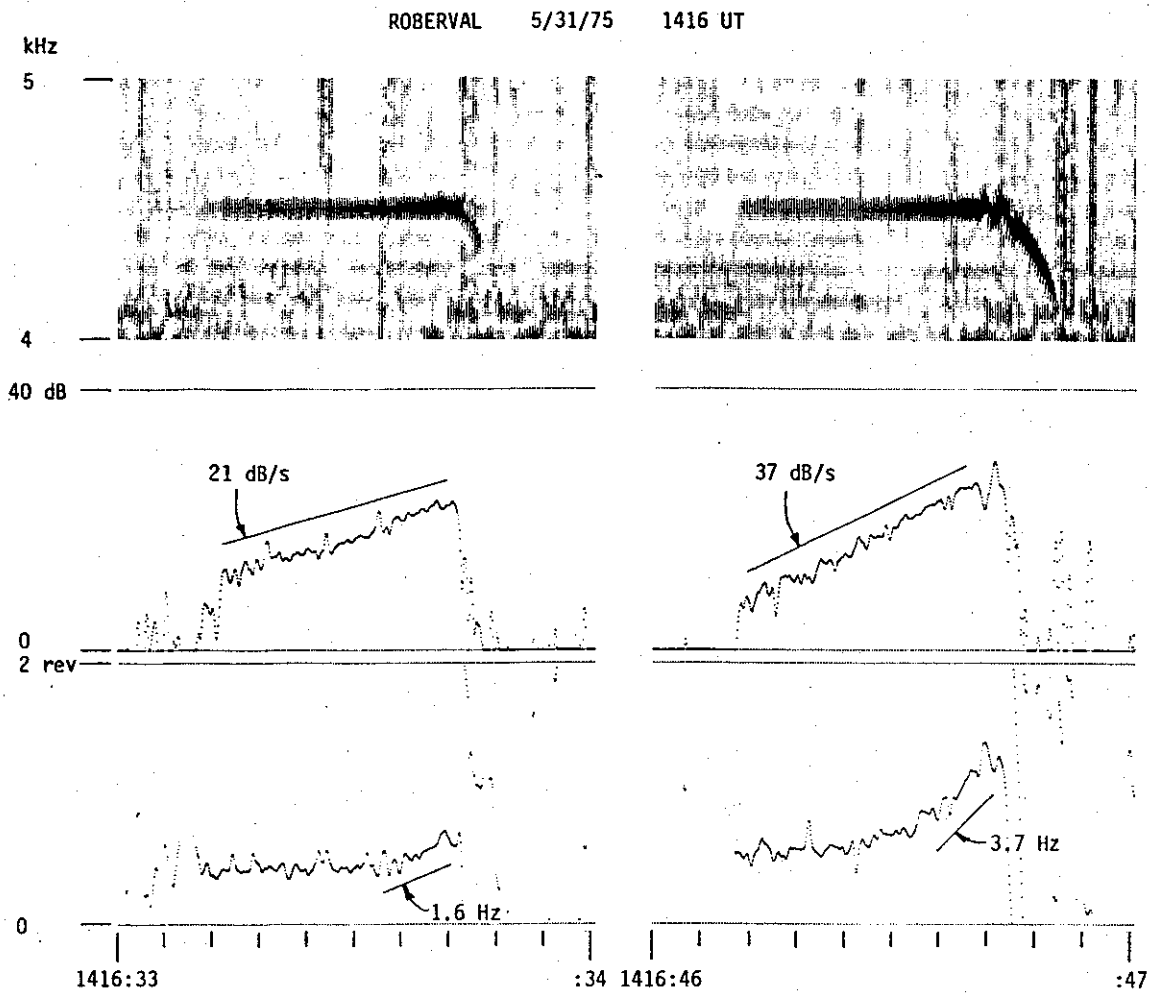

Figure 4.1. Spectrograms and magnitude-phase plots (BW = 80 Hz) of half-second pulses at 4500 Hz from the Siple Station VLF transmitter showing growth and phase advance. The first pulse shows growth at a rate of 21 dB/s with a phase increase of 0.25 rev during the last 100 ms. The pulse terminates in a brief falling emission. The second pulse shows faster growth, a phase advance of 0.65 rev during the last 300 ms, and a stronger emission.

is low to moderate, and this makes the pulses easier to interpret as well since their behavior is a little simpler. (In fact, low to moderate activity and predominately single-path propagation may be correlated. There may be a threshold in amplitude below which temporal growth does not occur. At times of low growth activity only one path may have conditions such that signals in it are above the threshold, and we observe single-path propagation. At times of enhanced activity signals on many paths may grow to saturation, and we see multipath propagation [*Helliwell et al.*, 1980].) Often we find that the level of growth activity changes over the course of tens of seconds, as it does in this record. While there is a general continuity in behavior, the exact amounts of growth and phase advance (and other related features) change from pulse to pulse. This is not due to any change in the transmitted signals but shows the inherent variability of the growth process. It is presumably caused by changes in the distribution function of energetic electrons that may be available at any given time to interact with the whistler-mode wave.

The group delay of the pulses in Fig.s 4.1 and 4.2 is $t_g = 2.18$ s. From whistlers that occurred

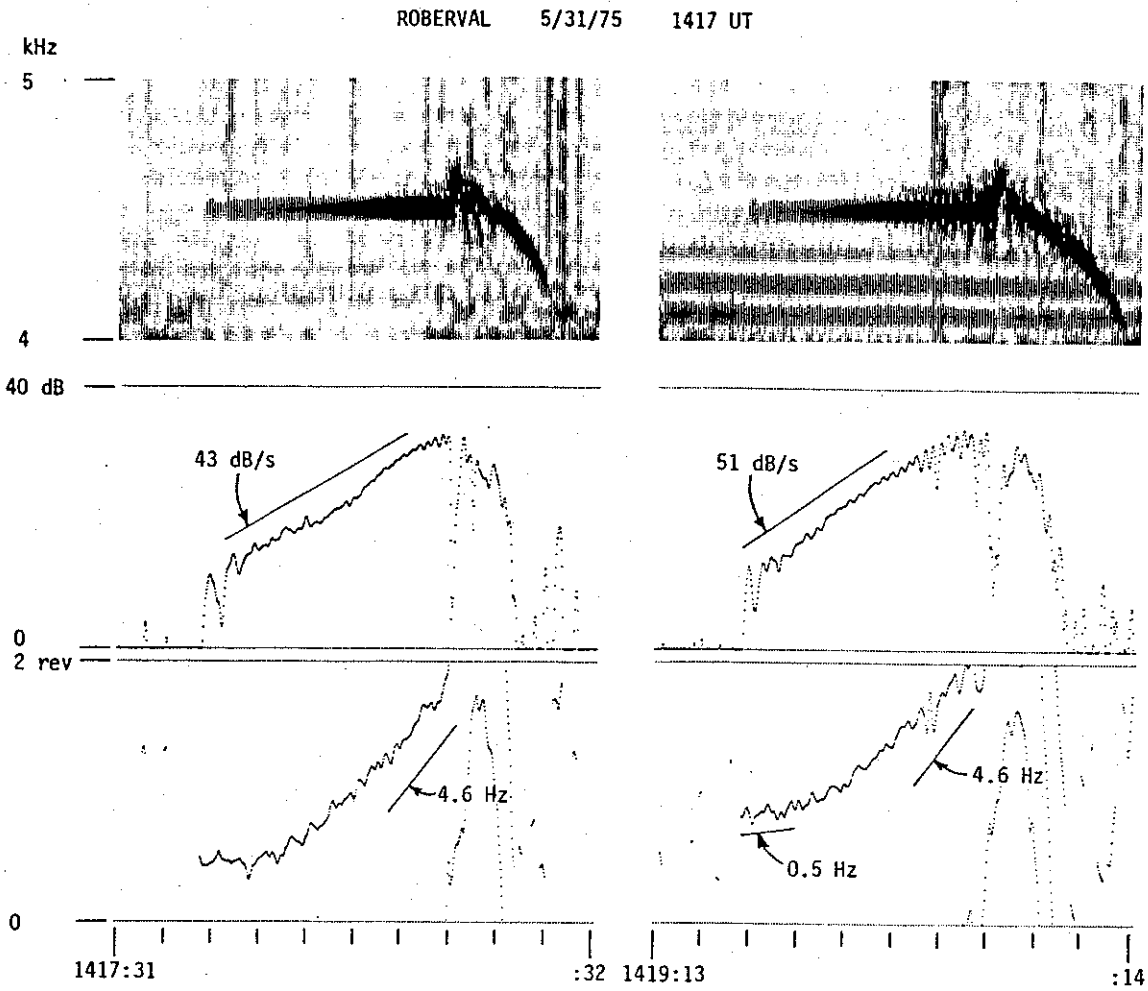

Figure 4.2. Two more half-second pulses like those shown in Fig. 4.1. The growth rate continues to increase as magnetospheric conditions change. Pulses now begin to show amplitude saturation as growth slows toward the end of the pulse. Amplitude ripples just before the end signal the formation of sidebands. The second pulse seems to show a small positive frequency offset (roughly 0.5 Hz) due to phase advance even during the first 100 ms. Phase advance occurs at an increasing rate and totals 1.4 revs at the end of each pulse. Both pulses show magnitude nulls and rapid phase advance as a termination emission begins above the transmitted frequency.

at this time we can identify the latitude and tube content of the path of the pulses. A group delay of 2.18 s at 4500 Hz matches a whistler component in the record with a nose frequency f_n of 3.9 kHz and a nose delay t_n of 2.1 s. From *Park* [1972] we find the path to be located at $L = 4.36$ with tube content $N_T = 3.4 \times 10^{13}$ electrons/cm^2 (equatorial electron density $N_{eq} = 260$ electrons/cm^3), typical of magnetically quiet conditions. The 4500 Hz pulses are at a frequency of $0.43 f_{Heq}$ on this path. Whistlers are seen on higher-latitude paths up to about $L = 5.0$. There is a very weak path near $L = 4.2$, but the $L = 4.36$ path is the one at the lowest latitude with strong whistler components.

Growth activity gradually increases during the 5 minutes of the ULF75 transmission, and the various growth-associated features begin to appear at different times. The 4500 Hz pulses are visible from the start of the record at 1415 but are weak and show no growth until 1415:40. By 1416:10

pulses grow about 7 or 8 dB (15 dB/s), but show no phase advance. By 1416:20 some pulses have short falling emissions at termination. At 1416:33 (the first pulse in Fig. 4.1) the growth rate is 21 dB/s, accompanied by a phase advance of 0.25 rev. Note that the magnitude scale in Fig. 4.1 is logarithmic (in decibels); a straight line on this plot means that the magnitude changes exponentially with time. Exponential growth is a very common feature, at least at the beginning of a pulse.

However, while the growth rate (in dB/s) is more or less constant throughout the first 0.5 s pulse in Fig. 4.1, the phase advance occurs only during the last quarter of it. An advance in relative phase with time means that the received signal is at a higher frequency than the transmitted signal. This, too, is a typical feature of growing pulses. In the first pulse in Fig. 4.1 the increase in frequency or *frequency offset* Δf reaches +1.6 Hz during the last 100 ms of the pulse. Note the advantage of phase analysis here. The phase plot shows a small but definite change in instantaneous signal frequency at the end of the pulse. This change would have to be much larger, or continue for a much longer time, to be noticeable in the *f-t* spectrogram.

At 1416:46 (the second pulse in Fig. 4.1) growth is 37 dB/s, with a phase advance of 0.65 rev mostly during the last half of the pulse. At the end of the pulse there is a small transient. This is an incipient band-limited-impulse or BLI [*Helliwell,* 1979b]. It is seen in the magnitude plot as a momentary amplitude fluctuation (a dip of 4 dB followed by a brief increase) at 1416:46.67 lasting 50 ms, and in the phase plot as an increase of 0.22 rev before a rapid drop. After the end of the pulse a fairly strong falling emission or "faller" continues for about 150 ms. This is a self-excited oscillation which was initiated by the transmitted signal but which now continues for a brief time after the input signal ends. Section 4.2 discusses termination emissions in more detail.

At 1417:31 (the first pulse in Fig. 4.2) the growth rate is 43 dB/s with a total phase advance of 1.4 rev. The phase advance now begins about 100 ms into the pulse and is roughly parabolic; that is, the relative phase ϕ is proportional to t^2. This means that the frequency offset $\Delta f = d\phi/dt$ increases linearly with time. A parabolic phase advance is a common feature of growing pulses, but is not as regular a feature as, say, the exponential growth in amplitude. Some pulses show only an approximately linear phase increase with time. By the end of this pulse its frequency is 4.6 Hz above that of the transmitted signal.

The BLI is well developed in the first pulse in Fig. 4.2, starting with a brief 30-dB dip in amplitude. After the BLI a strong faller begins. Note that the faller starts about 100 Hz above the transmitted frequency. In fact, the phase plot shows a rather rapid change in instantaneous frequency between the 4.6 Hz slope at the end of the pulse and the much faster increase at the emission frequency. The faller wraps up about 2 revs in phase before falling back through 4500 Hz. The faller now continues to drop until, at 1417:31.92, it is seen in the *f-t* spectrogram to intersect the third 50 ms–4100 Hz pulse of an idler signal with a brief increase in amplitude. This last effect is called *entrainment* and is discussed in Section 4.5.2.

At 1419:13 (the second pulse in Fig. 4.2) the growth rate has reached 51 dB/s during the first part of the pulse, but decreases toward the end of the pulse as saturation occurs. Total growth is about 20 dB, only slightly more than in the previous pulse. Saturation is another very common feature of growing pulses, and usually happens after 20–35 dB of growth. As saturation occurs, the magnitude develops 60 Hz ripples roughly 3 dB p-p. Associated with these are sidebands 60 Hz above and below the main signal (barely visible in the *f-t* spectrogram). These sidebands are discussed in Section 4.4.2. The falling emission at the end of the pulse is now even stronger than before. Unfortunately, any entrainment of the faller by idler pulses cannot be seen at this time because of increased local interference. (The strong signals at 4100, 4220, and 4340 Hz in the right-hand spectrogram of Fig. 4.2 are due to power line noise probably from rotating machinery.

This variable-frequency interference is especially annoying in this record, but is fortunately absent from 1415:45–1418:01 and after 1419:39. It is definitely local in origin and is not a magnetospheric signal.)

The total phase advance at the end of the second pulse in Fig. 4.2 remains about 1.4 rev, similar to the previous case, but it now starts at the beginning of the pulse, and seems to have an initial frequency offset of +0.5 Hz. The determination of the time of onset of phase advance in a growing pulse is important for the theory of growth, but it is made difficult because the amplitude at the beginning of a growing pulse is generally very low, and the plot of its phase is correspondingly noisy. There are many pulses in this record, especially after 1418:30, which seem to have a positive frequency offset even at the beginning of the pulse—i.e., within the first 20 or 30 ms. However, this is a somewhat subjective judgment.

Note that the offset frequency at the end of the pulse is still only about 4.6 Hz (or a little less), the same as in the previous pulse. Even though the initial growth rate is higher, and the phase advance starts earlier and at a faster rate, the final offset frequency is approximately the same. This is a relatively common effect. Many growing pulses show a phase which increases parabolically at the beginning but only linearly at the end (or until an emission is triggered). That is, the offset frequency increases at the start, but then saturates at some maximum value, much as we usually see saturation in amplitude. However, some pulses, as in the next example, show a steady increase in frequency offset without apparent limit until triggering takes place. The conditions under which a limit in frequency offset occurs are presently unknown, and remain questions for future study.

A Classic Pulse. Figure 4.3 shows *f-t* spectrograms and magnitude-phase plots of two one-second pulses at 2000 Hz. The pulse on the left exhibits most of the features associated with growth—everything except sidebands. It also illustrates some of the ambiguities of real data. We will see how non-growing or weakly-growing pulses can be used to separate real growth features from artifacts due to multipath.

The pulse in the left half of Fig. 4.3 shows typical growth behavior. Growth in magnitude is roughly exponential at 36 dB/s, though it seems faster during the first 100 ms or so. After about 600 ms the pulse develops ripples in magnitude, and a BLI occurs at about 700 ms. The BLI is seen in the spectrogram as a short, impulsive blip of energy extending about 100 Hz above the pulse, and as a short null in the magnitude and a brief skip in the phase plots. After the BLI a rising emission is triggered which slowly drifts away from the input signal. The pulse shows an initial positive frequency offset of +1.1 Hz, and the phase advances at an increasing rate with time. *Carlson* [1987, Fig. 1.4] has analyzed the phase of this pulse and finds that, prior to the emission, the offset frequency Δf increases nearly linearly with time at a rate of 9 Hz/s (*i.e.*, phase is proportional to t^2). The BLI and the triggered emission occur when the relative phase has wrapped up about 3 revolutions, even though we have not yet reached the end of the one-second input pulse. This example is different from those in Fig.s 4.1 and 4.2 where an emission occurs only at the end of the input pulse. The characteristics of such pre-termination emissions are discussed in Section 4.3. After the BLI, as the emission starts to separate from the input signal, the phase advance increases at an even faster rate, and at 750 ms the pulse is seen to be 32 Hz above the input frequency and rising.

However, the left-hand pulse in Fig. 4.3 also presents us with an ambiguity. After the emission separates we cannot see any signal at the input frequency. So where does the pulse end? For that matter, where exactly does it begin? The bar under the plot shows the assumed position of the pulse, but how was this determined? There is a weak blob of signal in the spectrogram just before

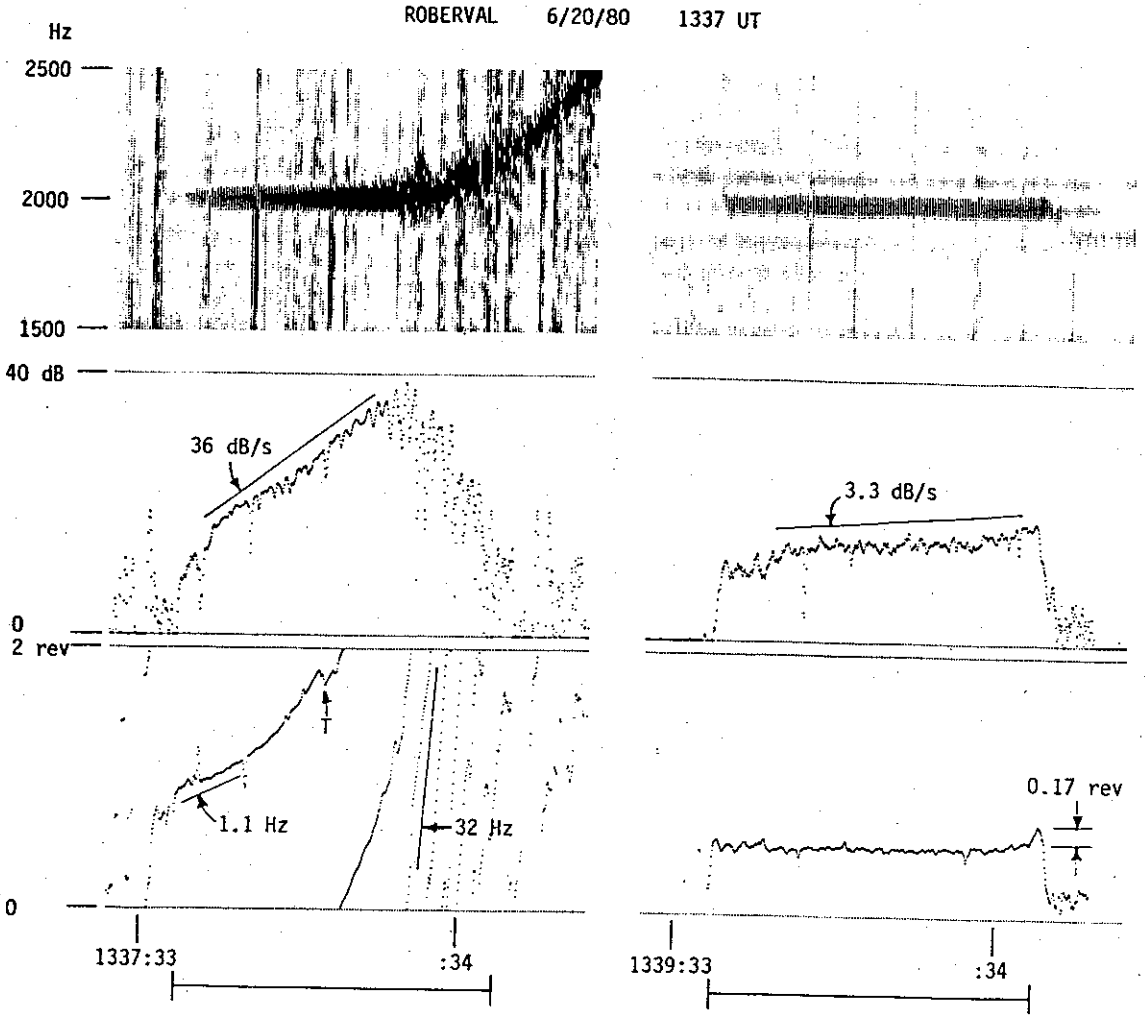

Figure 4.3. Spectrograms and magnitude-phase plots (BW = 80 Hz) of two one-second pulses at 2000 Hz. On the left is a pulse with temporal growth, increasing phase advance, initial frequency offset, and a pre-termination emission. "T" marks a spurious phase jump due to a pilot tone tracking error. On the right is a pulse two minutes later with little growth, used to identify the group delay time (and hence the beginning) of the first pulse. Bars below the plots show the duration of the transmitted signals. Pulse at left after *Paschal and Helliwell* [1984].

the assumed beginning of the pulse, and a corresponding plateau in the phase plot. Is this the real beginning? These questions are of some consequence because measuring a positive frequency offset at the beginning of a pulse assumes that we have accurately identified the beginning.

To answer these questions we turn to the right-hand pulse in Fig. 4.3. This is a one-second pulse at the same frequency, but it was transmitted two minutes later. Because magnetospheric conditions had changed, this pulse shows almost no growth. It shows a very weak termination faller with a phase wrap-up of 0.17 rev, but is otherwise almost undistorted. We see that there is a second path of propagation following the first one by 150 ms and about 10 dB lower in amplitude than the initial value. However, there does not appear to be any path with a signal preceeding the main pulse, and we fix the group delay of the leading edge of the pulse at $t_g = 3.11$ s. (We know from the transmission schedule that these pulses were sent exactly 30 s after the minute.) The bar under

the right-hand plot is exactly one second long, and starts at 1339:33.11. Now we can return to the previous pulse and identify its beginning. While growth activity can change radically from minute to minute, the group delay of a given path changes very little in such a short time, and we can be confident in assigning a delay of 3.11 s to the leading edge of the pulse at 1337:30 as well. The bar below the left-hand plot is again exactly one second long and starts with this delay at 1337:33.11. Also note that the magnitude at the beginning of the left-hand pulse is the same as that of the non-growing pulse (about +10 dB on the plot). Growing and non-growing signals typically start at the same level, so this confirms our identification of the group delay of the pulses.

More Complicated Phase Behavior. The pulses we have seen so far show a relatively constant growth in amplitude and a smooth parabolic (or linear in the limit) phase advance. Lest the reader think that all growing pulses are so well-behaved, I present Figure 4.4. This is an unusual case of phase behavior that proves (*i.e.*, tests) the rule.

The one-second pulse at 4050 Hz in the left half of Fig. 4.4 advances rapidly in phase by 0.9 rev in 240 ms, slows down and retards in phase by 0.2 rev over the next 120 ms, and then resumes the advance at a lower rate for the remainder of the pulse. This pulse has a very large initial frequency offset of 4.5 Hz. During later growth and saturation the frequency offset is only about 0.6 Hz. Saturation during the last third of the pulse is marked by strong magnitude and phase pulsations as sidebands are formed. Phase advance continues during saturation, as we have seen previously. The end of the pulse shows a brief BLI and a weak faller. The right-hand plots in Fig. 4.4 are gray-scale phase plots of four additional one-second pulses recorded at this time. Two of the pulses are again at 4050 Hz, and look very similar to the first one. The other two are at 3570 and 3810 Hz. The amount of phase lag in the 3810 Hz pulse is a bit less (0.15 rev) than at 4050 Hz, and the 3570 Hz pulse shows only a slowing of phase advance (0.1 rev over 120 ms) at that point, but their overall behavior is still similar.

The magnitude of the first pulse in Fig. 4.4 parallels its phase advance. The pulse has a large initial growth rate, 130 dB/s, which decreases as the initial phase advance slows down. During the 120 ms when the phase retards, the magnitude of the pulse decreases by 6 dB. When the phase advance resumes, the magnitude increases again, but at a slower rate, 35 dB/s, until saturation is reached and sidebands develop. The magnitude behavior of the additional pulses in Fig. 4.4 is more variable than their phase behavior. Since magnitude plots are not shown, I will describe them briefly. The next two pulses, at 3570 and 4050 Hz, also show a decrease in magnitude during the interval of phase lag, though the decrease is smaller than with the first pulse, only about 3 dB. The last two, at 3810 and 4050 Hz, do not show a magnitude decrease at this time. In fact, the 3810 Hz pulse shows fairly steady growth throughout, up to saturation. We will see more of these pulses when we discuss single-signal sidebands in Sec. 4.4.2. The pulse at 1351:30 was originally shown in *Paschal and Helliwell* [1984], and that at 1351:31 in *Park* [1981].

The pulses in Fig. 4.4 are unusual. Irregular magnitude behavior often occurs (see the first pulse in Fig. 4.10, for example), but such an irregular phase advance is not often seen. In fact, the phase decrease shown by these pulses is quite rare. Typical pulse behavior is more like that shown by the pulses in Fig.s 4.1–4.3. A single pulse with unusual features can be dismissed as a freak, a fortuitous coalescence of circumstances without any deeper meaning. In this case, however, the behavior shown is repeated by every one of the eighteen constant-frequency pulses that were sent over the course of a one-minute DIAG1 (DIAGnostic, version 1) transmission. The problem is to explain these odd features.

One possible explanation might be that this is caused by multipath propagation. Two signals,

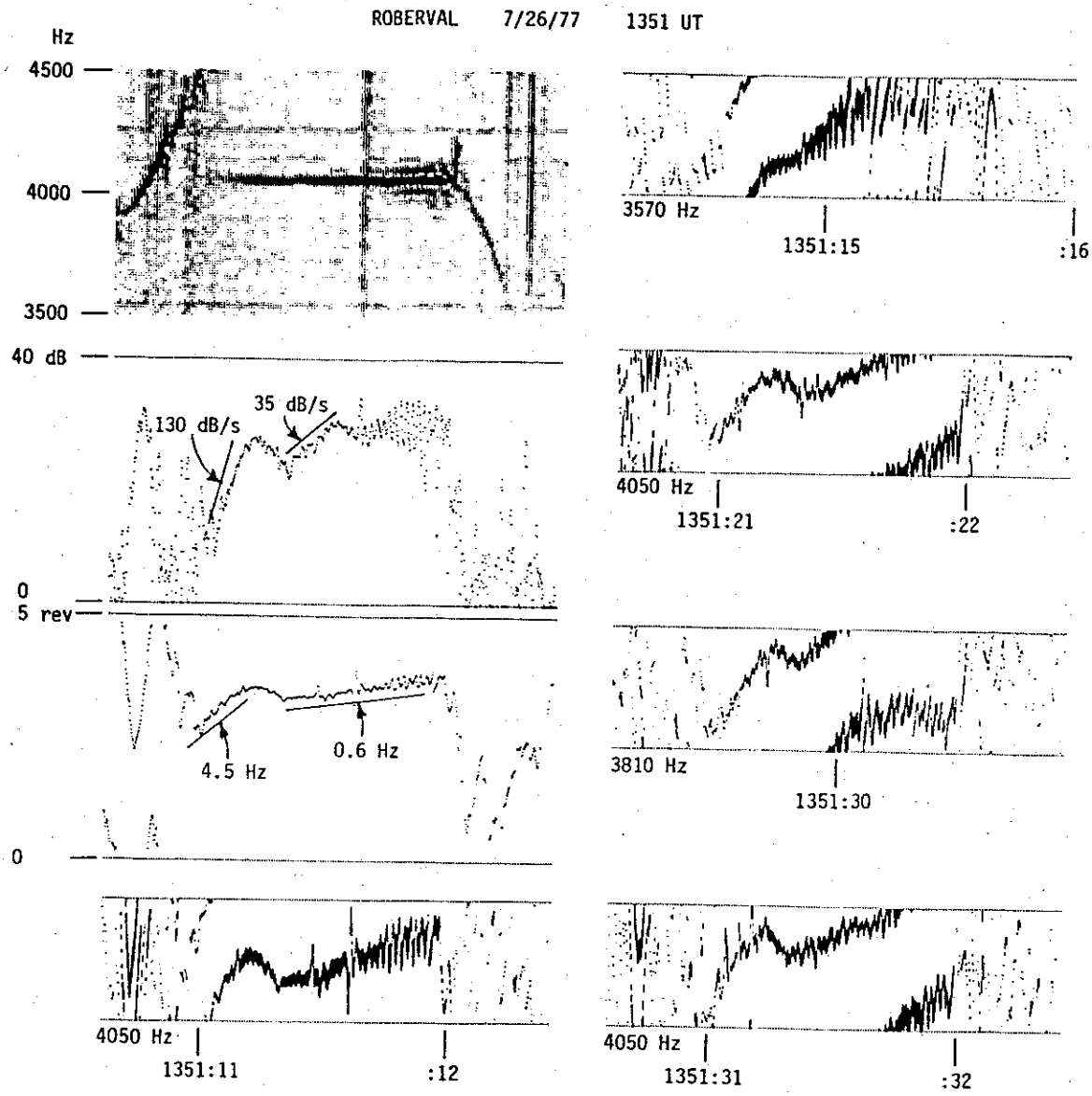

Figure 4.4. Left: spectrogram, magnitude-phase plot (BW = 80 Hz) and gray-scale phase plot (BW = 80 Hz, P_{span} = 1 rev) of a one-second pulse showing unusually complicated phase behavior during growth. Right: gray-scale plots of additional pulses showing the repeatability of the behavior at different frequencies and times.

growing at different rates, might be sufficient. Suppose that one path has more rapid growth, but stops early as it reaches saturation in about 240 ms. Or perhaps this path is better coupled through the ionosphere at the transmitter end so it starts with a higher input signal, and because of this reaches saturation first. The second path, perhaps with an exit point closer to the receiver so it will ultimately be heard stronger, or which couples better through the ionosphere at the receiving end, begins to dominate after about 360 ms, the start of the interval of slower phase advance. The phase decrease between the saturation of the first and the dominance of the second just depends on the difference in phase (modulo 1 rev) between these two paths. That is, at the beginning of the pulse we see only the first, rapidly growing signal, at its particular phase. Later we see primarily the second signal, at its phase, a fraction of a rev delayed with respect to the first. In between we see

a decrease in phase equal to half the difference between the two (that is, we see their mean phase when their magnitudes are equal).

There are several objections to this explanation. The frequency ramp signals in the diagnostic format do not show any evidence of multipath propagation, at least any additional paths with group delays differing by more than a few tens of milliseconds. Only single emissions are triggered at the ends of the various pulses, again evidence against multipath. And even if the momentary phase decrease were due to the addition of two different signals, it seems very unlikely that their relative phases should remain constant within a fraction of a revolution over the course of a minute. That is, any change in the relative phase delay between the two paths would change the amount of phase decrease (or increase) when the second signal begins to dominate, whereas this decrease actually remains fairly constant from pulse to pulse.

Of course, the two paths could be nearly the same length and have nearly the same total phase delay, so differential changes might be small. Nearly identical lengths are also necessary to explain why the phase difference is so similar for pulses at different frequencies. Since the difference is −0.4 rev at 4050 Hz (that is, twice the drop of 0.2 rev seen in the sum), and +0.2 rev at 3570 Hz, then (assuming values at intermediate frequencies are also in this range) the two paths must differ in phase length by about $\Delta t_p = (0.4 + 0.2)/(4050 - 3570) = 1.25$ ms. This corresponds, at $L = 4$, to a difference in nose delay of $\Delta t_n = 0.732$ ms, which gives a difference in path L-value of about $\Delta L = 0.002$, or a difference in latitude at the ends of the ducts of only 1 km. This is much less than the estimated size of a duct at the earth's surface, so the two ducts, while being at nearly the same L-value, must be separated in longitude. This is not implausible in itself, but it is difficult to explain why such similar paths should behave so differently in growth activity or trans-ionospheric coupling, especially in having opposite coupling behavior at each end.

One final objection to multipath interference is that the magnitude dips do not correspond to the observed phase differences. For the amplitude of the sum of two equal signals to be 6 dB less than that of the first signal alone, as in the first pulse in Fig. 4.4, they must differ in phase by 0.42 rev, approximately the amount seen. However, given the nearly identical phase behavior of the three pulses at 4050 Hz, we would expect similar magnitude behavior. There is a dip in magnitude of only 3 dB in the second pulse at 4050 Hz, and no corresponding dip at all in the last pulse.

Having ruled out multipath as a cause, we are left to conclude that the irregular growth and phase advance shown in Fig. 4.4 is an intrinsic feature of the growth process. Perhaps it is caused by slight motions of the interaction region about its nominal position near the equator (see following discussion), or by some more complicated feedback mechanism between different parts of an extended interaction region. These pulses are also unusual in having such a high initial frequency offset, and this may be a clue to their odd behavior. I think it will take an advanced model of wave-particle interactions to explain this case.

Theories of Growth. Growing signals, especially triggered emissions, have been observed for a long time, and theories to explain them were formulated as early as 1959 [*Gallet and Helliwell*, 1959]. For a review up to 1979 see *Matsumoto* [1979]. *Carlson* [1987, pp. 7–11] gives a brief review and lists some more recent studies. These theories have all attempted to explain the overall frequency characteristics of growing signals and emissions with varying degrees of success. Some have tried to explain other growth-associated features such as saturation, pulsations, and sidebands, with much less success. However, only those by *Nunn* [1974], *Dowden et al.* [1978], *Helliwell and Inan* [1982], and, most recently, *Carlson* [1987] have attempted to explain the phase characteristics of growing signals. This is undoubtedly because until recently there was very little experimental phase data for

comparison. (One of the main purposes of this report is to bring more phase data to light.)

The most promising theories of growth have been based on the doppler-shifted cyclotron or gyroresonance interaction, now generally accepted to be the correct model. In the cyclotron interaction the transverse field of a ducted whistler-mode wave travelling in one direction affects energetic electrons travelling in the other direction. Both the fields of the circularly-polarized wave and the electrons rotate about the earth's magnetic field line in the same direction, but the wave frequency is nearly always less than half the electron gyrofrequency. However, if the electrons are moving with respect to the wave, they see a local field whose speed of rotation depends both on the wave frequency at a given place (the frequency of the signal) and on the number of additional cycles of wave passed in a given interval due to the relative motion. If the electrons are travelling along a field line in the opposite direction to the wave they may see a transverse field which appears stationary to them, the condition of gyroresonance. Because of the inhomogeneity of the magnetosphere—both gyrofrequency and wave phase velocity vary with latitude along a given field line—different electron velocities are needed to resonate with a given wave at different positions. The inhomogeneity is smallest, and thus the interval when a given electron may be in resonance with a wave is longest, at the top of the path assuming constant-frequency waves; it is here in the equatorial region that the strongest wave-particle interactions are presumed to take place.

The longitudinal (or Landau) resonance has also been studied, though not as extensively [*e.g., Tkalcevic,* 1982]. In this interaction the longitudinal component of the wave field affects electrons whose velocity parallel to the earth's magnetic field matches the phase velocity of the wave. This interaction seems to be much weaker than the cyclotron interaction for ducted waves, possibly because their longitudinal field components tend to be much smaller than their transverse components. However, it is probably important for non-ducted waves, and may explain such things as whistler precursors (see Section 4.5.3). Čerenkov radiation has also been proposed to explain some VLF phenomena [*e.g., Kimura,* 1967], but seems unlikely to be very important for the present experiments.

However, agreeing on the proper wave-particle resonance does not mean we understand the process of growth any more than knowing the viscosity of air and the adiabatic lapse rate* enables one to predict the weather. Like the weather, actual wave-growth events are complex because of the sheer numbers of wavelets and particles involved, the nonlinearities in their interactions, and the heterogeneous conditions in the magnetosphere. Even worse, our knowledge of magnetospheric conditions at any given time is much less than our knowledge of that sparsely-sampled global network from which weather forcasts are made. Every author modelling growth has had to make various simplifying assumptions to render the problem tractable, and no study has explained all of the phenomena associated with growth in a completely satisfactory way. Here I will review the efforts of *Nunn* [1974], *Dowden et al.* [1978], *Helliwell and Inan* [1982], and *Carlson* [1987] mentioned above, to see how close current theories come to explaining real events. The first three studies are discussed in a bit more depth in *Paschal and Helliwell* [1984].

Nunn [1974] studies a transmitted pulse at $f = f_{Heq}/2$. The interaction takes place near the equator in an inhomogeneous medium. Nunn integrates the fields produced by three streams of resonant particles at different pitch angles and finds the amplitude and phase of the total field as the pulse passes through the interaction region. Strongly trapped particles play the dominant role in this narrowband interaction. *Nunn's* [1974] model successfully predicts the growth and general

* The adiabatic lapse rate is the change in temperature that occurs when a packet of air is moved vertically in the atmosphere without gaining or losing heat to its surroundings. The actual lapse rate is about 1.98 °C/1000 ft, and is of interest to pilots and mountain climbers as well as weathermen.

phase advance that is seen on a growing short pulse, but it also predicts an initial phase lag that is not apparent in real data. This model cannot account for amplitude saturation, or pulsations and sidebands.

Dowden et al. [1978] present observations, including relative phase plots made by the "phasogram" technique (see Sec. 1.3), of VLF transmissions at 6.6 kHz from Alaska as received in New Zealand. An interesting feature they observe is the "*N* event," a sudden decrease in phase of the output wave caused by beating between the amplified signal and the "embryo" emission, which tends to decrease the frequency of the output signal and keep the emission entrained by the input. I am not sure if their *N* event really is an independent growth-related feature, or whether it may not be just an example of the phase ripples that are often seen, say, during sideband formation. However, to explain *N* events *Dowden et al.* [1978] expand upon *Nunn's* [1974] theory with a model containing two interaction regions, one trapping particles resonant with the input wave and a second one trapping particles resonant with the developing emission. This improved theory is better at explaining some of the wider-bandwidth characteristics of growing signals. Still, it cannot explain saturation, and, because it is limited to trapped particles, may not be very realistic.

Helliwell and Inan [1982] describe a model in which the distributed interaction is simulated by two lumped elements: a buncher where the output wave field organizes the phases of entering electrons, and a radiator where the electrons radiate and add to the input wave field. They derive the magnitude and phase of the system loop gain by tracing the trajectories of a monochromatic stream of electrons with a single pitch angle but various phases with respect to the wave. Untrapped electrons are important here, and their model predicts saturation at large amplitudes. *Helliwell and Inan's* [1982] model shows the growth and phase advance that occurs during the initial part of a growing pulse. However, they predict that a negative phase shift may occur as the amplitude reaches saturation, an effect not seen in real data.

Carlson [1987] has produced the most comprehensive computer simulation of cyclotron-resonant growth to date. He develops a steady-state model which uses a range of energetic electron velocities $v_\parallel$ (the component of velocity parallel to the earth's magnetic field) and a full range of pitch angles α. He also develops a transient model similar in some ways to the two-port *Helliwell and Inan* [1982] model, but where the bunching and radiating processes are fully distributed in space. The transient model uses a full range of pitch angles α but only a single value of parallel electron velocity $v_\parallel$ (a function of α and position chosen to maximize growth) to keep computing time within bounds. Even with this limitation, it still reproduces very well the growth, phase advance, saturation, and even magnitude ripples during saturation that are seen in real data. The phase advance is parabolic, implying a linear increase in offset frequency, as often occurs, at least initially. However, there is a transient phase retardation at the very beginning of a pulse that is not seen in real data, and the model cannot reproduce an initial frequency offset. These discrepancies may be due to the limited range of $v_\parallel$ used. *Carlson* [1987] suggests that future models with a wider $v_\parallel$ range may be able to simulate features such as the BLI, emission triggering, positive frequency jump at pulse termination, and entrainment of emissions.

All four studies succeed in a qualitative way in explaining the increasing magnitude and advancing phase of a growing pulse, though *Nunn* [1974] predicts an initial phase lag that is not seen. *Dowden et al.* [1978] explain certain features of emission generation, such as the *N* event. *Helliwell and Inan* [1982] explain saturation, though with apparently incorrect phase effects. *Carlson* [1987] is the most successful of the four in simulating quantitatively both exponential growth and parabolic phase advance, as well as pulsations and saturation; but he also predicts an unobserved initial phase lag. None of these studies can explain the initial frequency offset that sometimes appears at the

beginning (or within 20–30 ms of the beginning) of a growing pulse.

Carlson's [1987] transient model is the most comprehensize and accurate with respect to its ability to reproduce the features of growing signals. It is also by far the most complex. I have a personal preference for theories with simple, analytic (if not closed-form) solutions. However, the trend to computer simulations is inevitable and probably necessary. The underlying processes are nonlinear and extremely complex. Facing a similar problem, weathermen use increasingly complex computer simulations with (though some may doubt it) gradually improving forecasting accuracy.

Summary. Whistler-mode signals often exhibit various growth-related features. Some of these features have been observed for a long time in magnitude plots and f-t spectrograms, such as an exponential increase in magnitude (temporal growth) followed by saturation, and the triggering of emissions.

Analysis of the spectral phase of constant-frequency signals from the VLF transmitter at Siple Station has now revealed several new growth-related features. The relative phase of a growing pulse increases with time, indicating that the received signal is at a higher frequency than when transmitted. When growth activity is weak, this phase advance may not begin until the magnitude has already increased a bit, say by 10 dB. As activity picks up, phase advance begins earlier and earlier. The advance in phase is often approximately parabolic with time (proportional to t^2), especially at the beginning of a pulse, meaning that the frequency offset increases linearly with time. However, often the advance is only linear (proportional to t) toward the end, evidence of a limiting or maximum value of frequency offset. Occasionally more complicated behavior occurs, but more than a momentary retardation in phase is never seen on a growing pulse. Received pulses usually begin at the frequency of the transmitted signal, but some seem to be offset in frequency even from the start, or at least within 20–30 ms of the start of the transmitted pulse.

The accepted mechanism for ducted whistler-mode growth is cyclotron resonance between the wave moving in one direction along a magnetic field line and energetic electrons moving in the other direction. Models using this mechanism have been developed to explain the magnitude and gross frequency characteristics of growing signals. However, only a few studies have been made which predict phase behavior. These particular studies have been successful in explaining phase advance once growth is well underway, but some predict a phase retardation at the beginning of a pulse that is not seen in the experimental data. No study has explained an initial frequency offset. While the physical laws governing the interactions of waves and particles are well-known, whistler-mode growth is a very complex problem for the same reason as weather forecasting: because of the sheer number of individual interactions which must be studied. The most successful models of growth use complex computer simulations.

4.2 Phase Behavior at Pulse Termination

In the previous section we saw examples of some phenomena associated with growing whistler-mode signals. Here we will look in more detail at what happens right at the end of a growing input pulse. This is a subject that has been studied several times before. In particular, *Stiles* [1974] examined triggered emissions using digital spectrum analysis, and was able to see features that had been hidden (or at least unnoticed) in analog spectrograms. Now we will look at some of these same features again, but with the benefit of phase analysis.

Emission Frequency and Phase Wrap-Up. Figures 4.5 and 4.6 show some variable-length pulses at 5500 Hz from Siple. These are some of the first signals received from the Siple Station VLF transmitter. This particular record has been presented many times before [*Stiles*, 1974; *Helliwell and Katsufrakis*, 1974; *Helliwell*, 1974; *Helliwell*, 1975; *Stiles and Helliwell*, 1975; *Stiles and Helliwell*, 1977; *Helliwell and Katsufrakis*, 1978; *Matsumoto*, 1979; *Helliwell*, 1983b], and may hold a record in this regard. Here it is once more.

All of the pulses in Fig.s 4.5 and 4.6 show exponential growth at about 100 dB/s. They all start from about the same level, and the peak amplitude reached depends on the length of the pulse. A falling emission is generated at the end of every pulse. However, not only do the longer pulses reach higher amplitudes due to prolonged growth, but they also trigger stronger emissions, and emissions which last longer and descend to lower frequencies before fading away. The spectrograms show that the 250-ms pulses each have a small BLI just before the emission. (The spectrograms also show the presence of multipath propagation with a second weaker path preceeding the main signal by about 50 ms. This second path doesn't affect the interpretation.)

The phase plots show a phase advance during growth that increases with time. The rate of relative phase advance (the frequency offset) also increases with time. It is difficult to measure the instantaneous frequency offset as a function of time for most of these pulses because the phase plots are just a bit too noisy. We can estimate the average frequency offset during each pulse, however, and we find offsets of 2.0, 3.2, and 4.1 Hz for the 150-, 200-, and 250-ms pulses, respectively. While these average frequency estimates are crude, they are consistent with an instantaneous frequency offset during growth which is zero at the beginning of each pulse and increases at 32 Hz/s. ($\Delta f = 0 + 32t$ gives 2.4, 3.2, and 4.0 Hz average frequencies.) So, while these pulses are shorter and grow faster than those we saw in the previous section, they have the same kind of magnitude and phase behavior.

When we look at the phase at the end of each pulse we discover a remarkable feature. All of the fallers, even those on the 150-ms pulses which seem in the spectrogram to just drop off the end of each pulse, start at a frequency *above* the input signal, even above that of the growing, phase-advancing signal. In fact, judging by the phase plot, the 250-ms pulses in Fig. 4.6 seem to jump almost instantaneously between the offset growth frequency and some higher frequency where the emission starts. The analysis step time in the magnitude-phase plots was $t_{step} = 2$ ms. The first and second pulses in Fig. 4.6 seem to change frequency in about twice this time (*i.e.*, twice the interval between individual dots in the plot) or about 4 ms. (This is roughly the resolution time of the BW = 80 Hz filter used.)

The benefit of signal phase information here can be understood by comparing our results with the earlier work of *Stiles* [1974, pp. 123–132]. He used spectrograms and A-scan plots (amplitude *vs.* frequency at a given time) of station NAA at 14.7 kHz to determine how emissions separated from the triggering signal. He decided that an emission began at the frequency of the input signal, and then rose rapidly over some 30 ms to its own frequency which might be some 250 Hz higher. However, as he mentioned, any model in which the emission begins within 50 Hz or so of the input

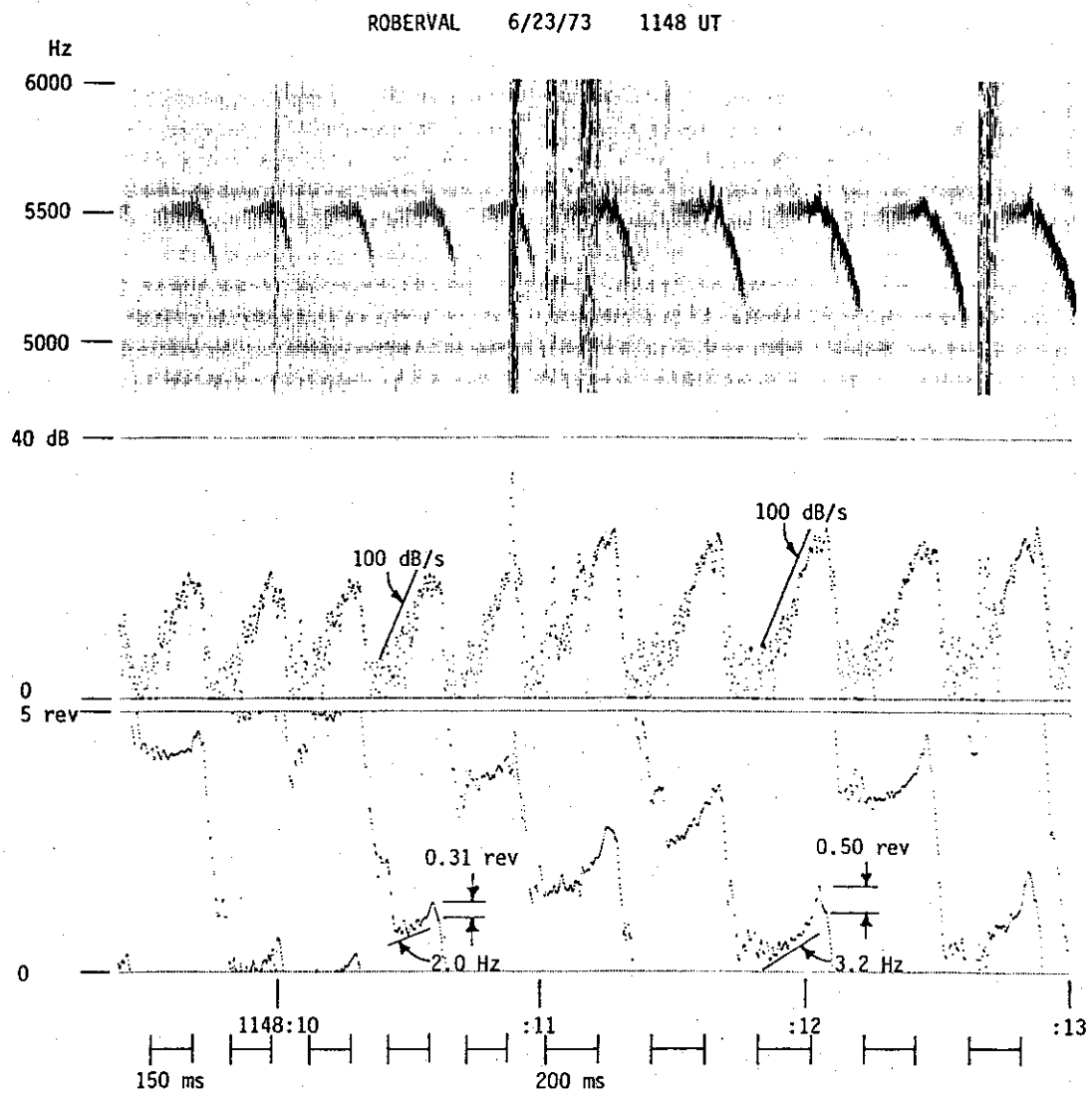

Figure 4.5. Spectrogram and magnitude-phase plot (BW = 80 Hz, t_{step} = 2 ms) of variable-length pulses at 5500 Hz with termination fallers. Both 150-ms and 200-ms pulses show the same growth rate, but the longer pulses show a larger average frequency offset and a larger phase wrap-up at termination.

frequency would agree fairly well with the data, because this was the frequency resolution of the analysis filters. With phase information we can resolve changes in instantaneous frequency with a rapidity limited only by the signal-to-noise ratio of the data. We will use this technique in the examples which follow.

However, while the initial frequency of the emission is above that of the input signal, the slope of the emission is always downwards in frequency with time. The phase plots show the beginning of each emission as a small downward-concave arch at the end of each pulse. The instantaneous frequency offset from the transmitted frequency is equal to the slope of this curve. Since the phase is downward-concave, its second derivative $d^2\phi/dt^2$, and thus the rate of change of frequency df/dt, is negative.

The arch in the phase plot at the end of each pulse rises as long as the emission remains

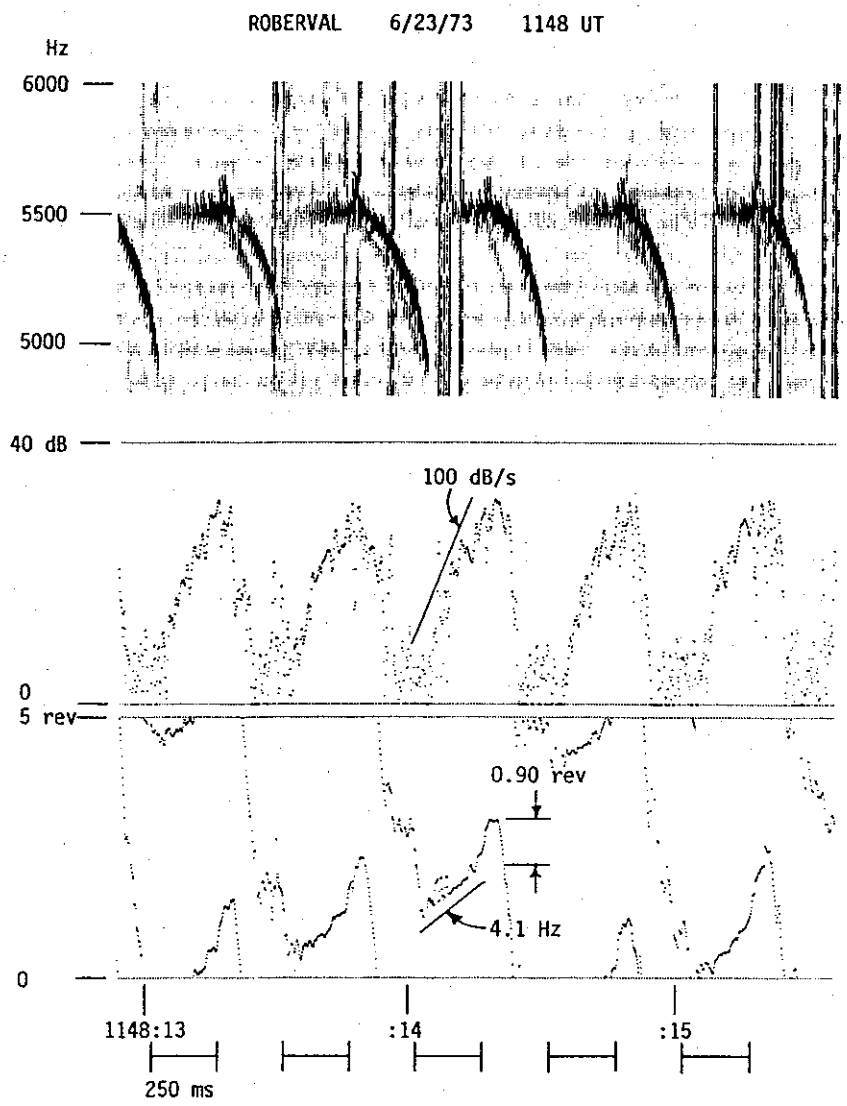

Figure 4.6. Variable-length pulses as in Fig. 4.5. 250-ms pulses show an even larger average frequency offset and termination phase wrap-up.

above the input frequency, and reaches its peak value just as the emission falls back through this frequency. This peak value is termed the *phase wrap-up*, and depends on both the initial frequency of the emission and on how long it remains above the input frequency. (It is, of course, just the integral of the offset frequency Δf over the time it takes the emission to fall through the input frequency.) The wrap-ups for the 150-, 200-, and 250-ms pulses are seen to be about 0.31, 0.50, and 0.90 revs, respectively. So longer pulses, which grow more, reach larger amplitudes and frequency offsets, and generate stronger emissions, also show bigger phase wrap-ups. It is not clear from the figures whether this is because the emissions on longer pulses start at higher frequencies or just take longer to fall back through the input frequency. It looks as if both factors are at work.

The 5500 Hz pulses in Fig.s 4.5 and 4.6 are at a fairly high frequency for signals from the Siple transmitter. An interesting question is how high in frequency can Siple signals be and still show growth? *Carpenter* [1968] has shown experimentally that most ducted whistler-mode signals do not propagate at frequencies above half the equatorial gyrofrequency, presumably because signals

above $f_H/2$ leak out of the duct as predicted by ray theory. For the usual Siple-Roberval path at $L = 4.2$, the equatorial half-gyrofrequency is $f_{Heq}/2 = 5.90$ kHz using a dipole field model. (It may more realistically be about 5.0 kHz using *Seely's* [1977] distorted model magnetosphere.) Are these 5500 Hz pulses close to $f_{Heq}/2$? There are some swishy whistlers in this record, but they are too diffuse and ill-defined to be much help. However, some of the risers triggered by the longer pulses in this transmission continue up to at least 6500 Hz, suggesting that the half-equatorial gyrofrequency is at least this high. We conclude that this particular transmitter signal was probably on a path at $L = 4.07$ or lower, at a frequency of no more than about $5500/(2 \times 6500) = 0.42 f_{Heq}$.

Speed of Transition to Emission Frequency. Figure 4.7 shows two segments of a NOSI (NOise SImulation) transmission, including the last second of CW (single frequency) signal at 2600 Hz. The NOSI format was designed to simulate band-limited noise while running the transmitter at full power [*Helliwell et al.*, 1986b]. Instead of modulating the amplitude of the transmitter (with a decrease in average power), the frequency was changed every 10 ms in a "random" manner by selecting the next entry from a table of precalculated frequency offsets. The intent was to generate a wide-band noisy signal to mimic mid-latitude hiss, and to study the response of the magnetospheric growth mechanism to signals of constant power but varying bandwidths. The signal was transmitted in five-second intervals at two frequencies. In each interval the bandwidth (peak frequency excursion from the table) was reduced step by step, until during the last second a constant-frequency (zero-bandwidth) tone was transmitted. Figure 4.7 shows the end of two intervals of transmission at 2600 Hz, including the one-second CW tone, followed a second later by the beginning of an interval of transmission at 2200 Hz. The signal at 2200 Hz from 1406:19–:20 in the left-hand spectrogram at top is a two-hop echo of a previous CW tone at that frequency. A weak whistler crosses the signal in the right-hand spectrogram.

The interesting feature here is the behavior at the end of each CW pulse. The first pulse is seen to trigger a riser that starts above the 2600 Hz input frequency, remains at about the same frequency for a quarter of a second, and then drifts up and dissipates, lasting almost 1.5 s. The second pulse generates a brief faller, which also starts above the input frequency. The expanded spectrograms and magnitude-phase plots in the bottom of Fig. 4.7 show the behavior at termination in more detail. The end of each pulse is marked by a brief, weak BLI-like transient (stronger in the first pulse) followed by the emission. Note that the magnitude of the emission is almost the same as that of the pulse before it. This is a common feature of emissions. After a period of growth the total signal energy in the interaction region in the magnetosphere will be much larger than the transmitted input signal controlling it. When the input signal disappears, this energy (now a termination emission) does not vanish instantaneously.

However, the frequency of the emission is quite different from that just before the end of the pulse. The phase plots show a rapid change from the offset frequency of the growing pulses (+1.9 and +2.8 Hz) to the frequency of the emissions (+56 and +71 Hz). The change in frequency is seen to take place in a very short time. The change from 1.9 to 56 Hz in the first pulse takes about 50 ms. The change from 2.8 to 71 Hz in the second pulse takes only about 10 ms. This latter change is very fast (though still not instantaneous).

I might add that our resolution in measuring the speed of frequency change here is not limited by the 320 Hz bandwidth of the analysis filter. Tests with a synthetic signal show that the phase plot could reveal transition times as short as 1.25 ms. The DFT filter spacing is $f_D = 1/NT = 200$ Hz in this case. The overlap correlation for windows spaced by $0.25NT = 1.25$ ms is only 57.4% from Table 2.5.

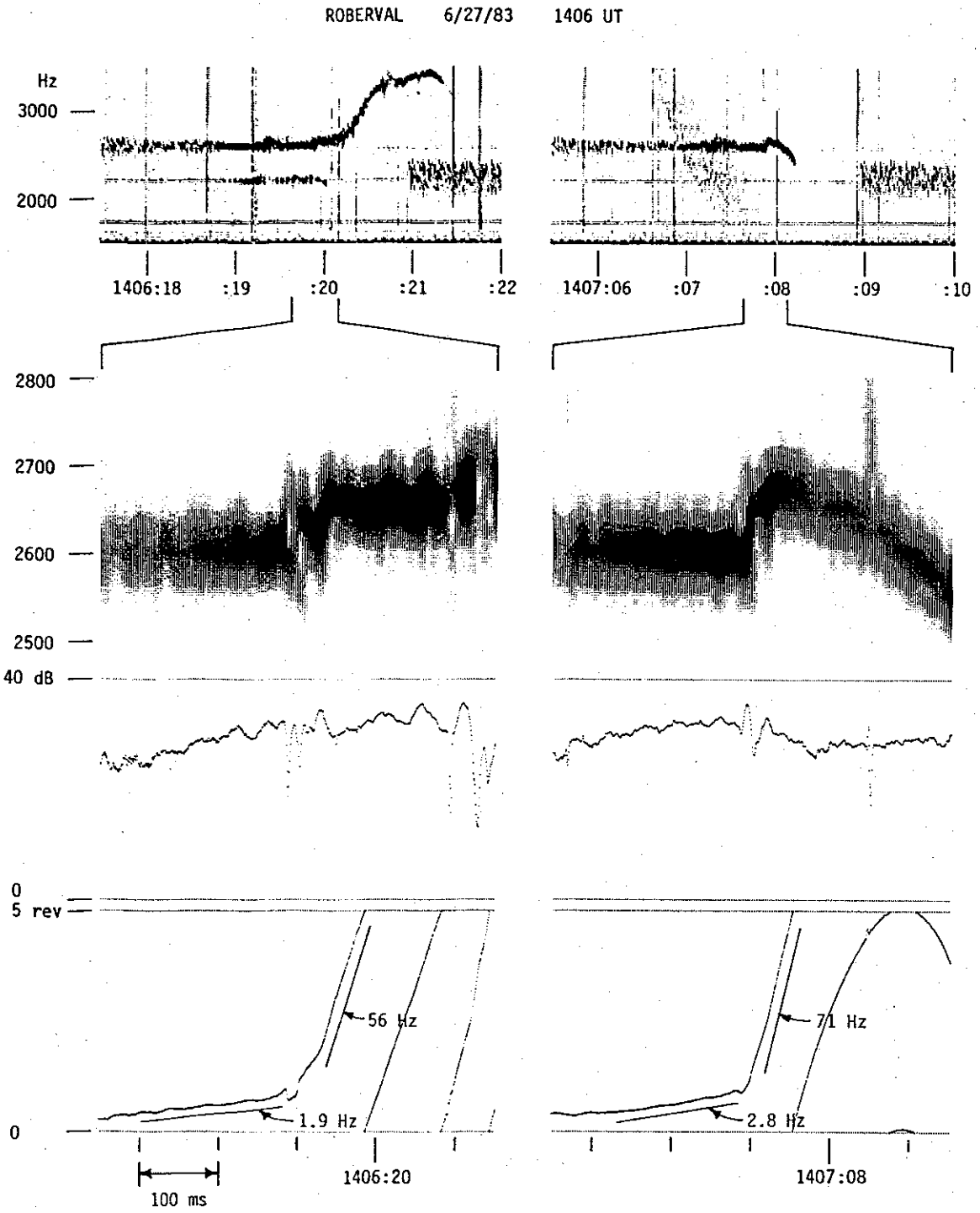

Figure 4.7. One-second CW pulses at 2600 Hz at the end of a NOSI transmission showing the speed of change in frequency at the beginning of a termination emission. BW = 320 Hz in the magnitude-phase plots. In the first case the transition from the growing signal to the emission takes about 50 ms; in the second case less than 10 ms.

Right at the point of transition between the end of the growing pulse and the beginning of the emission, both pulses show a brief phase jump of about 0.2 rev that lasts about 10 ms. It coincides with dips in the magnitude plot and (at least in the first pulse) a weak BLI-like event in the spectrogram. One of the possibilities we must always consider in a case like this is the effect of dispersion. Perhaps the change from end-of-pulse to emission is instantaneous at the interaction region near the equator, but being at different frequencies these two signals are overlapped at the ground, and the transient features are caused by beating between the two signals. The dispersion of the whistler in the right spectrogram in Fig. 4.7 is -3 kHz/s at 2600 Hz. Since the whistler is a two-hop signal, the dispersion for the half-hop path from equator to ground will be -12 kHz/s. The 68.2 Hz frequency jump at the end of the second pulse would cause the beginning of its emission to overlap the pulse end by $68.2/12000 = 5.7$ ms. During this time the two signals will differ by $68.2 \times 0.0057 = 0.39$ cycles, or undergo 39% of a beat. While this is significant, it does not seem quite enough to explain the transients in Fig. 4.7. We may also be seeing the effects of multipath propagation involving two very closly spaced paths.

One final interesting point in Fig. 4.7 is the difference between the generation of a riser in one case and a faller in the other. It is not known what factors determine whether a termination emission will be one or the other. Generally speaking, the greater the level of growth activity and the larger a signal is allowed to grow, the more likely it is to trigger a riser at termination. However, in this case both pulses reached approximately the same amplitude at the end. The first pulse was growing a bit faster just before the end; is recent growth rate the determining factor? On the other hand, the second pulse had a larger frequency offset, and its emission began at a higher frequency, characteristics we might normally associate with stronger emissions. Why the first should be a riser and the second a faller is not clear.

Instantaneous Change from Pulse to Termination Emission. Finally, in Figure 4.8 we see the ends of two half-second pulses in the ULF75 transmission described in Sec. 4.1. These two pulses occurred at times between the first and second pulses shown in Fig. 4.2. The pulses in Fig. 4.8 are plotted to the same expanded scale used in Fig. 4.7. Both pulses show a strong BLI, much stronger than those in Fig. 4.7, followed by emissions that are only slightly above the input frequency. The phase plots show the frequency offsets at the end of the growing pulses to be 3.5 and 3.2 Hz, and to change almost instantaneously to the 100 Hz offset of the BLI. The BLI's last 10–20 ms, and then the emissions begin. The first emission is about 20 Hz above the input frequency, and the second one is only slightly above the input.

From the phase plots we can estimate the time of transition from the end of the pulse to the BLI. The analysis step time here is $t_{step} = 0.56$ ms, and the transitions take place within the space of three or four dots on the plot, or in about 2 ms. Because of the lower signal-to-noise ratio here compared to the signals in Fig. 4.7 (though it is still quite good), a narrower analysis filter had to be used, with a corresponding decrease in time resolution. In this case the filter bandwidth was BW = 160 Hz, with a sequence length $NT = 10$ ms. The step time for 50% window correlation is about one-quarter of this, or 2.5 ms (twice that of Fig. 4.7). This is effectively the time resolution of the analysis. Even if the frequency change from pulse to BLI were instantaneous, it would still appear to take about 2.5 ms in the phase plot. Thus, as far as we can ascertain, the transition in this case is instantaneous.

Concurrent with the start of the BLI is a sudden dip (or dips) in magnitude. Such a dip is seen repeatedly throughout this particular record. Each dip lasts 15–30 ms. The dip is only a few dB for the first fallers after 1416:20, but is typically 10-15 dB for later pulses, and sometimes over 25 dB.

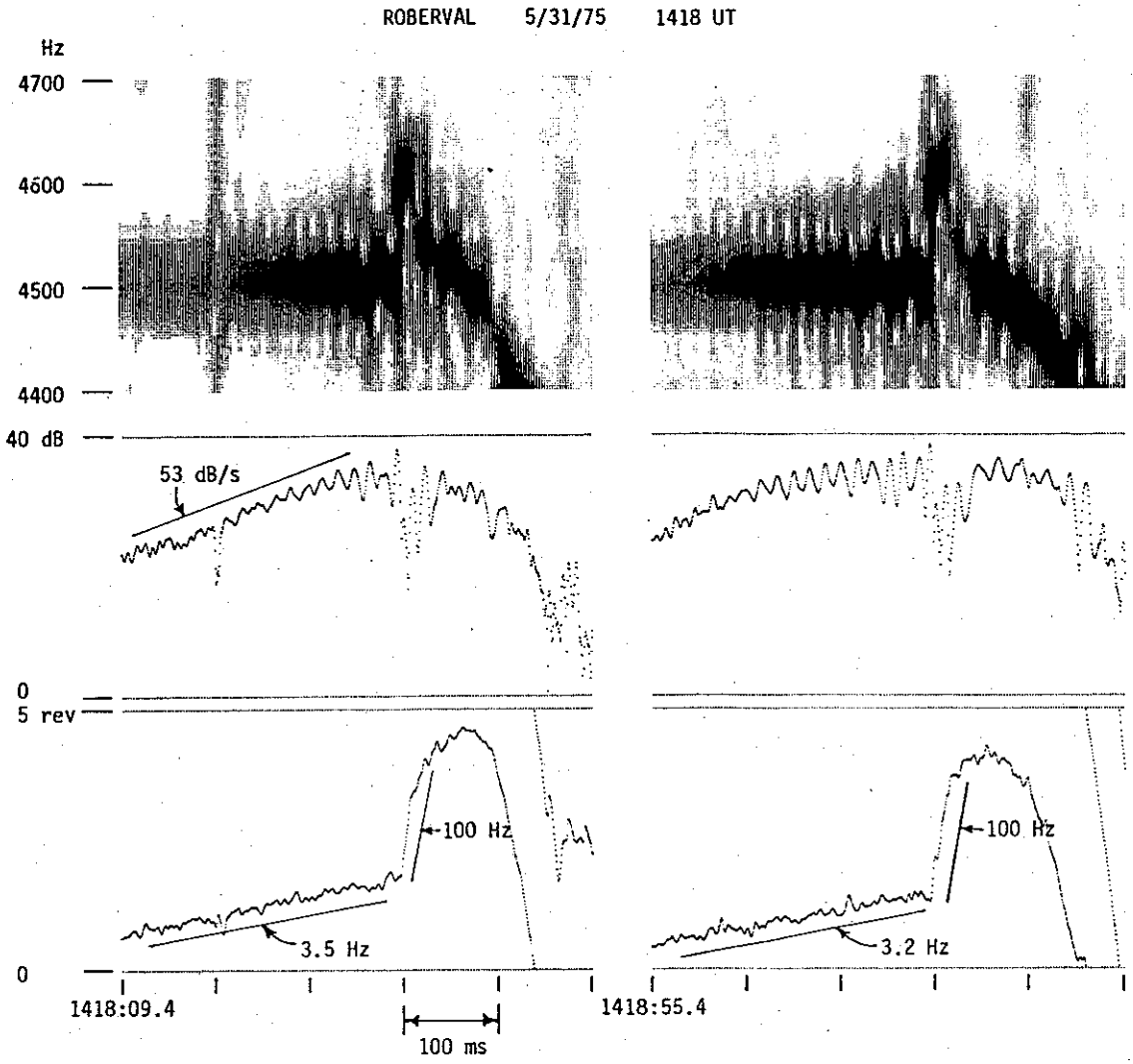

Figure 4.8. Emission behavior at the end of two half-second pulses at 4500 Hz. BW = 160 Hz and t_{step} = 0.56 ms in the magnitude-phase plots. Compared to the pulses in Fig. 4.7, the change in frequency at the end of the pulse occurs instantaneously (in $\approx$ 2 ms). The falling emission after each pulse begins with a very strong BLI.

The dips in Fig. 4.8 are roughly 20 dB. Is this caused by the triggering mechanism, or is it the result of dispersion? In this case the 4.5 kHz signals are above the nose frequency, which is roughly 3.9 kHz. Dispersion here will delay the higher-frequency BLI with respect to the growing pulse, and so separate them by the time they reach the ground. This might cause the magnitude to dip between the pulse and the BLI. The two-hop whistler dispersion at 4.5 kHz is about +7 kHz/s, so the half-hop dispersion will be +28 kHz/s. With a frequency jump of 97 Hz this gives a separation between the end of the pulse and the BLI of 97/28000 = 3.5 ms. This is small compared to the duration of the magnitude dips, and we conclude that they are not due to dispersion but may be inherent in triggering. (The pulses in Fig. 4.8 show three or four sharp dips at the BLI. This might be caused by dispersion acting on signals on three or four paths with slightly different group delays. However, in this case the null at each dip, if all paths contribute equally, should only be about $2/3 = -3.5$ dB. This is much less than the dips actually seen.)

The observent reader will notice that the pulse-to-BLI transitions in Fig. 4.8 occur at 1418:09.700 and 1418:55.700, whereas in Sec. 4.1 we determined a group delay of $t_g = 2.18$ s for these pulses. So the beginnings of these pulses must have been received at 1418:09.180 and 1418:55.180; that is, 0.520 s before the putative ends. Indeed, these pulses are about 20 ms too long.

There are at least two possibilities here. One is that we have a case of multipath propagation with two paths only 20 ms apart in time. If this is true, we must ask why the transition to the BLI occurs so rapidly. Two signals, growing at approximately the same rate, would make for a more complicated transition, and, most importantly, one which would start 20 ms earlier at the end of the first signal. This objection would be overcome if only the later path showed growth, as it would dominate the earlier signal by the end of the pulse. This might be possible, though paths so close together in group delay (and presumably physically close in the magnetosphere) often show similar growth activity. However, any suggestion of multipath propagation must be examined in light of Fig.s 4.1 and 4.2. While there are some magnitude irregularities at the beginning of the pulses shown, only the one in the right-hand pulse in Fig. 4.2 might be as short as 20 ms duration. And none of the pulses shows a step in phase after 20 ms as is typically seen when a second signal arrives.

Another possibility is that the transition to the BLI, though well defined, does not occur until about 20 ms after the end of the input signal. That is, the growth process continues as if nothing had happened for 20 ms, and then generates a BLI. This possibility seems a bit bizarre. All in all, I have no explanation for the 20 ms stretching of these pulses.

Summary. When a growing pulse is received, the signal is found to continue past the end of the transmitted pulse, usually in a falling emission, though rising termination emissions are occasionally seen if growth is sufficiently developed. The faller (or riser) begins at a frequency above that of the growing signal, itself offset a few Hertz above the frequency of the input or transmitted pulse. The faller is often, though perhaps not always, preceeded by a band-limited-impulse or BLI, a short transient event whose energy is mostly above the input frequency. The BLI may be heralded by a momentary (few millisecond) dip (or dips) in magnitude. Except for any transients, the magnitude of the signal at the beginning of the faller is about the same as it was at the end of the pulse, though it may change rapidly from then on.

Since the faller starts above the pulse frequency, the relative phase of the signal advances after the end of the pulse until the faller has drifted back through the input frequency. The total advance in phase at this point, the phase wrap-up, depends on both the offset frequency at the beginning of the faller and the time it takes to drift back down. For short pulses which terminate before saturation, the phase wrap-up is correlated with pulse length, being bigger for longer pulses. Both the initial emission offset frequency and the time to drift back through the input frequency seem to increase with pulse duration. Longer pulses also reach higher magnitudes and show larger frequency offsets prior to termination, of course.

The change from the frequency of the growing pulse to that of the BLI or emission may occur very rapidly. In Fig. 4.8 the change from the 3 Hz offset of the pulse to the 100 Hz offset of the BLI occurs in less than the 2.5 ms resolution time of the analysis filter, and is effectively instantaneous.

The transients that mark the beginning of a BLI or emission, though brief, are too long to be explained as the results of dispersion. They may be inherent in the triggering process. They may also be due partly to contamination by multipath propagation, which is always hard to rule out. The pulses in Fig. 4.8 seem about 20 ms too long, but the cause of this is not clear.

4.3 Pre-Termination Emission Triggering

All of the pulses we have considered so far show a more-or-less continuous increase in relative phase during growth. However, a steady-state signal cannot have a continuously advancing phase any more than a model airplane with a rubber-band motor can have its propellor wound up without eventually snapping. In the case of growing whistler-mode signals, snapping occurs after a phase advance of two to three revolutions and results in the generation of a pre-termination emission. These emissions are always risers; fallers are never generated in the middle of a growing signal.

Pre-termination Emissions and Regrowth. Figures 4.9 and 4.10 show four one-second pulses at 2790 and 3030 Hz from the DIAG1 (DIAGnostic, version 1) transmission format. This is a sequence of constant-frequency pulses at various frequencies interspersed with rising and falling ramps with several slopes. It is very useful for monitoring growth activity as a function of frequency and checking for the presence of multipath propagation. The tail ends of some of the falling ramps (and a riser triggered by one) can be seen just preceeding the constant-frequency pulses in the figures. This particular record was selected because growth rates are high enough here (and pulse durations long enough) to trigger pre-termination emissions, and yet activity is not so high that multiple triggering on several paths at once will obscure the underlying processes. There is a bit of multipath propagation at this time, with a weak path preceeding the main one by about 100 ms, but these signals are still fairly well-behaved.

The two pulses in Fig. 4.9 show growth and phase advance, and each triggers a pre-termination emission after about 400 ms. The growth rate for the first pulse at 2790 Hz is somewhat less than for the second one at 3030 Hz. This is a frequency effect in this case and not due to changing activity with time. Growth often is a fairly narrowband process. At this time, pulses below 2600 Hz showed no growth at all.

The first emission is triggered in each case when the phase has advanced about 2 or 3 revs. The exact point where the emission starts is hard to define in these pulses because there is not such a sharp change in frequency as we saw before in the case of termination emissions. The phase plots show that shortly after each emission starts it is about 44 Hz above the input signal. The emission is a riser in each case, and slowly drifts away from the input. (A second, weaker emission is also triggered at about the same time on the 3030 Hz pulse, another indication of multipath.) Once the emission has drifted sufficiently far from the input signal, the pulse is seen to start growing all over again with about the same growth rate as it had initially. A second riser is triggered after an advance of about 2 revs near (or just at) the end of the pulse. The second emission seems very much like the first one.

After the first emission separates, the signal amplitude as shown in the magnitude plot is reduced to approximately the same level as at the beginning of the pulse, though it may be slightly higher. The level when growth restarts is difficult to determine here because the first emission is about 25 dB stronger than the input signal, and takes some time to drift out of the passband of the analysis filter, by which time the second growth period is already underway. Just when the emission drifts out of range, there are some brief but deep nulls in the magnitude plot, presumably caused by beating between the emission (on the skirts of the filter) and the regrowing signal. However, there is also evidence that in some cases of pre-termination triggering the input signal is suppressed below its initial value immediately following the emission [*Helliwell*, 1983a, Fig. 3; or *Helliwell et al.*, 1985, Fig. 6]. The suppression typically lasts up to 100 ms. Suppression can be seen in the spectrograms in Fig. 4.9 as a brief white space at the input frequency just before regrowth. (The spectrogram filter bandwidth, 40 Hz, was smaller than that used in the magnitude-phase plots, 160 Hz, allowing

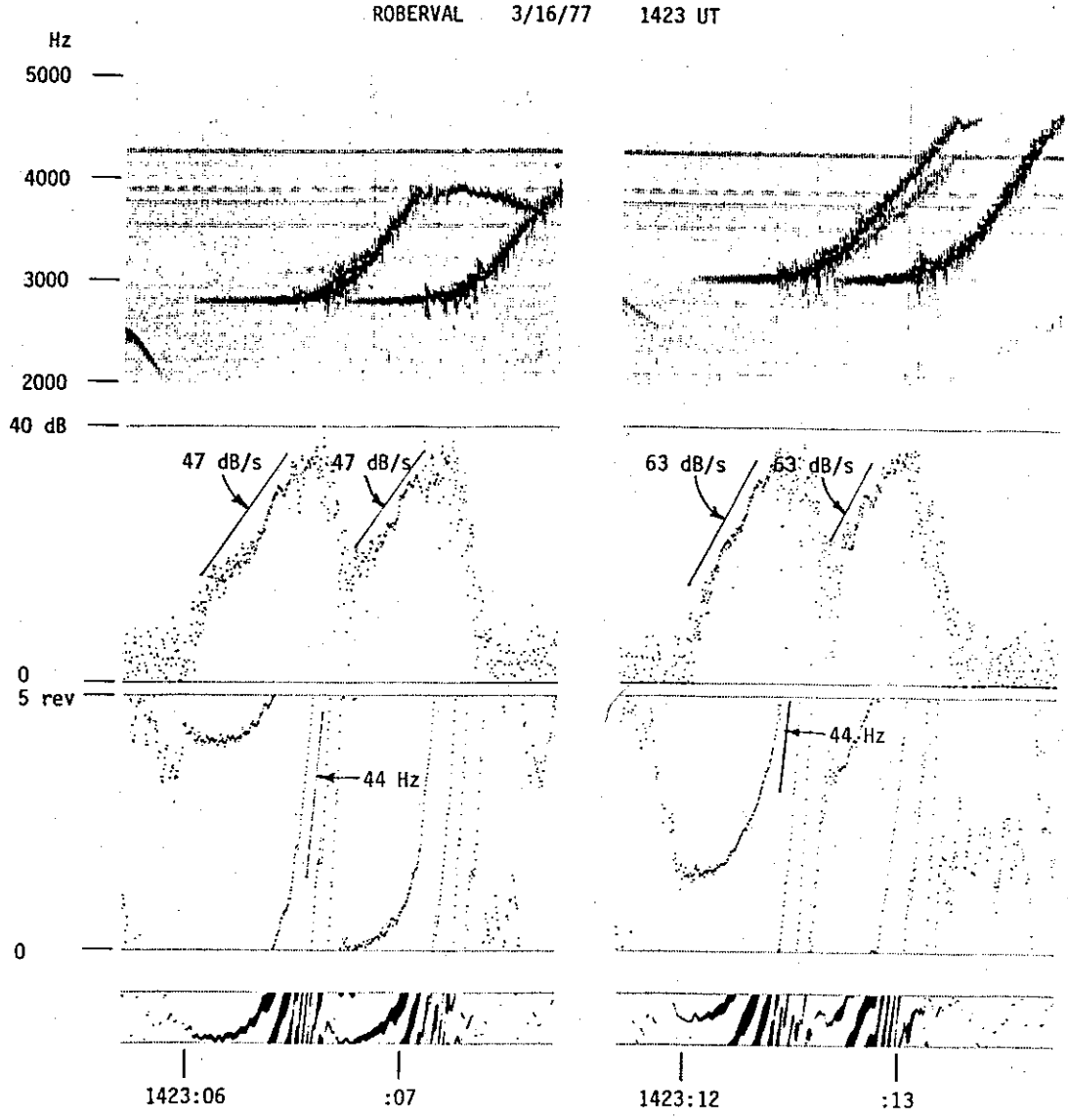

Figure 4.9. One-second pulses at 2790 Hz (left) and 3030 Hz (right) showing the triggering of pre-termination rising emissions. The magnitude-phase plots (BW = 160 Hz) show that successive risers are similar, and begin when the growing signal has advanced in phase about 2 revs. The gray-scale phase plots at the bottom (BW = 40 Hz, P_{span} = 1 rev) show that growth restarts at the initial phase after separation of the first riser. Spectrogram BW = 40 Hz.

us to see this effect.)

There is little doubt about the phase behavior of the signal. The gray-scale phase plots at the bottom of Fig. 4.9, plotted with a range of P_{span} = 1 rev, show that the regrowing signal starts with the same phase it had during the first period of growth. That is, as far as phase is concerned, once the first emission has separated, conditions are reset and growth begins *ab initio*.

The pulses in Fig. 4.10 (both at 2790 Hz) are similar at the beginning to the previous ones. Again there is growth with phase advance, and the triggering of a pre-termination riser. In the first pulse the riser begins after a phase advance of 2 revs, and makes a rapid change in frequency to an offset about 44 Hz above the input frequency. In the second pulse the emission begins after a slightly

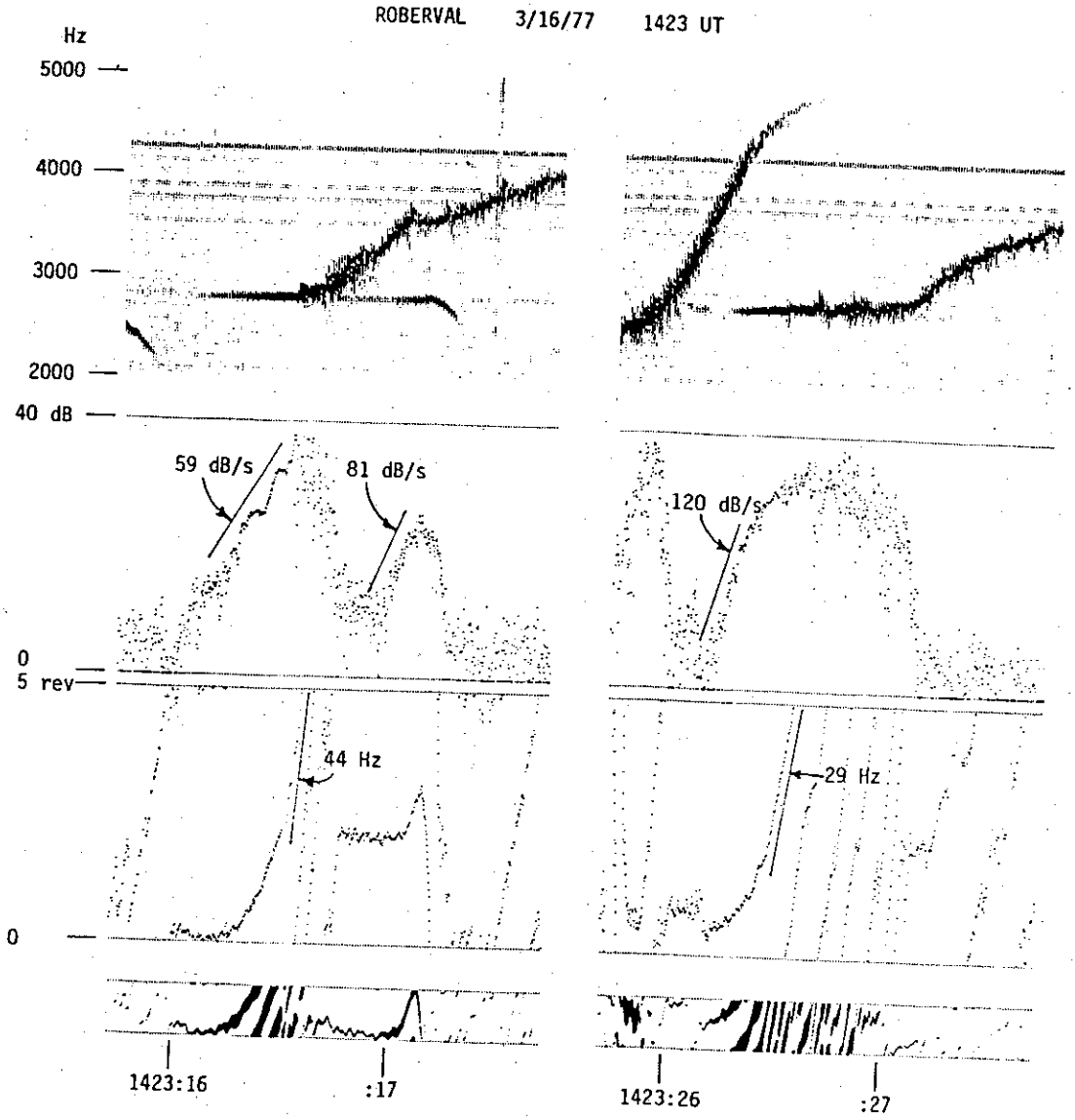

Figure 4.10. Two one-second pulses at 2790 Hz as in Fig. 4.9. In the left-hand case, the signal has not regrown sufficiently by the end of the pulse to do more than trigger a short faller. In the right-hand case the first emission never completely separates from the input signal and suppresses all subsequent growth.

smaller advance and is only 29 Hz above the input frequency. The growth rates here are different than those in Fig. 4.9, at least according to the labels in the figure, but some of this difference may be due to changes in the rate of growth with time. Especially in the last pulse, growth is less regular and tapers off faster as saturation is approached. The fact that the initial growth rate seems higher is compensated by a lower growth rate later on.

Once the emissions have separated, the pulses in Fig. 4.10 show behavior different from those in Fig. 4.9. The left-hand pulse begins to grow again, but by the end has only grown by about 15 dB and has advanced very little in phase. Instead of triggering another riser, this pulse can manage only a weak faller as a termination emission. Because regrowth comes later in this case, it is easier to see that the pulse does start regrowing at its initial amplitude. The gray-scale phase plot below

confirms that regrowth starts at the initial phase.

The right-hand pulse in Fig. 4.10 shows even less regrowth. In this case the first emission never separates far enough from the input signal to allow the input to regrow at all. It remains about 30 Hz above the input until almost the end of the pulse. The narrower filter (BW = 40 Hz) in the gray-scale plot may show a bit of the input pulse reappearing during the last 100 ms.

Generalizations about Triggering. From the pulses in Fig.s 4.9 and 4.10 and others seen in the record at this time we can generalize as follows. A pre-termination emission is triggered on a growing pulse when the phase has advanced 2–3 revs. Such an emission is always a riser. It may start with a BLI.

When multiple emission triggering occurs before the end of a pulse, the first two events (and sometimes later ones) are similar. The magnitude of the first riser may be slightly higher (say 2 dB) than that of the second one, but otherwise they look alike. (Subsequent risers may be slower to form and reach lower peak magnitudes.) Phase behavior during the intervals of growth prior to the first two emissions may be very similar. Of course, it is often hard to see what happens at the beginning of the second growth period because of energy from the previous riser leaking through the passband of the analysis filters. On gray-scale plots like those at the bottom of Fig.s 4.9 and 4.10 where the filter bandwidth is only 40 Hz, the phase traces during the first and second periods of growth and their subsequent emissions often look identical. The conclusion is that once an emission has separated from the driving signal by, say, 50 Hz or more, it leaves few aftereffects and a second period of growth can start almost from scratch. Each subsequent period of growth starts at the same phase and approximately the same amplitude as at the beginning of the pulse.

In some cases of pre-termination triggering the input signal seems to be suppressed below its initial amplitude for a brief interval immediately after each emission. Only after this interval, lasting up to 100 ms, does the signal again appear and begin to regrow. It is not known how common post-triggering suppression is. Pre-termination triggering usually occurs during active growth conditions, when multipath propagation is also common. It is possible that post-triggering suppression always occurs, but is only seen when it is not masked by concurrent signals from other paths.

Behavior at the end of a pulse depends on the amount of growth that has occurred at that point. First, if the pulse is in the process of regrowth (any previous pre-termination emission has drifted sufficiently far away in frequency) but the phase advance is not more than about 1 rev, at the end of the pulse we will see the generation of a faller with a phase wrap-up of a few revs at most. Behavior in this case is identical to the termination fallers shown in Sec. 4.2.

Second, if the pulse has grown such that its phase advance is 2 revs or more, at the end of the pulse we will likely see a riser much like the pre-termination ones. In other words, if the emission has started at the end of a pulse, or is just about to start, then the end doesn't seem to have much effect. If the emission is already underway when the end of the pulse is reached but has not yet drifted outside the passband of the analysis filter (or if post-triggering suppression occurs), the actual end may not be observable.

4.4 Sideband Generation

We have already seen several examples of sidebands, signal components that appear spontaneously at frequencies near the frequency of a transmitted signal. For instance, in Fig. 3.6 the LICO1 signal with two tones 30 Hz apart was seen at the receiver to have additional components 30 Hz above and below the transmitted pair. These sidebands are relatively constant with time, and show phase changes from duct motion identical to the phase changes at the transmitted frequencies. That is, the sidebands are phase coherent with the transmitted signal.

Sidebands may also occur on single-frequency input signals. In Fig. 4.4, a one-second growing pulse was seen in the spectrogram to develop symmetrical sidebands for a brief time as saturation is reached. These sidebands are transient, and appear about 60 Hz either side of the transmitted signal, or "carrier." Other growing pulses were seen to develop regularly-spaced ripples in amplitude at saturation, for instance, the 50–70 Hz ripples in Fig.s 4.2 and 4.8. If we had used sufficiently narrow analysis filters in the spectrograms, we would have found that these ripples corresponded to sidebands. Temporal growth of the input components is not necessary for sideband generation. The linearly-propagating pulses in Fig. 3.5, which showed no growth and no phase changes except those due to duct drift, still develop amplitude ripples, in this case at an 11 Hz rate.

Phase analysis is useful in two ways when studying sidebands. First, as we will see in the first example, signal magnitude and phase behavior reveals the general nature of the mechanism that causes sidebands, though without revealing all the details of the process. Second, relative phase measurements allow us to determine the instantaneous frequency relationships between input signals and sidebands. As we will see, multiple sidebands are often harmonically related, and some sidebands seem to occur at particular frequency offsets.

4.4.1 Sidebands Due to Two-Tone Transmissions

As we saw in Section 4.1, the growth process in the magnetosphere distorts a whistler-mode signal, changing its amplitude and frequency in complicated ways. An amplifier which distorts a signal waveform can cause intermodulation distortion, mixing signal components at different frequencies to produce new components, including sidebands. With this in mind it isn't surprising that magnetospheric distortion might cause a two-tone input signal to develop sidebands. However, the distortion produced by magnetospheric growth is not the same as distortion in an electronic amplifier. The growth mechanism is too slow to distort the individual cycles of an input waveform. It is the envelope of the whistler-mode signal (a complex-valued function of time) which shows growth and phase advance. Thus magnetospheric distortion is really envelope distortion.

Waveform distortion in an amplifier is usually instantaneous. The output voltage $v_{out}(t)$ at a given time depends only on the input voltage $v_{in}(t)$ at that time and the amplifier gain g, as $v_{out}(t) = g \cdot v_{in}(t)$. Distortion occurs because the gain g is not constant, but is a function of the input voltage, $g\big(v_{in}(t)\big)$. The gain function of magnetospheric growth is different in that it depends not only on the present input signal, but on past conditions as well. Growth exhibits a kind of memory. This is because cyclotron-resonant waves and particles are travelling in opposite directions through the growth interaction region. An output signal can affect particles moving into the interaction region, which will in turn affect output signals at some later time, as in the feedback model of *Helliwell and Inan* [1982]. Magnetospheric gain is really some function $g\big(v_{out}(t-T),\ldots,v_{out}(t);v_{in}(t)\big)$ that depends not only on the present input, but on previous output signals back to some time T before the present. The time T is the *loop delay* of *Helliwell and Inan's* [1982] model, the time for a particle to transit the interaction region going one way plus the time for a wave to travel back going the other way. It is probably in the range 50–100 ms for typical Siple signals. (The magnetospheric gain

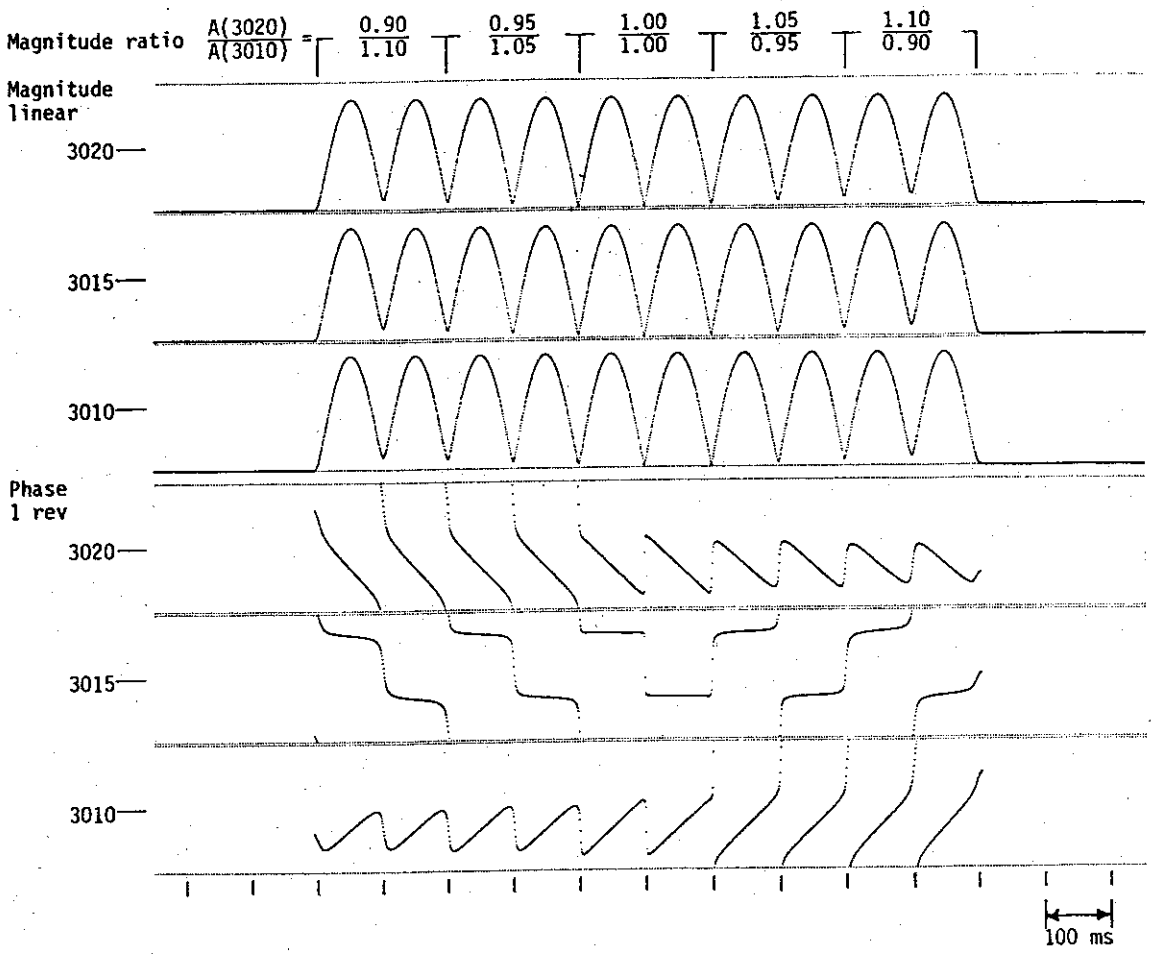

Figure 4.12. Magnitude and phase plots of a synthetic two-tone signal to aid in interpreting Fig. 4.13. Scale at top shows the relative magnitudes of the two components at 3010 and 3020 Hz, which was changed every 200 ms. Filter bandwidths (BW = 80 Hz) are wide enough that all magnitude plots are identical. However, the phase plots can be used to estimate the relative component magnitudes as explained in the text. Note the 1/2-rev alternation in phase with each beat at the mean frequency.

but since the average is one-sided in time it causes transient events to last longer than expected and to be asymmetrical. For instance, the reader will notice that the group of spherics near 1324:27 lasts longer in the gray-scale plot that it does in the spectrogram. The pulses themselves are extended in time and die away slowly because of averaging.

Before we examine the magnitude and phase structure of the pulses in Fig. 4.11 in detail, let's review what an undistorted two-tone signal should look like. Figure 4.12 shows magnitude and phase plots of a synthetic two-tone signal containing components at 3010 and 3020 Hz, just like the first two pulses in Fig. 4.11. Analysis filters have been synthesized at both of the input frequencies, and at the mean frequency of 3015 Hz as well. The bandwidth of the filters is 80 Hz, so all filters will pass both signal components and their magnitude traces are nearly identical. The phase traces are different, however, because of the different frequencies of the phase references. The amplitudes of the two components were changed every 200 ms, while keeping their sum constant, as labeled at the top of the figure.

The waveform $s(t)$ of a two-tone signal with components of equal amplitude A and frequencies f_a and f_b (say 3010 and 3020 Hz) can be expressed as

$$s(t) = A\cos(2\pi f_a t) + A\cos(2\pi f_b t) = 2A\cos\left(2\pi\frac{f_b - f_a}{2}t\right)\cos\left(2\pi\frac{f_b + f_a}{2}t\right). \qquad (4.1)$$

In the right-hand expression we have written the sum of two independent tones as the product of two signals: the carrier signal at the mean frequency, $(f_a + f_b)/2 = 3015$ Hz; and the modulation signal at half the difference frequency, $(f_b - f_a)/2 = 5$ Hz. When we plot the magnitude at the mean frequency we will see the envelope of the signal, the absolute value of the modulation. This is the term $\left|\cos(2\pi(f_b - f_a)t/2)\right|$, a rectified cosine wave with beats at a rate of $f_b - f_a = 10$ Hz. When we plot the phase at the mean frequency we will see a straight line (since the carrier is at this frequency) interrupted by half-rev steps as the modulation term changes sign with each beat. That is, alternate beats in the waveform are out of phase.

This is just what we see in Fig. 4.12. Midway through the signal, where the amplitudes at 3010 and 3020 Hz are equal (1.00/1.00), the magnitude at the mean frequency is a rectified sinewave with beats every 100 ms. Indeed, all the magnitude traces look like rectified sinewaves at this point because, as mentioned above, the analysis filters at 3010 and 3020 Hz are wide enough to pass the entire signal with little attenuation. The phase at 3015 Hz is a squarewave with 1/2-rev steps. The phase traces at 3010 and 3020 Hz are not squarewaves but sawtooths 1/2 rev peak-to-peak every 100 ms. These are merely the squarewave at 3015 Hz tilted by an additional ± 5 rev/s due to the differences in reference frequency.

The situation when the amplitudes of the two tones in the signal are not equal is shown at the beginning and end of the signal in Fig. 4.12. In this case, the nulls between beats do not go quite to zero magnitude since there is still a bit of the stronger tone left at the bottom of the null. The sawtooth phase of the stronger tone is a bit smoother now (reaching the limiting case of a straight line when the weaker tone vanishes completely). The phase of the weaker tone is rougher and shows whole-revolution lags or advances in phase between beats. Here is an interesting application of phase information—determining the relative magnitudes of signal components. While the magnitude traces are similar at all frequencies, the phase traces show which component is stronger.

Now we can look at the CB793 pulses in more detail. Figure 4.13 shows an enlarged spectrogram and magnitude-phase plots of the second pulse in Fig. 4.11. Magnitude and phase are plotted at the two transmitted frequencies, 3010 and 3020 Hz, and at the mean frequency, 3015 Hz, just as in Fig. 4.12. Figure 4.14 shows similar information for the third pulse in Fig. 4.11, a pulse with components 20 Hz apart. Examining these two figures, and comparing them to the synthetic signals in Fig. 4.12, we can make the following observations:

1. The received signal is periodic. At each null in the envelope of the transmitted pulse (every 100 ms in Fig. 4.13, every 50 ms in Fig. 4.14) the magnitude of the received signal goes to zero, or close to it, and the process is reset. However, the output signal takes a little time, say 200–300 ms at the beginning of each pulse, to reach steady state.
2. For the tones with 10 Hz separation, as the input magnitude increases after each null so does the output magnitude. The output magnitude continues to increase after the input signal has peaked, giving the output beats a sawtooth shape instead of the rectified sinewave shape of the input. The beats with 20 Hz separation are not appreciably sawtooth shaped, though they look flatter on top than rectified sinewaves. The 10 and 20 Hz pulses show ripples in magnitude

4. The phase at the mean frequency shows 1/2-rev changes from one beat to the next, as expected. If the amplitudes of the upper and lower tones were equal we would expect the phase at the mean frequency to be constant during each half-cycle interval. Since the lower component is actually a bit larger, we expect the phase at the mean frequency to slowly decrease during each interval, as at the beginning of Fig. 4.12. In fact, the phase at the mean frequency increases during each interval, a most important effect, but in line with the general phase advance seen with growing signals.
5. The rate of phase advance (instantaneous frequency offset) is not constant during each beat, but peaks near the middle of the interval. For the 10 Hz pulses, the peak rate of phase advance is about +8 Hz. The phase advance itself peaks about two-thirds of the way through each beat, reaching about 0.2–0.4 rev with 10 Hz separation, and 0.15–0.20 rev with 20 Hz separation. After the phase peak is reached, the phase rapidly retards as the next magnitude null approaches. In most cases the phase retards 1/2 rev, as expected, to meet the phase inversion in the following interval. However, in several instances in Fig. 4.13 the phase retards 1.5 revs. That is, the faller generated at the end of one beat inserts an extra whole revolution of phase lag before the start of the next beat.

We can think of the two-tone pulses above as a series of separate beats, each beat a short pulse that grows, advances in phase, and triggers a terminal faller much like the short pulses in Fig. 4.5. However, those pulses had constant input magnitudes (rectangular envelopes) whereas the two-tone beat envelopes here are rectified sinewaves. Still, the behavior is quite similar in both cases.

Helliwell et al. [1986a] describe a sideband generation model based on this similarity. Each beat in a two-tone signal is assumed to start an exponentially growing wave that reaches its maximum magnitude at the null before the next beat. The signal is then suppressed as the next beat begins to grow. The phase is assumed to advance in a roughly parabolic manner during each beat. This model describes many of the important features that are seen. It is not complete, however, for the following reasons:

1. The phase during each beat does not advance monotonically, but shows the parabolic advance followed by termination wrap-up and decrease typical of pulses with fallers, as seen in Section 4.2. But before we can incorporate such details into the sideband generation model we need to develop a good model of termination triggering.
2. The beats are not independent, since the process takes several beats to reach steady state. That is, the output depends not just on the current beat but on those up to 200–300 ms in the past.
3. The emission triggered at the end of one beat continues for some time, in some cases through the next beat and into the second one following. A complete model must account for the presence of at least two signals at once, the current beat and the previous emission.
4. Finally, we must account for interactions between the growing beats and the fallers. In Fig. 4.11 the lower sidebands are seen to be phase-coherent more than 100 Hz below the input tones. It is possible that this coherence is just due to the constancy of the fallers themselves—each faller moves at exactly the same rate just because the growth environment remains stable throughout the pulse. However, since the amplitudes and durations of the fallers change during the pulse, it seems more likely that they are coherent because they remain phase-locked to the input tones in some manner. (Note that coherence is not always present. Sidebands in the first pulse show some drift at 2970 Hz and below.)

Two-Tone LICO Signals with Sidebands. Now we will look at a more typical case of two-tone signals with sidebands. Figure 4.15 shows gray-scale phase plots of parts of a LICO (LIne COupling) transmission, 10 s segments with two signals separated by 30 Hz. This transmission is similar to the LICO1 case shown in Fig. 3.6, except here the two-tone segments are only 10 s long, alternating with single-tone segments. The LICO format also contains two-tone segments with 60, 120, and 240 Hz separations, but the 30 Hz segments shown here are the most interesting.

The two transmitted signals at 4090 and 4120 Hz are fairly stable, with no evidence of temporal growth or its associated phase advance. The interesting features here are the coherent sidebands at 30 Hz spacings above and below the transmitted tones. All intervals show the first two upper sidebands, most show the third, and the last interval at 1013:44 even has a weak fourth upper sideband at 4240 Hz, 120 Hz above the upper input tone. All intervals show the first lower sideband and some show the second lower one, but very weakly. There is also a weak signal at 4120 ± 2 Hz. This is a local power line harmonic and not a magnetospheric signal.

The relative amplitudes of the two input tones and the various sidebands change slowly with time. As a general rule, the upper sidebands are stronger than lower ones, and those closer to the transmitted tones are stronger than those farther away. In fact, the first upper sideband at 4150 Hz is sometimes stronger than one or both of the input tones. All components show phase changes at a rate of about ± 0.1 Hz, presumably due to duct drift. All the sidebands remain phase coherent with the input tones throughout each interval, showing nearly identical phase changes, though they are sometimes a bit noisier.

There may be small differential phase changes occurring, particularly with the higher-order sidebands. For instance, during the fifth interval at 1013:24, the two input tones and most of the sidebands show a phase advance of 1 rev. However, the third upper sideband at 4210 Hz advances 2 revs during this time. The second lower sideband at 4030 Hz seems to advance about 1.5 rev, though it is a bit noisy. As another example, in the last interval at 1013:44 all components advance roughly 1 rev except for the upper transmitted tone whose phase changes very little. We might expect to see differential phase changes since the relative amplitudes of the different components are changing, presumably because of changes in the sideband generation mechanism. However, the data are not good enough to say whether this is the case, or whether the differential phase changes are just due to multipath fading or some other cause unrelated to sideband generation.

Figure 4.16 shows magnitude-phase plots of one-second intervals from the center of each of the six LICO segments in Fig. 4.15. The output of only one analysis filter is plotted, centered at the mean frequency of the transmission, 4105 Hz. The filter bandwidth is 160 Hz, so it passes most of the sidebands in Fig. 4.15. For comparison, the response of the filter to a synthetic signal like the LICO signal as transmitted is shown at the bottom of the figure.

The various intervals in Fig. 4.16 seem to look quite different. They are certainly more complicated than the mean-frequency filter plots in Fig.s 4.13 and 4.14. However, we can pick out the following features:

1. Each signal is periodic, with a period of $1/15th$ second. Phase alternation between successive beats in the input signal (two beats per period) is still occasionally evident, as in the third plot starting at 1010:47.

2. The rate of magnitude ripples changes from one plot to another. In the first plot there are 60 major ripples per second, in the fifth plot about 30/s, and in the sixth plot there are 90/s. Turning back to Fig. 4.15 we see that the number of major ripples per second just reflects the frequency spacing of the dominant components in the spectrum, being 60 Hz in the first case

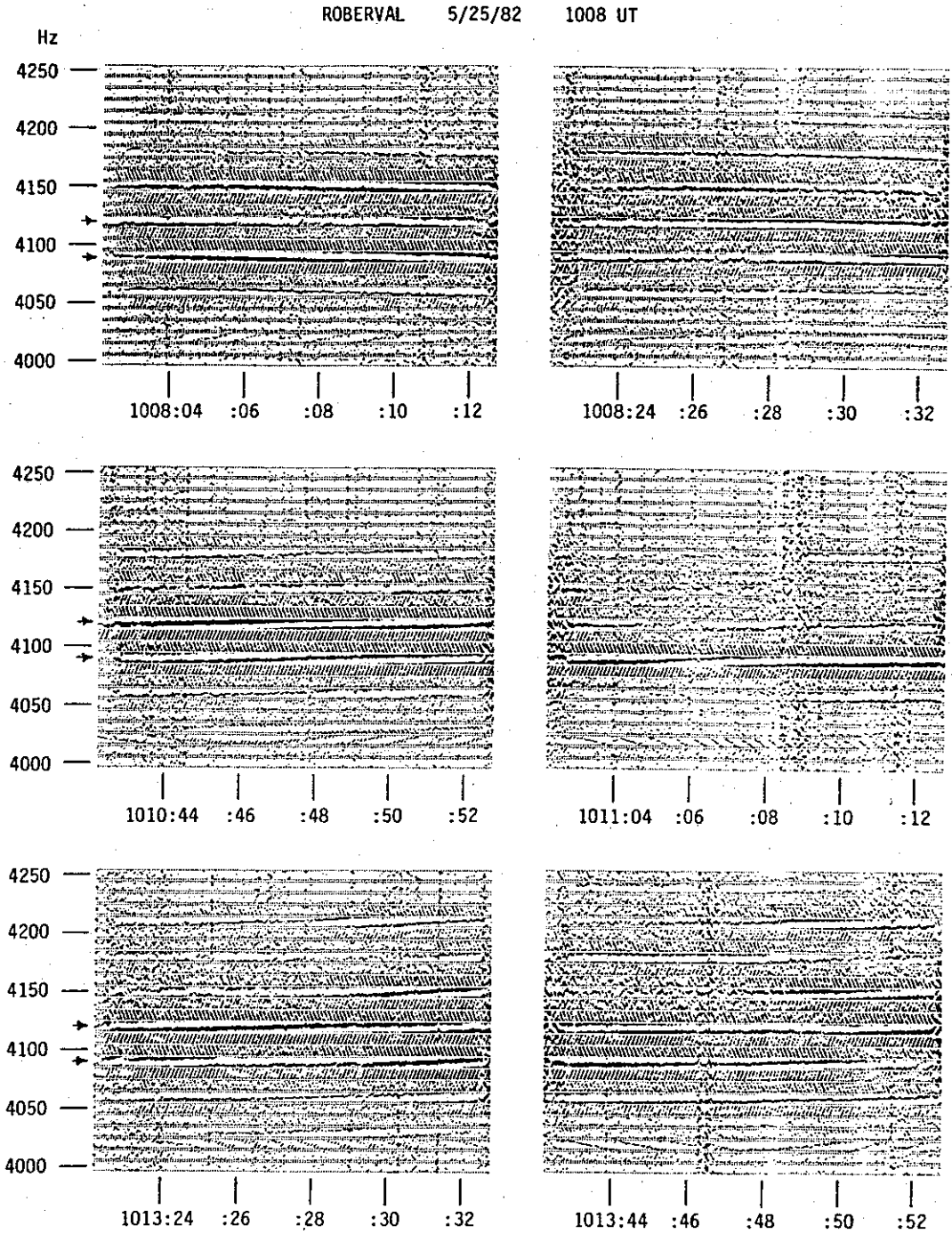

Figure 4.15. Gray-scale phase plots (BW = 20 Hz, P_{span} = 1 rev) of six ten-second segments of two-tone signals with 30 Hz separation from a LICO transmission. The two transmitted components at 4090 and 4120 Hz are indicated by arrows. Coherent sidebands can be seen, at various times, at 30-Hz intervals from 4030 Hz up to 4240 Hz. The input tones and generated sidebands show slow phase drifts due to duct motion, but all drift in phase together. A local power line signal near 4020 Hz is unrelated to the whistler-mode signals.

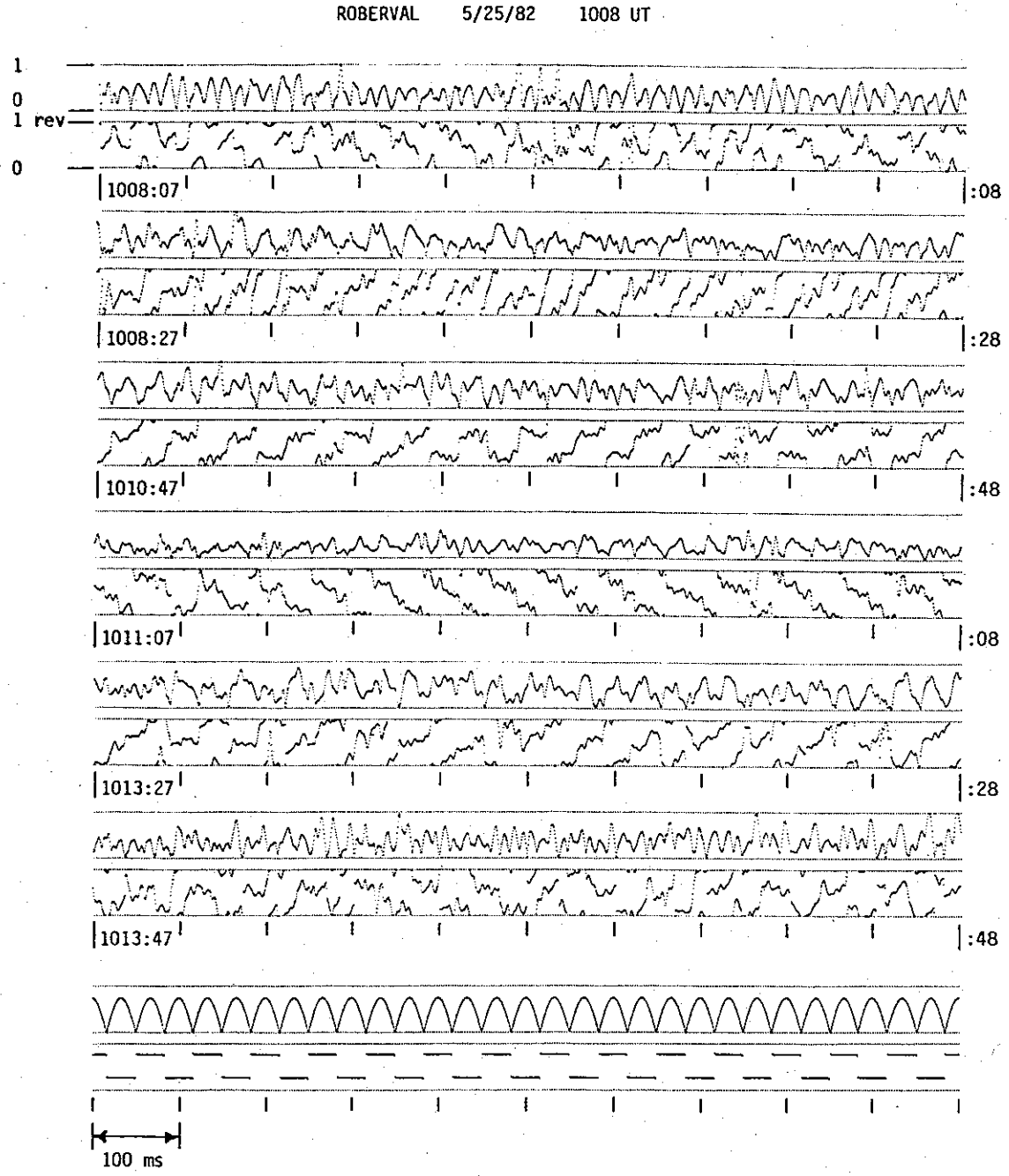

Figure 4.16. Magnitude-phase plots (BW = 160 Hz, linear magnitude scale) of one-second intervals from the six LICO segments in Fig. 4.15. Signals are analyzed at the mean frequency, 4105 Hz. The bottom plot shows a synthetic two-tone input signal for reference. The LICO signals show periodic behavior which changes slowly with time. The rate of magnitude ripples depends on the frequency separation of the major components of the signal at any given time, and the average slope of the relative phase depends on the mean frequency of those components.

(lower input tone and first upper sideband), 30 Hz in the fifth interval (two input tones), and 90 Hz in the last case (four major components from first lower sideband to first upper sideband).

3. The average rate of phase change is different in different plots. For example, in the second plot the phase often advances 2 revs every $1/15th$ second (+30/s), in the third plot there is little or not net advance, and in the fourth plot the phase lags 1 rev every period (−15/s). Again referring to Fig. 4.15, we see that the average rate of phase advance reflects the effective frequency, the "center of mass" of the various components in the spectrum, with respect to the analysis filter reference frequency. In the second plot this is about 30 Hz above the mean frequency (two input tones and two upper sidebands), at the mean frequency in the third plot (two input tones), and 15 Hz below the mean frequency in the fourth plot (lower input tone).

These features are just what one would expect to see, given the gray-scale plots in Fig. 4.15. Figure 4.16 doesn't really tell us much more than Fig. 4.15 did, except to confirm the phase-coherent nature of the process. However, remembering the signals in Fig.s 4.13 and 4.14, it is important to note what we don't see in this case. We don't see beats with sawtooth-shaped magnitudes characteristic of individually growing pulses. And we don't see the phase advance and wrap-up at each beat characteristic of growing pulses with terminal fallers. (There appear to be instances of phase wrap-up in some beats, but the signals are too complex in general to be sure of this.)

Summary. Two-tone signals with fallers like the CB793 pulses in Fig.s 4.11–4.14 are rare. Two-tone signals with sidebands usually look much more like the LICO signals in Fig.s 4.15 and 4.16. However, the CB793 pulses are invaluable in trying to understand the mechanism of sideband generation. They can be viewed as an intermediate form, a "missing link" combining the features both of growing pulses and of multi-tone signals with coherent sidebands. Because of their special form, we are able to see how cyclotron-resonant growth can explain, at least in a qualitative manner, the creation of sidebands. Each beat in a two-tone signal can be seen as an independent pulse, which grows and advances in phase until growth is damped out by the following beat, as proposed by *Helliwell et al.* [1986a]. Most two-tone signals with sidebands are a good deal more complex. Yet the same mechanism must surely be responsible even in more complicated cases.

4.4.2 Spontaneous Sidebands on a Single Tone

The formation of sidebands around a two-tone signal as discussed in the previous section is straightforward. Sidebands form at frequencies offset from the two input tones by multiples of the input tone spacing, and can be viewed as a kind of intermodulation distortion. However, sidebands are also seen to form around single-tone signals. The cause in this case is less clear.

Sidebands on a single tone were reviewed by *Park* [1981]. He summarized the characteristics of these sidebands as follows [*Park*, 1981, p. 2289]:

> "1. Although sidebands generally tend to appear when the carrier is strong, there is no simple relationship between the carrier amplitude and sideband frequency separation or sideband amplitude.
>
> 2. Sideband separations from the carrier range from $\sim$ 2 to 100 Hz, but at any given time, sideband separations tend to remain constant even as the sidebands switch on and off. No case has been observed where the sideband separation varies smoothly with the carrier amplitude.
>
> 3. Sideband amplitude may be symmetrical or asymmetrical about the carrier. In the asymmetrical case it is usually the upper sideband that is stronger.
>
> 4. Multiple sidebands are often observed, and their frequency separations from the carrier may or may not be harmonically related.
>
> 5. Sideband amplitude is usually 10 dB or more below the carrier amplitude, but sometimes it can exceed the carrier amplitude and can also trigger emissions."

From my own experience, this is an accurate summary of spontaneous sideband features. *Park* [1981], of course, was not able to use signal phase information in his study. In the rest of this section we will take a second look at some examples of spontaneous sidebands, and see how phase analysis can help measure their instantaneous frequencies. Based on these observations we will add two more characteristics to the list above.

Symmetrical Sidebands on Growing Pulses. Figure 4.17 shows two one-second pulses from a DIAG1 transmission, as seen above in Fig. 4.4. The first pulse, at 3810 Hz, grows for about half a second and then develops roughly symmetrical sidebands with energy both above and below the carrier. Toward the end of the pulse the lower sideband disappears and a second upper sideband appears at an intermediate frequency, between the carrier and the original upper sideband. At the end of the pulse there is little energy remaining at the carrier frequency and only a weak terminal emission is triggered.

The gray-scale phase plot of the first pulse shows that the sidebands initially appear about $\pm$50 Hz from the carrier, then drift out to $\pm$60 Hz. The upper 60 Hz sideband continues until the end of the pulse. Note that during its 400 ms existance this sideband shows a differential phase change with respect to the carrier of no more than 0.1 rev. That is, its offset frequency from the carrier is 60 ± 0.25 Hz. The second upper sideband, which persists for 250 ms near the end of the pulse, is seen to be exactly 30 Hz from the carrier—a subharmonic of the 60 Hz upper sideband.

The second pulse in Fig. 4.17, at 4050 Hz, develops symmetrical sidebands during the last 400 ms, but still has enough energy left to generate a BLI and trigger a terminal faller. The gray-scale plot shows the sidebands to be about $\pm$41 Hz from the carrier. Note that the sidebands are symmetrical about the frequency of the growing output signal at 4050.6 Hz, and not that of the input at 4050 Hz. There is also a hint of a sideband at -82 Hz for about 200 ms.

The magnitude-phase plots at the bottom of the figure were made with a filter bandwidth of 160 Hz, wider than the 80 Hz bandwidth in Fig. 4.4, in order to include all sideband energy. Both pulses show ripples in magnitude, at a 60 Hz rate in the first pulse, with an emphasis on alternate beats while the +30 Hz sideband is present; and at a roughly 40 Hz rate in the second pulse. The

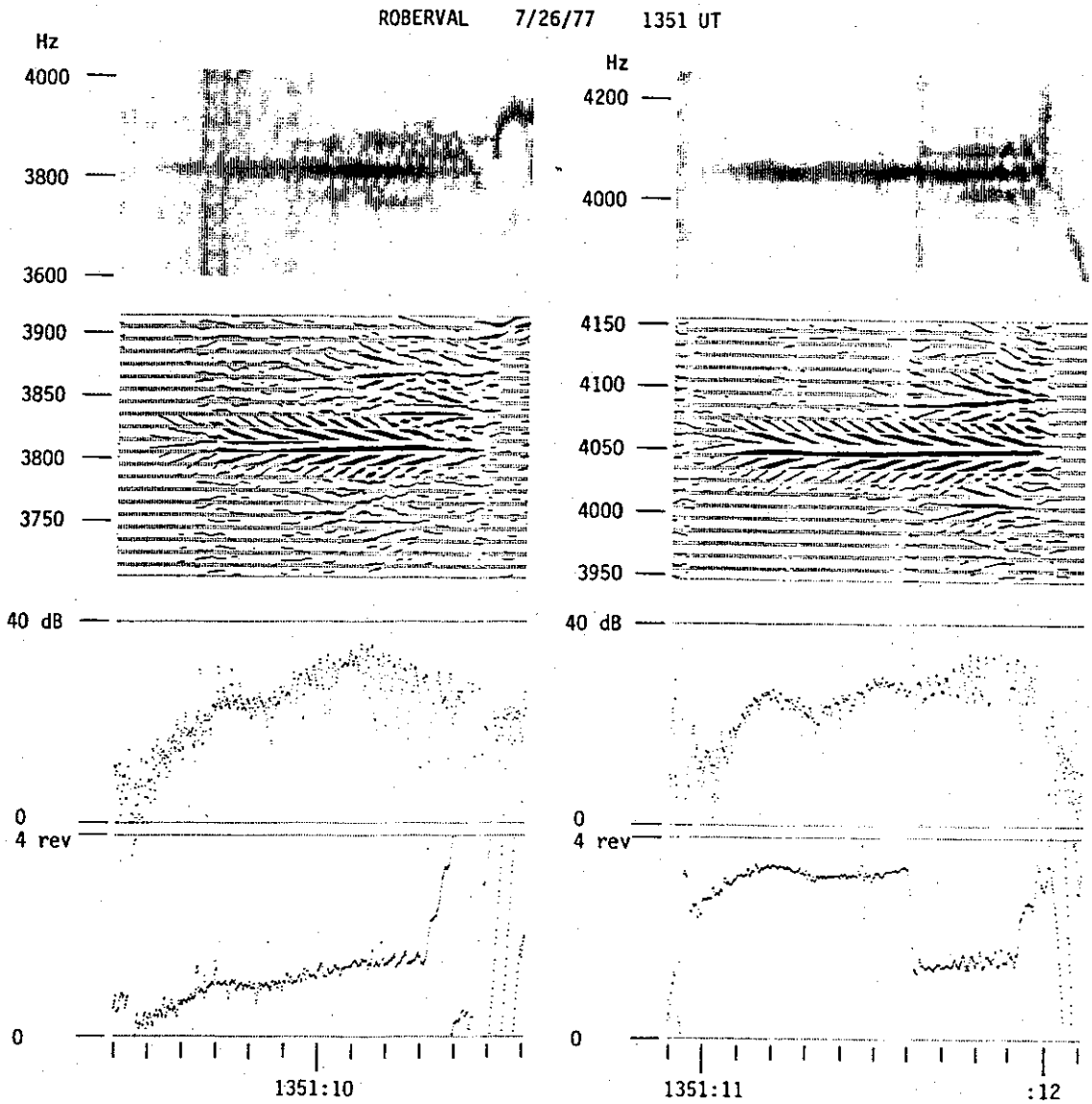

Figure 4.17. Two one-second single-frequency pulses with sidebands. Spectrograms (BW = 20 Hz) show symmetrical sidebands developing midway through each pulse. Gray-scale phase plots (BW = 20 Hz, P_{span} = 1 rev) show the sidebands to be at 30 and 60 Hz in the first (3810 Hz) pulse, and at $\approx$ 41 Hz in the second (4050 Hz) pulse. Magnitude-phase plots (BW = 160 Hz) show marked magnitude pulsations and significant, though less regular, phase pulsations. See also Fig. 4.4.

size of these ripples is significant, with dips as much as 20 dB below the peaks. There are also some phase ripples, about 0.4 rev peak-to-peak, but they are not as regular as the magnitude ripples. (The large phase jump in the middle of the second pulse is caused by the spheric, of course, and should be ignored.) If we saw magnitude ripples with constant relative phase, we would interpret the sidebands as resulting from amplitude modulation of the carrier tone. Since there are phase ripples, there must be a certain amount of phase modulation as well, though it may just reflect unequal amplitudes of the upper and lower sidebands.

All the other pulses in this transmission behave similarly to those in Fig. 4.17. Pulses at

3570 Hz show noisy sidebands, mostly about ±40 Hz, but with significant signal energy closer to the transmitted frequency. The 3810 Hz pulses have sidebands very close to ±60 Hz with some subharmonics. The 4050 Hz pulse sidebands are near ±40 Hz, and better separated from the transmitted signal than the other pulses. That is, there is less energy at frequencies between the transmitted signal and the sidebands. Magnitude plots show than many pulses develop pronounced ripples beginning halfway through the pulse. The ripples are initially quite symmetrical, occur at a 40–60 Hz rate, and show a peak-to-valley ratio of about 2:1. Often the ripples become more irregular toward the end of the pulse. There are several cases where every other ripple is emphasized, indicating signal energy at half the ripple frequency. Phase plots show ripples at the sideband frequencies, but they are less regular than the magnitude ripples. In several pulses there are cases of whole-revolution advances in phase as upper sidebands start to predominate. These advances may be the same phenomenon as the *N* events of *Dowden et al.* [1978].

More Growing Pulses. Figure 4.18 shows spontaneous sidebands developing on some ULF75 pulses. The first pulse is one of that series of 0.5 s pulses at 4500 Hz that was introduced at the beginning of the chapter. The second pulse is one of a few 1 s pulses at the end of this transmission. The data record analyzed here was played back at twice normal speed so the effective sampling rate was 12800 samples/s, half the usual rate. This allowed analysis filters to be used which had bandwidths as small as 10 Hz. As a result, the *f-t* spectrograms and gray-scale phase plots show finer frequency details than can be seen in Fig. 4.17.

The spectrogram of the first pulse in Fig. 4.18 is typical of most 4500 Hz pulses in the record at this time. Many of these pulses, especially from 1418 onward, develop more or less symmetrical sidebands. The sidebands arise after growth has saturated, around 300 ms into each pulse. The largest sidebands seem to be about 60 Hz above and below the main signal, or carrier. Sometimes further sidebands appear during the last 100 ms, situated between the 60 Hz sidebands and the carrier.

Gray-scale phase plots show that the *major* sidebands are indeed close to ±60 Hz, as shown by the first pulse in Fig. 4.18. While the phase of the carrier increases with time, the phases of the sidebands increase at very nearly the same rate. The difference between the phase advance of the carrier and the sidebands is often less than 0.2 rev over the 200 ms duration of the sidebands. That is, the sidebands are often offset from the carrier by 60 ± 1 Hz, even though the growing carrier may be several hertz above the transmitted frequency. The upper 60 Hz sideband is usually somewhat stronger. The first pulse in Fig. 4.18 also shows inner sidebands at about ±20 Hz, and a weak sideband about 80 Hz below the carrier. Other pulses in this record have sidebands near ±30 Hz. All these additional sidebands appear later than the 60 Hz sidebands, and their phase coherence with the carrier is not as great.

The beginning of the second pulse in Figure 4.18 is obscured by a rising emission from a preceeding one, but otherwise this pulse grows in a fashion almost identical to the first pulse in the figure. By the middle of this one-second pulse we see sidebands very similar to those of the first pulse. In fact, we suspect that if the first pulse had lasted longer than 0.5 s both might look much the same. After the midway point, the sideband structure becomes much more complicated, with energy at seven or eight discrete frequencies. The upper sidebands are stronger than those below. By the end of the pulse most of the energy has been transferred to the main upper sideband, and very little remains at the carrier. Only a weak terminal emission is triggered. The gray-scale plot shows the initial sidebands in this case to be at ±62 Hz. However, especially in the middle of the pulse, there is significant signal energy about ±25 Hz from the carrier. The remaining sidebands

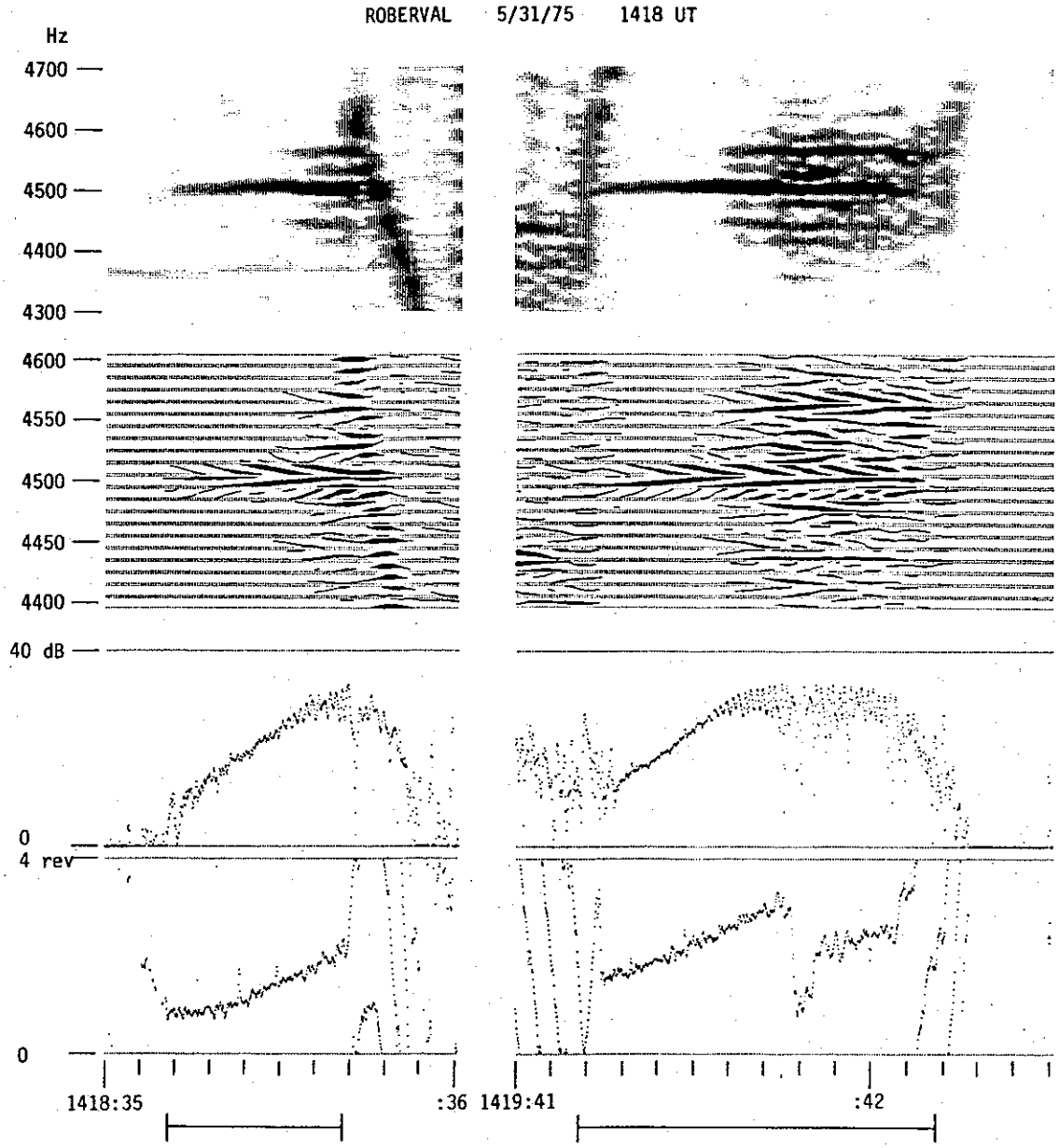

Figure 4.18. Single-frequency pulses at 4500 Hz with sidebands. Spectrograms (BW = 10 Hz) show complicated sideband structure developing after about 300 ms of growth. Gray-scale phase plots (BW = 10 Hz, P_{span} = 1 rev) show the strongest sidebands to be at $\approx \pm 62$ Hz, but also at $\approx \pm 25$ Hz. Magnitude-phase plots (BW = 160 Hz) show complicated magnitude and phase pulsations. Bars below the plots indicate the duration of the 0.5 s and 1.0 s transmitted signals. See also Fig.s 4.1, 4.2, and 4.8.

are less stable and their frequencies are hard to estimate. They may or may not be harmonically related.

The magnitude-phase plots show sideband ripples developing as the pulses reach saturation. Again, the magnitude ripples are large and fairly regular whereas the phase ripples are less significant and noisier. However, there are significant low-frequency phase ripples that reach about 0.7 rev p-p in the middle of the second pulse when the strong ± 25 Hz sidebands appear. There are also some

very deep magnitude nulls at this time, and some whole-revolution jumps in phase.

60 Hz Sidebands on a CW Signal. The preceding examples showed sidebands developing as growing pulses reached saturation. In the next example we will see sideband activity on a continuous single-frequency signal. Figure 4.19 shows 37 seconds of a POLIN (POwer Line INterception) continuous-wave transmission. The POLIN format consists of four minutes of single-tone signal, starting at 4420 Hz and stepping up in frequency by 10 Hz once per minute. Figure 4.19 shows the end of the second minute and the beginning of the third minute of transmission. A spectrogram of three minutes of this signal is given by *Park* [1981], and *Paschal and Helliwell* [1984] show a similar gray-scale phase plot, made with a slightly wider analysis filter.

During the first 108 s of the transmission the received signal is stable, without growth, and shows a slow decrease in phase due to duct drift which varies from −0.05 to −0.4 Hz. This is the condition at the beginning of Fig. 4.19. At 1332:48 the signal amplitude increases and the phase advances. After a second or two, when the phase has advanced about 1 rev, growth stops and sidebands appear 60 Hz above and below the carrier. They last about 1.5 s, disappear, and the carrier phase retards. This cycle of growth and phase advance, sidebands appearing and fading away, phase lag and amplitude decrease continues every few seconds for the rest of the transmission.

The received signal steps from 4430 Hz up to 4440 Hz at 1333:02.4 (after a one-hop delay of 2.4 s). Fig. 4.19 shows that the 60 Hz sidebands also increase at this time, to remain ±60 Hz from the carrier. The upper 60 Hz sideband is stronger than the lower one. There is also sideband energy closer to the carrier, from 20 to 40 Hz away, also stronger above than below. These closer sidebands seem very noisy and show no phase-coherent structure.

Because of the irregular variation in the phase of the carrier in Fig. 4.19, it is difficult to estimate the exact frequency of the coherent sidebands. To overcome this, a second gray-scale plot was made tracking the carrier as a pilot tone in order to measure the phase of the sidebands with respect to it. The phase of the sidebands relative to the carrier sometimes seems to be constant over periods of several seconds, so the sidebands are at 60 ± 0.1 Hz, at least in some cases. However, the amplitude of the carrier is not always very high, and there are some tracking errors. Because of this, it is difficult to compare the phases of sideband segments separated by more than a few seconds. Also, in a few spots there is evidence of a drift in the relative phase of a sideband by as much as 0.25 Hz. So we can say that the coherent sidebands in Fig. 4.19 are offset from the carrier by a frequency close to 60 Hz, certainly within 0.25 Hz of it. Whether the offset is exactly 60 Hz, possibly with some small phase variations as sidebands come and go, or whether it may have some other value, or even change with time, is uncertain.

Figure 4.19 also shows signals induced in the receiving loop from local power lines. There is a weak signal near 4260 Hz, which is the $71st$ harmonic of the 60 Hz power grid. The frequency of this signal shows changes as large as 1 Hz. This is a relative change of 0.02%, typical of the stability of the power system. There is another weak signal (barely visible) at 4140 Hz, the $69th$ harmonic, which is coherent with the previous one. Finally, there is a stronger signal near 4120 Hz. This signal, which varies in frequency by over 5 Hz, is not a 60 Hz harmonic but is probably caused by an electric motor somewhere in the local power system. This type of interference is common at Roberval, but it can be identified by its wide frequency variation, intermittent nature, and occasionally by frequency ramps when the motor is first turned on. The point to notice here is that neither the Siple carrier nor its 60 Hz sidebands have any obvious connection with either the local power line harmonics or the motor noise.

Noisy Sidebands with Coherent Carrier. Not all single-frequency signals with sidebands show

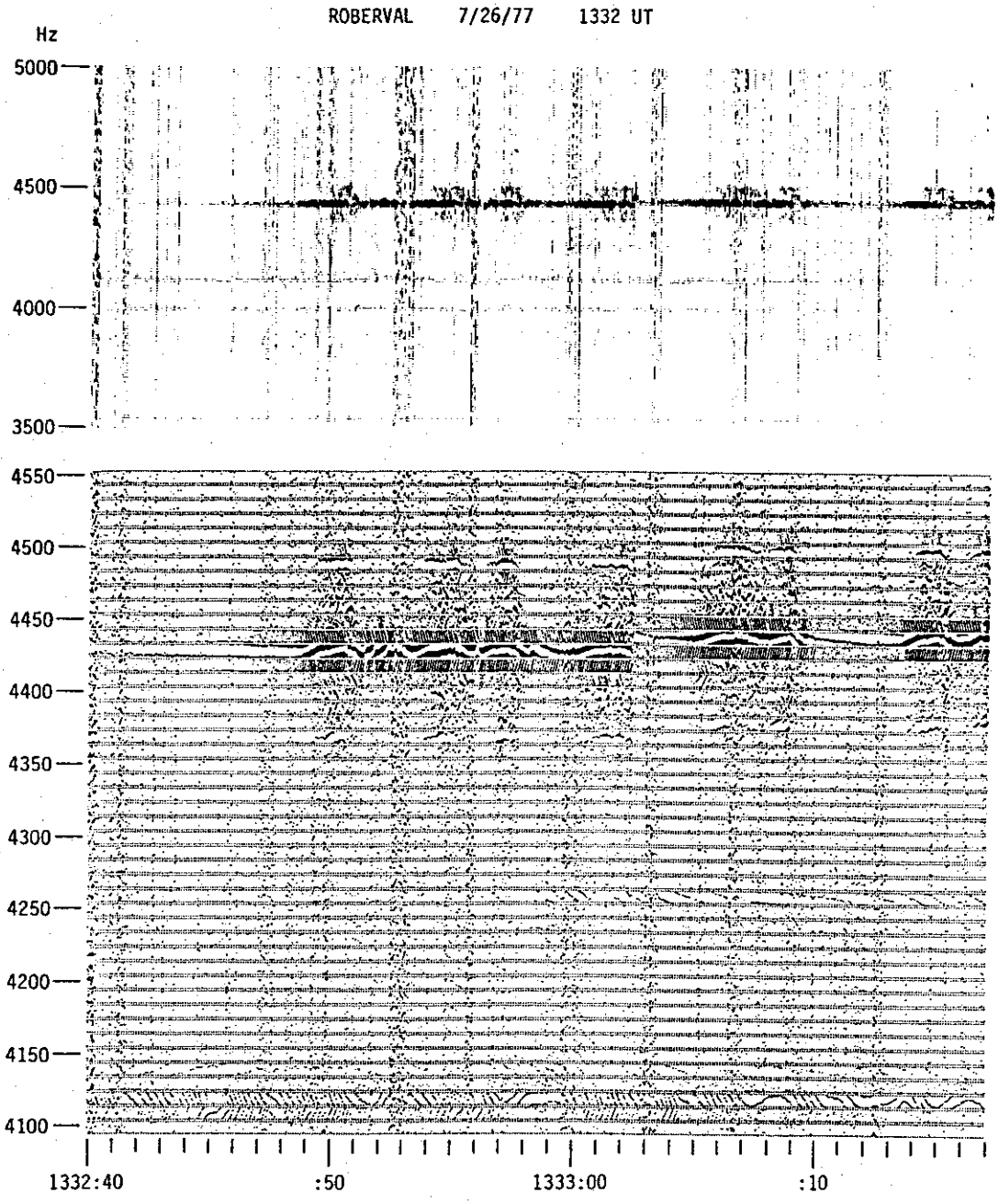

Figure 4.19. Spectrogram and gray-scale phase plot (BW = 16 Hz, P_{span} = 1 rev) of a CW signal starting at 4430 Hz. Note the growth and phase advance starting at 1332:48, and the subsequent bursts of coherent sidebands ±60 Hz from the growing input signal. At 1333:02 the input frequency in increased by 10 Hz, and the sidebands also increase by that amount. After *Park* [1981] and *Paschal and Helliwell* [1984].

coherent components at 60 Hz offsets, of course, or at any other well-defined frequency. In Figure 4.20 we see an example of coherent single-tone carrier signals with noisy, incoherent sidebands. This is a small segment of a CBST (Coherence Bandwidth STair) transmission [*Helliwell*, 1983a]. The CBST format contains two tones, both initially at the same frequency. One tone moves up in 10 Hz steps once per second until the frequency separation is 270 Hz, then it steps back down again. The second

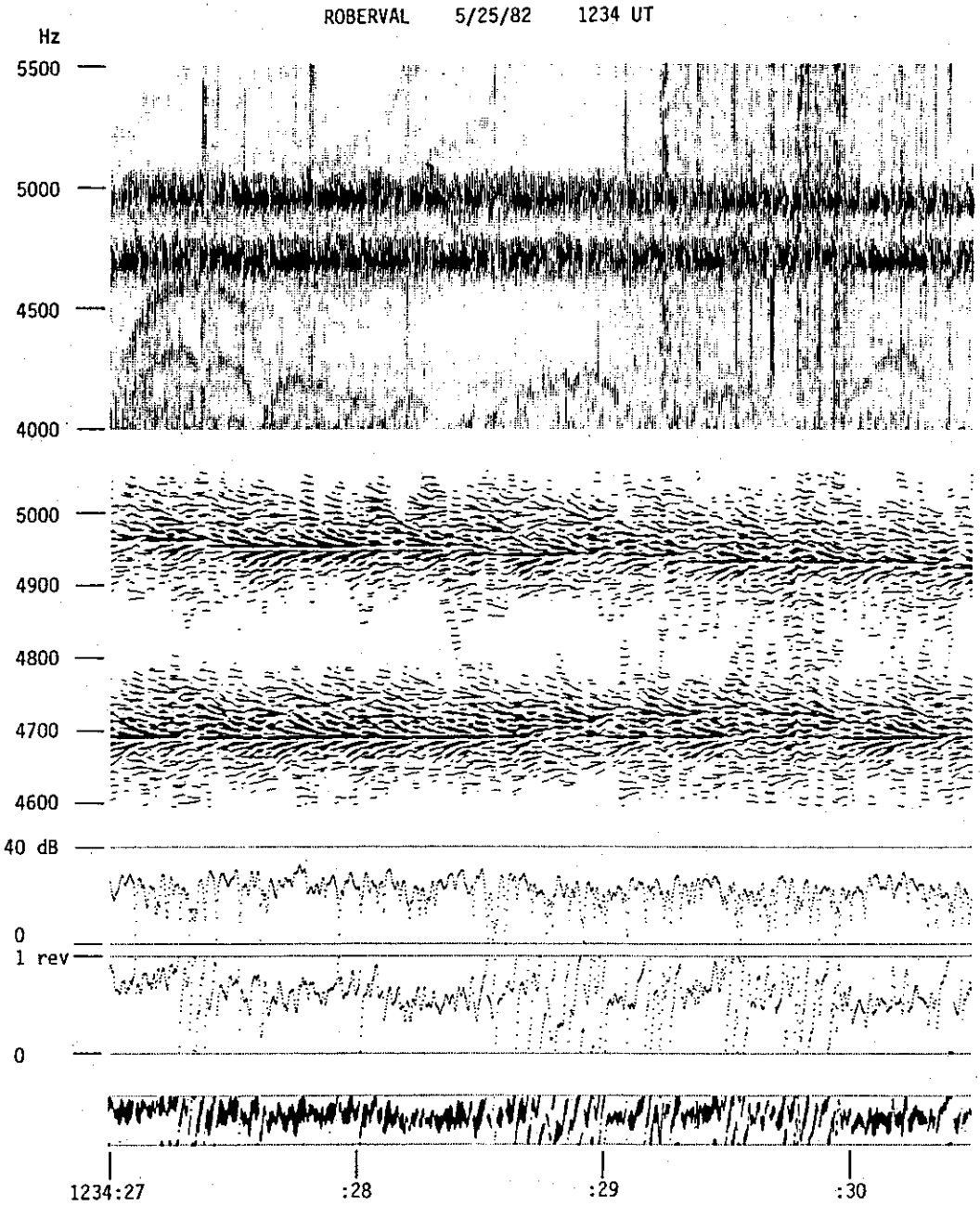

Figure 4.20. Portion of a two-tone CBST transmission showing noisy amplification typical of more active conditions. The upper tone steps down by 10 Hz every second, the lower one is constant in frequency at 4690 Hz. Discrete chorus elements rise from below to over 4500 Hz. The center gray-scale phase plot (f_I = 10 Hz, BW = 20 Hz, P_{span} = 1 rev) shows coherent signals with mostly upper sideband noise. The magnitude-phase plot and lower gray-scale phase plot (both BW = 80 Hz) show the noise at 4690 Hz.

tone remains constant in frequency. The CBST format was designed to investigate the interaction of two single-frequency signals as their separation changes.

In Figure 4.20 we see the signal as the upper tone is starting to step down from its maximum frequency. At the beginning of the plot, the upper tone is at 4960 Hz, and by the end of the plot has

just reached 4920 Hz. The lower tone remains at 4690 Hz throughout. The spectrogram shows the upper edge of a band of chorus, a sign of growth activity, that extends from 2200–4000 Hz with a few discrete elements reaching up to the Siple signal. However, there are few risers triggered by the CBST signal, normally another sign of growth activity. Instead, most of the signal growth seems to be channeled into the sidebands.

The gray-scale phase plot in the center of Fig. 4.20 was made with a filter spacing of $f_I = 10$ Hz, but without index dots separating the filter traces so the plot would be a bit more compact. We see coherent signals at both input frequencies (viewing the figure edge-on helps), and can pick out the frequency steps made by the upper tone. However, the sidebands just look like noise. There is more sideband power above each input signal than below, but there are very few discrete sideband elements. This is a common type of behavior. Much of the time sidebands on single-frequency signals appear noisy and incoherent. Noisy sidebands become more common as growth activity increases. When conditions are moderately active, a coherent carrier signal may still be seen among noisy sidebands, as in Fig. 4.20. When conditions are very active, even the carrier may disappear into noise. This increase in noisiness could be caused by increasing instability in the sideband generation process, but it might also be due to increased multipath interference.

The magnitude-phase and gray-scale phase plots at the bottom of Fig. 4.20 show the behavior of the lower tone at 4690 Hz. The magnitude plot shows variations over times scales from 10 ms to 1 s. The phase plot shows fluctuations around 0.4 rev peak-to-peak interupted by numerous whole-revolution phase advances. The gray-scale plot at the bottom shows that some of these phase jumps resemble the N events of *Dowden et al.* [1978].

The sideband noise we see in Fig. 4.20 occurs independently around each input tone. No sidebands at harmonics or subharmonics of the frequency separation are seen. The CBST format demonstrates that there is little if any interaction between two signals when they are separated by more than, say, 100 Hz. However, harmonic sidebands do occur in CBST signals with smaller separations, just like the two-tone signals in Sec. 4.4.1 above. In this particular record, harmonic two-tone sidebands occur with separations from 20 to 50 Hz, and occasionally at larger values. One interesting feature is that both the carrier amplitudes and the level of incoherent sidebands are reduced when the input tones are close together (but more than 10–15 Hz apart), especially at separations of 20 and 30 Hz [*Helliwell*, 1983a].

Discussion. *Park* [1981] reviews various theories which have been proposed to account for spontaneous sidebands. All are based on postulated instabilities inherent in the growth process. The examples of sidebands above offer little information to help us choose between these various theories, except possibly (as *Park* [1981] also observes) to rule out some which predict that the frequency offset of sidebands will depend on the carrier amplitude.

However, we have seen an effect which is not predicted by any theory—the occurrence of sidebands at offset frequencies very close to 60.0 Hz. Measuring signal frequencies from spectrograms, *Park* [1981, Fig. 3] shows that many sidebands in the POLIN case above (Fig. 4.19) have offsets clustered in the range 57–63 Hz. (Others are in the range 25–40 Hz.) Using phase analysis we found those sidebands to actually be within ± 0.25 Hz, and probably ± 0.1 Hz, of 60.0 Hz. We also saw sidebands near 60 Hz on growing pulses in Fig.s 4.17 and 4.18. There is no known natural process which might account for this particular value, and the obvious inference is that these sidebands are related in some way to the 60 Hz frequency of the power system.

There are several conceivable causes of 60 Hz sidebands. One is that they are due to background signals in the magnetosphere, "power line radiation" or PLR at harmonics of the 60 Hz power

frequency, which are somehow amplified in the proximity of a Siple signal. However, the sidebands in Fig. 4.19 cannot be PLR signals *per se,* because when the transmitter frequency steps up by 10 Hz, the sidebands also increase to remain 60 Hz away from it.

As a second possibility, *Paschal and Helliwell* [1984] suggested that "beats between amplified power line harmonics could temporarily trap longitudinally resonant electrons as in the *Park and Helliwell* [1977] explanation of whistler precursors, causing a periodic density bunching of these electrons. This bunching might then cause cyclotron-resonant waves to be modulated at the 60 Hz beat frequency." This theory has the advantage that it is only the difference in frequency between power line harmonics that is important, and not the harmonic frequencies themselves. We will discuss the evidence for and against PLR in Section 4.6.

A third possibility is that the 60 Hz sidebands were transmitted, at a low level, along with the intended single-frequency signal. 60 Hz is, of course, the nominal frequency of the Siple Station generators as well as the North American power grid. It is certain there will be 60 Hz sidebands at some level on the Siple signal, and the question really is how far below the carrier they are. The transmitter, "Zeus," that was used in the POLIN transmission in Fig. 4.19 created its output waveform as a series of half-sinewave pulses generated by switching currents through tuned circuits with silicon controlled rectifiers (SCR's) [*Helliwell and Katsufrakis,* 1978]. The analog output of a frequency synthesizer was converted into control pulses for the SCR's by the transmitter exciter. Any 60 Hz hum at the input to the exciter would have been added to the synthesizer signal and caused low-level 60 Hz phase modulation. It is likely that transmitted 60 Hz sidebands on Zeus signals were at least 40 dB below the level of the carrier, and probably even lower. Unfortunately, this transmitter is no longer in existence, and its sideband level cannot be measured.

The current transmitter, "Jupiter," is a surplus Omega Navigation System unit and is just a high-power class B audio amplifier. Its signal-generation equipment is more complex than Zeus's, and there are several places where low-level 60 Hz modulation might occur, as well as modulation at 120 Hz from power-supply ripple. Attempts to measure the sideband level at its output have been inconclusive, though some measurements suggest they may be as high as −30 dBc. Both transmitters are also expected to have significant sidebands at 360 Hz, the ripple frequency of their 3-phase power supplies.

If low-level transmitter modulation is the cause of sidebands, we must explain why they are so strong in the received signal. Perhaps the sidebands grow independently of the carrier—they appear after carrier growth has saturated because it takes them longer to reach a noticeable level. Or, perhaps low-level sidebands are strong enough to synchronize a spontaneous sideband oscillation that would otherwise have appeared at a nearby frequency. Several transmissions have been made with the Siple generators running at 58 Hz. If low-level transmitter modulation is important, we would expect to see received sidebands near 58 Hz instead of 60 Hz. The results of this experiment are not yet in.

Transmitter modulation might explain another effect—the variability in frequency of some 60 Hz sidebands. The frequency of the Siple diesel generators may fluctuate by as much 1.5 Hz depending on their load, which can change in large steps during pulse transmissions. On the other hand, sidebands might be locked to or initiated by a signal at exactly 60 Hz and still show phase (and temporary frequency) fluctuations because of changes during the growth process. A study of relative sideband phase *vs.* amplitude might clarify this.

Summary. From our observations of the phase behavior of spontaneous sidebands we can add the following two features to *Park's* [1981] list of sideband characteristics given above:

6. When sidebands appear both above and below the carrier, they are often offset by the same frequency and are phase-coherent. These symmetrical sidebands are not due entirely to amplitude modulation of the carrier, but show some phase modulation as well (some of which may be due to dispersion). Sidebands at multiple frequencies are sometimes harmonically related and phase-coherent.

7. Sidebands often appear offset from the carrier by frequencies very close to 60 Hz. Sidebands at subharmonics of this frequency are also seen.

There are several conceivable causes of sidebands with 60 Hz offsets and their subharmonics. One which cannot be ruled out yet is that these received sidebands are due to low-level 60 Hz modulation at the transmitter.

4.5 Wave-Wave Interactions

In this section we will look at some of the effects of one whistler-mode wave on another. We have already seen some examples of one effect—the generation of sidebands by two-tone signals in Section 4.4.1. Now we will see examples of others: suppression, where one signal inhibits the growth of another; entrainment, where one signal synchronizes the growth of another; and precursors, where a whistler enhances the growth of a second signal.

4.5.1 Growth Suppression by Nearby Signals

Various forms of suppression of one whistler-mode signal by another have been observed for some time. *Helliwell and Katsufrakis* [1974, Fig. 2] show an example where a whistler suppresses the growth of a pulse from the Siple Station transmitter. *Raghuram et al.* [1977a] discuss the "quiet band" effect, where a Siple signal suppresses the amplitude of natural mid-latitude hiss in a band up to 200 Hz wide just below the transmitted frequency. They propose that cyclotron resonance with the Siple signal alters the velocity distribution of interacting electrons, and this prevents the hiss. *Raghuram et al.* [1977b] report the suppression of growth of Siple pulses by two-hop echoes of previous pulses. In this case they think the effect is due to echoing emission components, above the transmitted frequency, which reduce the coherence of the input signal and thus reduce its growth.

Mutual growth suppression between components in a multi-tone signal has also been observed. *Chang* [1978] and *Chang et al.* [1980] report experiments using frequency-shift keying of the Zeus single-tone transmitter at Siple to generate multiple components. They find that transmitted components spaced 50 Hz or less show mutual suppression and energy coupling. This 50 Hz separation is the *coherence bandwidth*, the range in frequency over which waves may resonate with a given energetic electron in the interaction region. Two wave components separated by less than this frequency will resonate with some of the same electrons, and can thus affect each other. Two components further apart will interact independetly with separate populations of electrons. (The authors use a homogeneous interaction model with phase-trapped particles, and predict that the coherence bandwidth will be proportional to the square-root of the wave intensity. This is probably not an appropriate model for low-level signals.)

Helliwell [1983a] describes the results of the CBST two-tone transmission where the separation between two equal-amplitude signals is varied in steps from 0 to 270 Hz (see Fig. 4.20). He finds two types of suppression. One is the mutual suppression of growth and triggering of two components when their separation is 20 or 30 Hz. The second is an asymmetrical suppression of a lower frequency component by an upper component up to 100 Hz away. *Helliwell* [1983a, 1983b] also compares the effects of two-tone signals with signals whose components are due to 100% amplitude modulation or to frequency modulation with unity modulation index (frequency deviation = modulation frequency). Signals with 0 and 5 Hz modulation show approximately the same level of growth. Maximum

suppression occurs with 20 Hz modulation, and is about 25 dB with two-tone signals, slightly less for AM and FM. Components spaced 100 Hz or more show growth comparable to that of single tones.

Coherence Bandwidth from AM Sideband Phase Behavior. The studies above by *Chang* [1978], *Chang et al.* [1980] and *Helliwell* [1983a, 1983b] determined the coherence bandwidth either from the amplitude effects of growth suppression, or from changes in the rate of pre-termination emission triggering. However, growth suppression also affects the phase behavior of signal components, and we can estimate the coherence bandwidth of wave-particle interactions from this. In the following example we will study signals with amplitude modulation.

We can write the waveform of an amplitude-modulated signal as

$$s(t) = A[1 + mx(t)]\cos(2\pi f_c t), \tag{4.2}$$

where $A\cos(2\pi f_c t)$ is the unmodulated carrier at frequency f_c, $m \leq 1$ is the modulation index, and $x(t)$ is the modulating waveform. We require that $|x(t)| \leq 1$. The case $m = 1$ is known as 100% modulation. If the modulating signal is a tone at frequency f_m, as $x(t) = \cos(2\pi f_m t)$, and the modulation is 100%, we can write

$$\begin{aligned} s(t) &= A[1 + \cos(2\pi f_m t)]\cos(2\pi f_c t) \\ &= A\cos(2\pi f_c t) + \frac{A}{2}\cos[2\pi(f_c - f_m)t] + \frac{A}{2}\cos[2\pi(f_c + f_m)t]. \end{aligned} \tag{4.3}$$

Comparing this to the two-tone signal in Eq. (4.1), we notice the following differences. The two-tone signal had two equal-amplitude components. The 100% tone-modulated AM signal has three components, one at the carrier frequency and two sidebands a distance f_m on either side. Each sideband has half the amplitude of the carrier. The envelope in the two-tone case was a rectified cosine wave with alternate beats out of phase. The envelope of the AM signal is $[1 + \cos(2\pi f_m t)]$, a raised cosine wave. Each such "beat" reaches zero amplitude smoothly at the bottom of the cosine, and there is no phase reversal from one beat to the next. However, we can still think of an AM signal as a series of pulses, as in the case of the two-tone signals in Sec. 4.4.1, and we expect that magnetospheric growth will have similar effects, such as creating additional sidebands.

Figure 4.21 shows a segment of a CB792 (Coherence Bandwidth, 1979, version 2) transmission, a set of one-second pulses at 3270 Hz with 100% amplitude modulation at frequencies of 0, 5, 10, 20, 50, and 100 Hz. The pulse with 0 Hz modulation (a CW pulse) shows growth and the triggering of multiple pre-termination emissions. The gray-scale plot shows phase advance at the beginning typical of a rapidly-growing pulse. Note the multipath propagation here. The spectrogram shows a weak signal preceeding the main pulse by about 200 ms, and perhaps another weak one on a path following the main pulse by 200 ms. The main pulse itself triggers emissions on at least three closely-spaced paths. The one-hop group delay for the main pulse is 2.6 s. There are also two-hop echoes. The weaker signal with rising emissions between the 10 and 20 Hz pulses is a two-hop echo of the first, 0 Hz pulse, delayed by 5.2 s.

The pulse with 5 Hz modulation also shows growth, but not as much as the first pulse. Emissions are triggered by the last four beats in the signal, though they don't last as long as those triggered by the first pulse. A strong impulse, a BLI, occurs at the end of the pulse. The 200 ms beat spacing must be large compared to the loop delay T of *Helliwell and Inan* [1982] since the beats behave almost like independent pulses.

The 10 and 20 Hz pulses are weaker, and show almost no growth or triggering (the emissions are echoes). The 20 Hz pulse is the weakest. However, they do generate additional sidebands, as

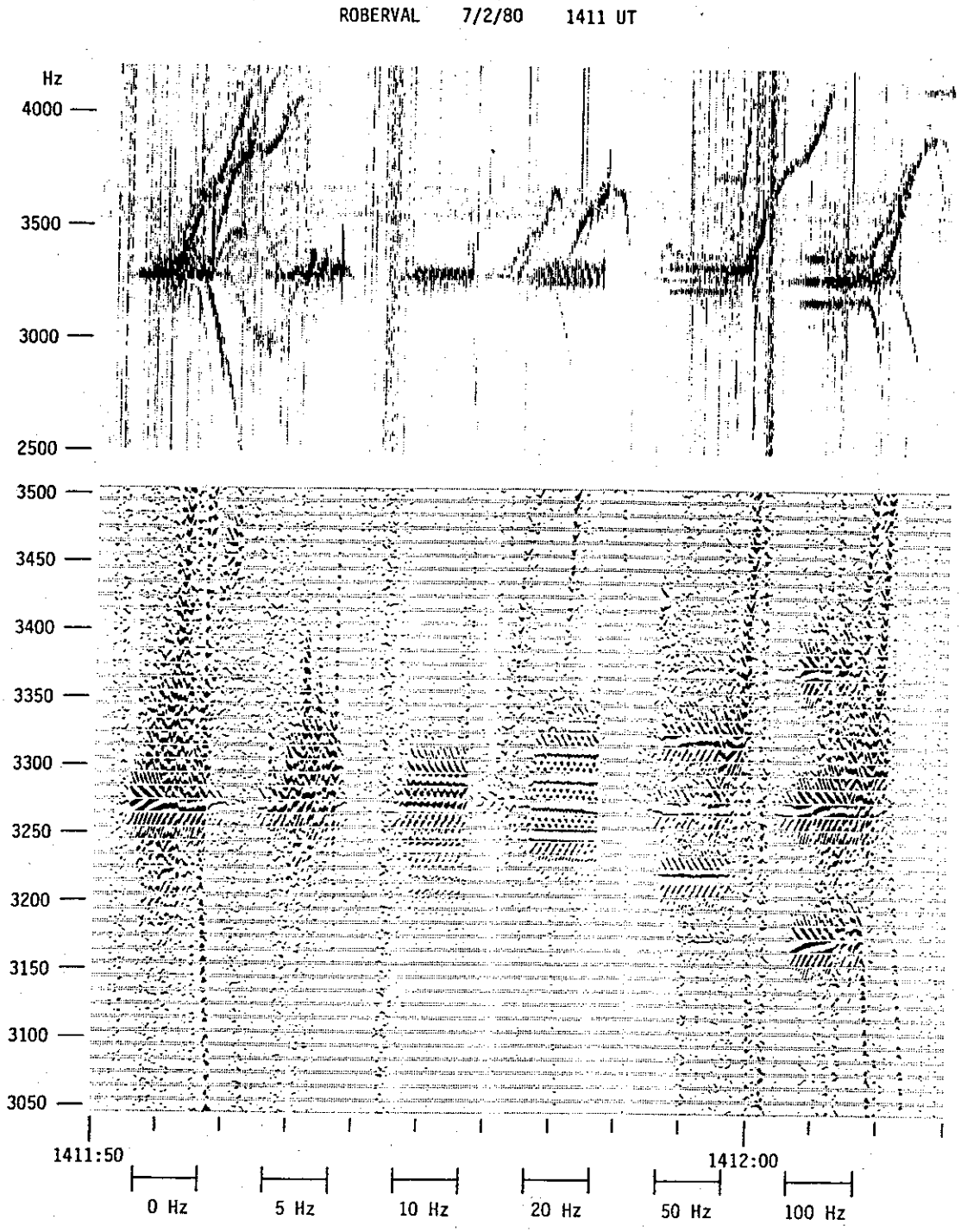

Figure 4.21. Amplitude-modulated one-second pulses whose behavior depends on sideband spacing. Bars below the plot show modulation frequency and duration of transmitted pulses. 0 and 5 Hz pulses show growth. 10 and 20 Hz pulses develop sidebands instead. 50 Hz pulse has sidebands and some growth. 100 Hz components grow independently. (Gray-scale phase plot BW = 20 Hz, $P_{span} = 1$ rev.)

we might have expected. The 10 Hz pulse shows coherent sidebands every 10 Hz from at least 3190 to 3310 Hz; that is, at least seven lower sidebands and three upper ones in addition to the three

transmitted components. The 20 Hz pulse has one additional lower sideband and two upper ones. In both pulses, sidebands further away take slightly longer (up to 100 ms) to appear.

The 50 Hz pulse still shows fairly stable phase behavior, though its phase is noisier than the preceeding two pulses. It generates additional sidebands ±100 Hz from the carrier, at 3170 and 3370 Hz, which take about 300 ms to develop. The upper transmitted component at 3320 Hz shows a modest amount of growth, with a slight phase advance. There are noisy signals between the various phase-coherent components. Noise during the first half of the pulse is an echo of the earlier 5 Hz pulse. Noise toward the end appears as the carrier and upper component trigger several emissions. This pulse shows a combination of features—growth suppression and sidebands, but also termination triggering.

The final pulse, with 100 Hz modulation, shows no coherent behavior. No additional sidebands are generated, and each of the three transmitted components shows growth and triggers emissions. In the gray-scale plot, each component shows an advance in phase associated with its growth, but the advance is different in each case. That is, not only do the three transmitted components grow, but phase information shows they grow independently of each other.

The independent phase behavior of the components in the 100 Hz pulse is an important observation. We might imagine, as their frequency separation increases, that different signal components will interact with increasingly separate populations of energetic particles until a point is reached where growth is no longer suppressed. Yet at this point there might still be enough overlap in populations that growing components remain in phase. This may be what happens in the 50 Hz pulse, where some growth is observed yet the phase behavior is largely coherent. At 100 Hz separation, even component phases behave independently, and we can be sure that the particle populations interacting with each component are almost entirely disjoint. Judging from the phase behavior, as well as the amplitude and emission behavior, the coherence bandwidth in this record is greater than 20 Hz but less than 100 Hz; it is probably about 50 Hz.

Effects of Relative Component Amplitude on Suppression. Figures 4.22–4.25 show four segments of a CBVA (Coherence Bandwidth, Variable Amplitude) transmission from Siple. This was a transmission with two tones separated by 20 Hz to provide maximum growth suppression. The objective of the CBVA format was to determine the effect on growth and triggering of varying the amplitude of one of the tones relative to the other. The CBVA format consists of a series of two-second two-tone pulses. Sixteen different pulses result from combinations of the following: 1. either the amplitude of the upper tone or the lower tone is varied, 2. the variation occurs during the first or the second second of the pulse, and 3. the variation is either an infinite attenuation (the remaining tone by itself) or is an amplitude ramp from −20 dB at the respective end of the pulse to full-power at the middle. The diagrams at the bottom of Fig.s 4.22–4.25 show the various combinations. (The CBVA format also includes two-tone pulses where one tone is a frequency ramp to vary the separation; but these are of no interest here since the phase of frequency-varying signals is difficult to interpret.) The pulses in Fig.s 4.23 and 4.25 appear in *Helliwell et al.* [1986a, Fig.s 2b, 6–8], which also shows other CBVA pulses.

We will consider four catagories of pulses in turn:

1. *Pulses with a single tone preceding a two-tone interval.* These are the first two-second pulses in Fig.s 4.22 and 4.23. During the single-frequency interval at the beginning, each pulse shows rapid growth and phase advance, and one or two triggering events. Whether the single-frequency transmission is at 3770 (Fig. 4.22) or 3750 Hz (Fig. 4.23) makes no difference, of course. The initial growth rate is 80–90 dB/s. Phase advance is limited to about 2 revs before triggering occurs.

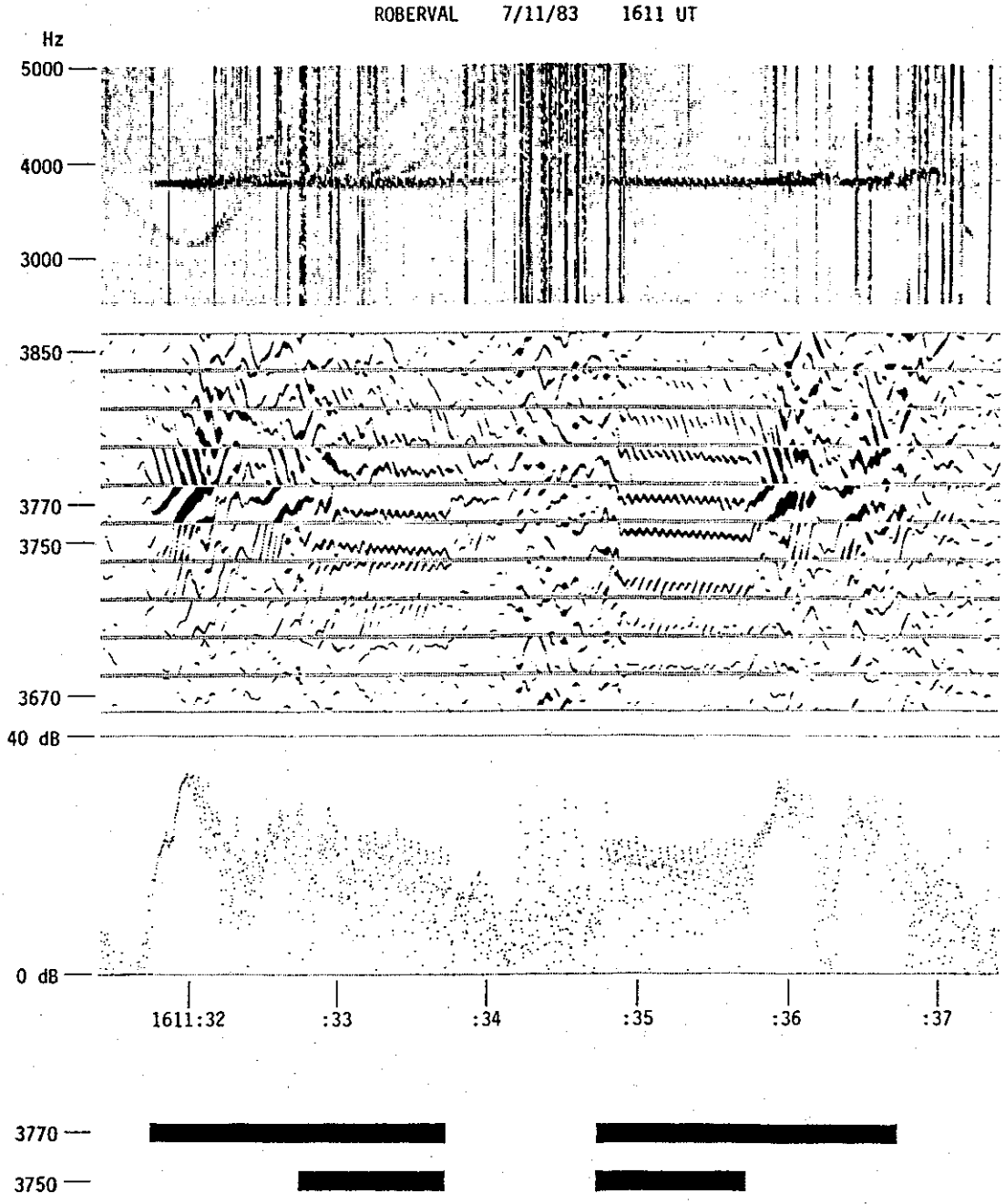

Figure 4.22. Spectrogram, gray-scale phase plot (BW = 20 Hz, P_{span} = 1 rev), and magnitude plot (BW = 80 Hz) of a pair of pulses from a CBVA transmission. Diagram at bottom shows amplitudes of the transmitted components at 3750 and 3770 Hz. The lower tone at 3750 Hz is off during the first half of the first pulse, and the second half of the second pulse.

Because of multipath, the triggering is a bit complicated. The main path seems to generate some impulsive noise and stop, but some weaker paths trigger risers. About 700 ms into each pulse a second interval of growth, phase advance, and triggering begins, a bit noisier than the first. This initial behavior is what we expect of typical single-tone signals.

Soon (100 ms) after two-tone transmission begins in the middle of the pulse, the system settles

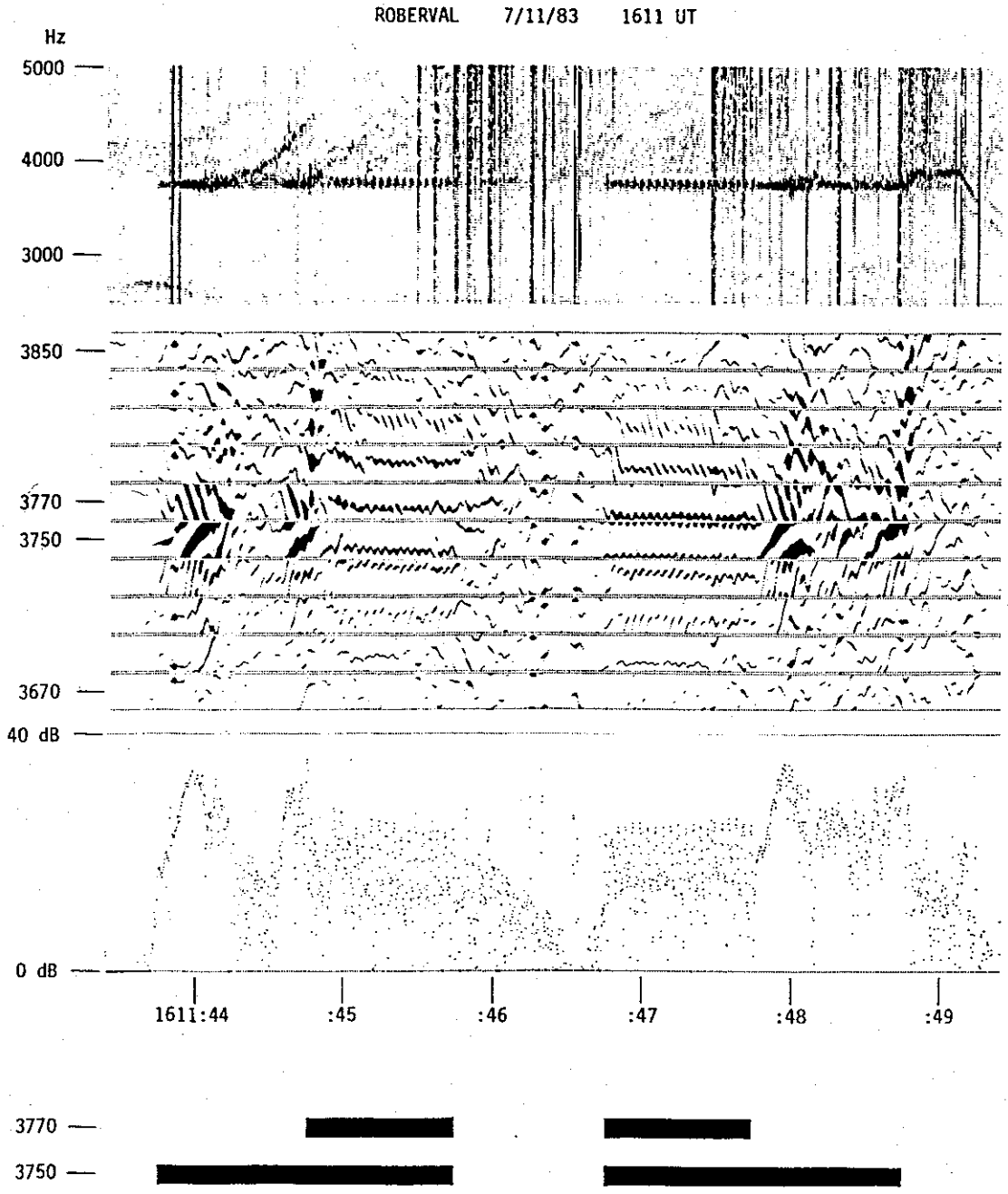

Figure 4.23. Another CBVA pulse pair as in Fig. 4.22. Now, the upper tone at 3770 Hz is off at the beginning of the first pulse and the end of the second pulse.

down into a state of phase coherence with multiple sidebands. The initial phase of the single-tone signal at the beginning of the pulse is nearly the same as its phase during this two-tone interval, as best as can be determined. There is a little slow phase drift due to duct motion, but it can be ignored. The peak magnitude during the two-tone interval is about 10 dB less than the maximum attained with only one tone.

At the end of the pulse there is very little triggering, at least on the dominant path. Termination triggering is suppressed by the two-tone signal. This may be because triggering requires a phase

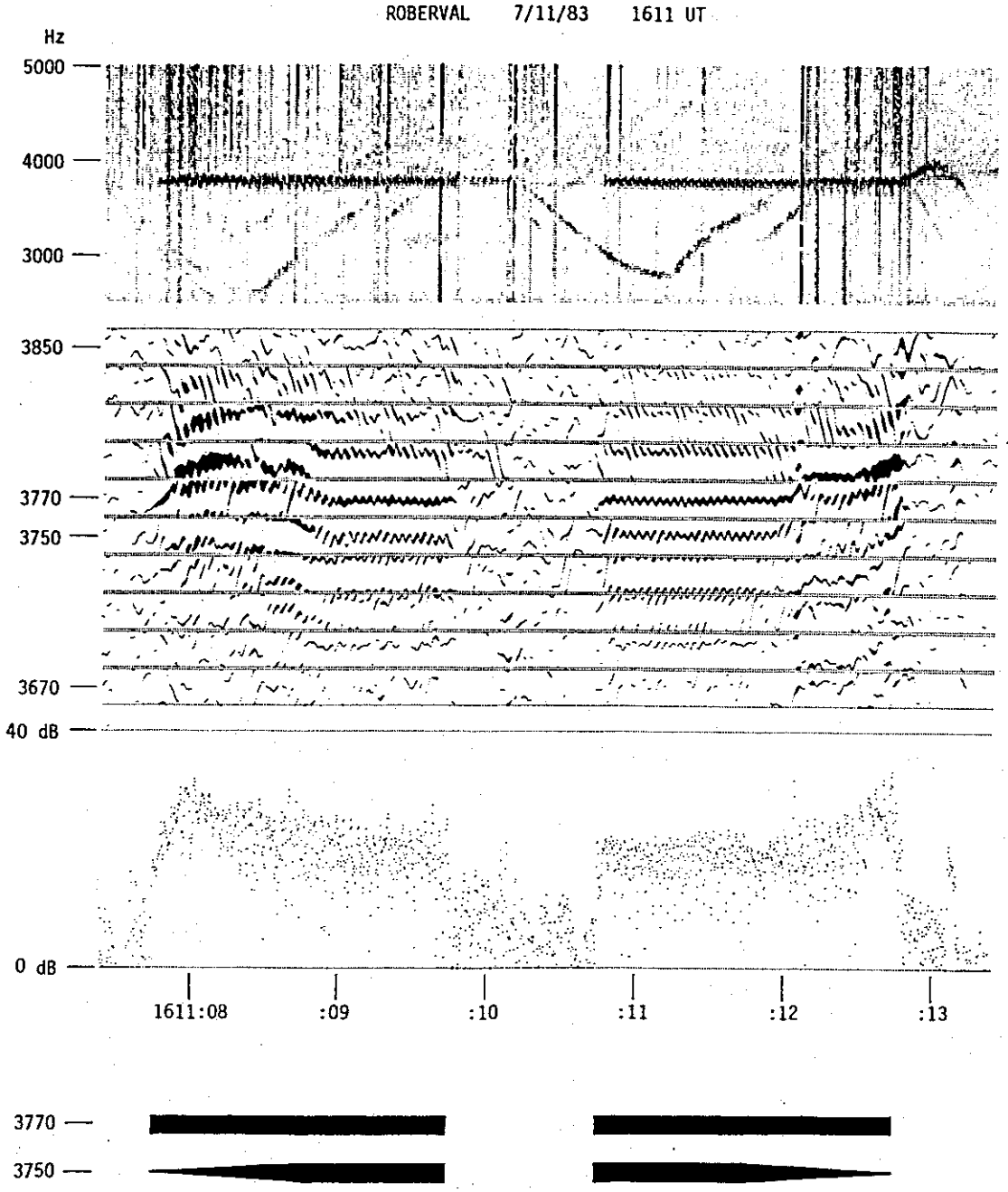

Figure 4.24. A third CBVA pulse pair as in Fig. 4.22. The amplitude of the 3750 Hz tone is turned up slowly, at a rate of 20 dB/s, during the first second of the first pulse, and turned down slowly at −20 dB/s at the end of the second pulse.

advance whereas the two-tone signal output remains in phase with the input. It may also be an amplitude effect, since the peak of the two-tone signal stays about 10 dB below the single-tone saturation level.

2. *Pulses with a single tone following a two-tone interval.* These are the second pulses in Fig.s 4.22 and 4.23. They behave similarly to the first pulses. During the two-tone interval, growth and phase advance are suppressed, and sidebands are generated. As soon as the second tone is

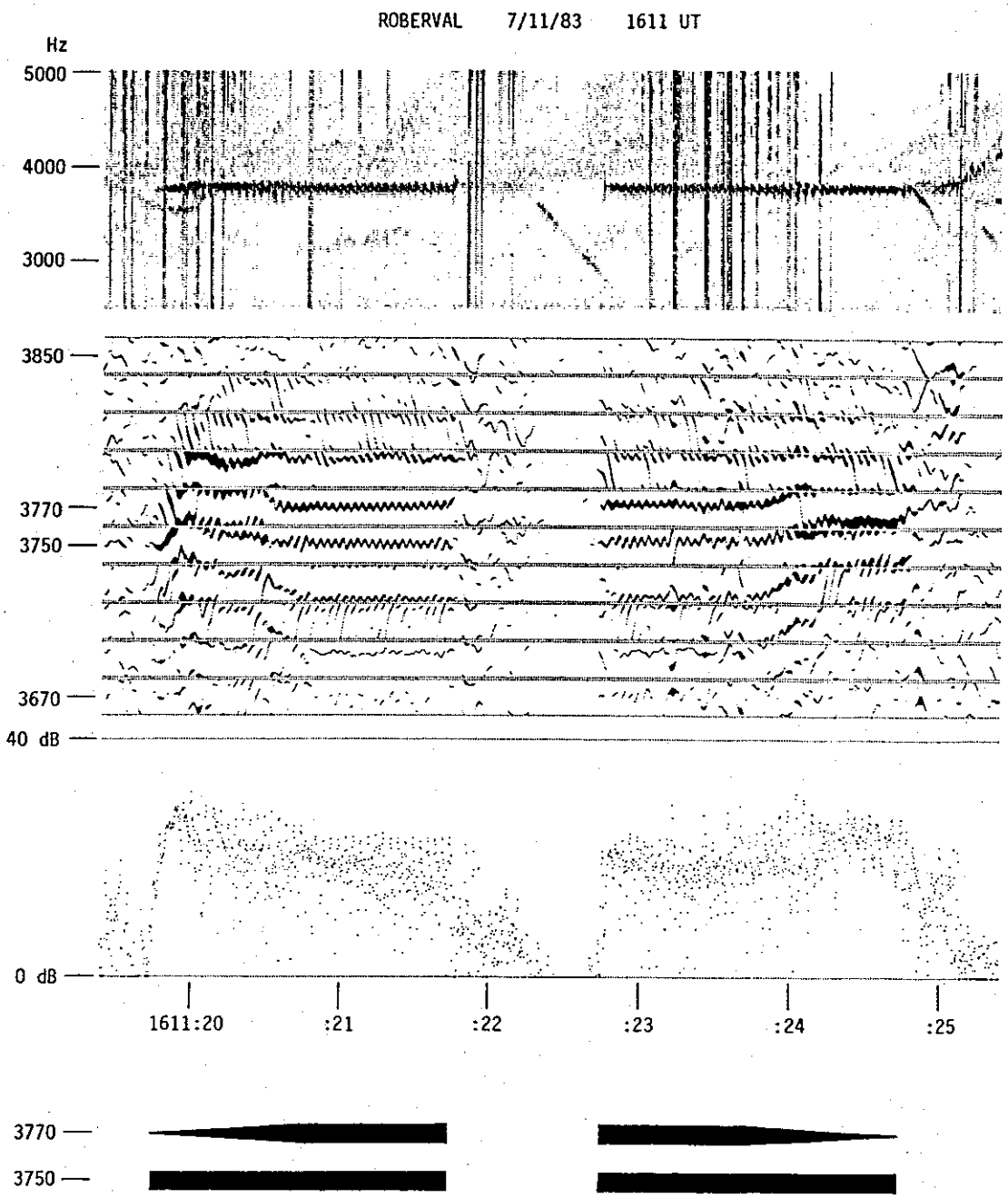

Figure 4.25. A fourth CBVA pulse pair as in Fig. 4.22. In this case, the upper-frequency tone at 3770 Hz is turned on and off slowly.

turned off, the remaining signal grows, shows phase advance, and triggers. Triggering occurs after a wrap-up of 1.5 rev, or at the end of the pulse, whichever is reached first. Again, it makes no difference whether the remaining tone is the upper or lower one—growth begins at the remaining frequency, and phase advance starts off from the phase maintained during the two-tone interval.

3. *Pulses with an amplitude ramp following a two-tone interval.* Now we will consider the second pulses in Fig.s 4.24 and 4.25. These pulses start with 1 s of two-tone signal. Each component is stable in phase, shows no growth (at least after the first 100 ms), and generates sidebands. Three

upper and three lower sidebands at 20 Hz spacing can be seen in Fig. 4.24.

After one second, the amplitude of one input component is attenuated at a rate of 20 dB/s, and this produces some very interesting effects. Nothing happens until the amplitude of this component has decreased 3 or 4 dB. Then both input components show a gradual phase advance. By the time the variable component is about 10 dB below the constant one, the phase advance has peaked at about 3/4 rev. The phase then stabilizes and remains relatively steady for the rest of the pulse, even though the input amplitude of the variable component continues to decrease to −20 dB. This 3/4 rev phase advance affects all signal components to some degree though it may affect the variable input component the most.

At the same time as the phase advance occurs, the total output magnitude grows about 5 dB above its equal-input level and remains there during the last 0.5 s of the pulse. This level is still roughly 10 dB below the saturation level seen during the one-tone intervals in Fig.s 4.22 and 4.23 (it's a bit higher in Fig. 4.24, a bit lower in Fig. 4.25). Some components seem to grow more than others as the variable input component is turned down. The biggest growth occurs on that component which lies just above the constant-amplitude input tone. If this is the first upper sideband (the lower input tone is being decreased while the upper one is constant), it may become the strongest component during the last 0.5 s of the pulse. If the first component above the constant input tone is the variable one itself (as in Fig. 4.25), we have the paradoxical situation where that component whose input amplitude is attenuated becomes the strongest component in the output.

At the end of the pulse there is termination triggering. The emission generated has the same amplitude as those terminal emissions generated at the end of the single-tone pulse segments, but is not as strong as the pre-termination emissions on the single-tone pulses.

4. *Pulses with an amplitude ramp preceding a two-tone interval.* These are the first two pulses in Fig.s 4.24 and 4.25. Here the constant-amplitude tone, which is initially 20 dB stronger than the variable one, shows growth and phase advance during the first 200 ms. The growth and advance are suppressed, however, compared to single-tone pulses. The phase advance is only about 1 rev and the amplitude does not quite reach saturation.

After the first 200 ms, sidebands appear and gradually the signal settles down to a period of phase coherence much like that during the amplitude decrease in the pulses discussed above. During this interval, which lasts approximately until amplitude parity is reached, phases are somewhat (3/4 rev) advanced from their initial (and final equal-input) values. The strongest component is again the one above the constant input tone. (In other similar pulses in this record, the second higher component is often the biggest.)

Finally, during the last second, the system settles down to the phase coherence and sidebands seen in all the equal-amplitude two-tone signals. At the end of the pulse there is no phase wrap-up and no appreciable termination triggering.

Figure 4.26 shows these last two pulses in more detail. Note that the steady-state phase during the equal-input interval at the end of the pulse is not exactly the same as the initial phase of the growing constant-amplitude component, but is slightly advanced from it. (Similar advances occur in the first pulses in Fig.s 4.22 and 4.23, but are somewhat obscured by the larger and more irregular phase drifts due to duct motion.) That is, even though two-tone transmissions suppress temporal growth, there seems to be a small phase advance at the output of the interaction region compared to the input. This may be indirect evidence of steady-state signal amplification. On the other hand, lacking an absolute phase reference, we cannot be sure that the steady-state suppressed phase is not the actual phase of the input signal. In this case, what we really see is a transient lag in phase at the

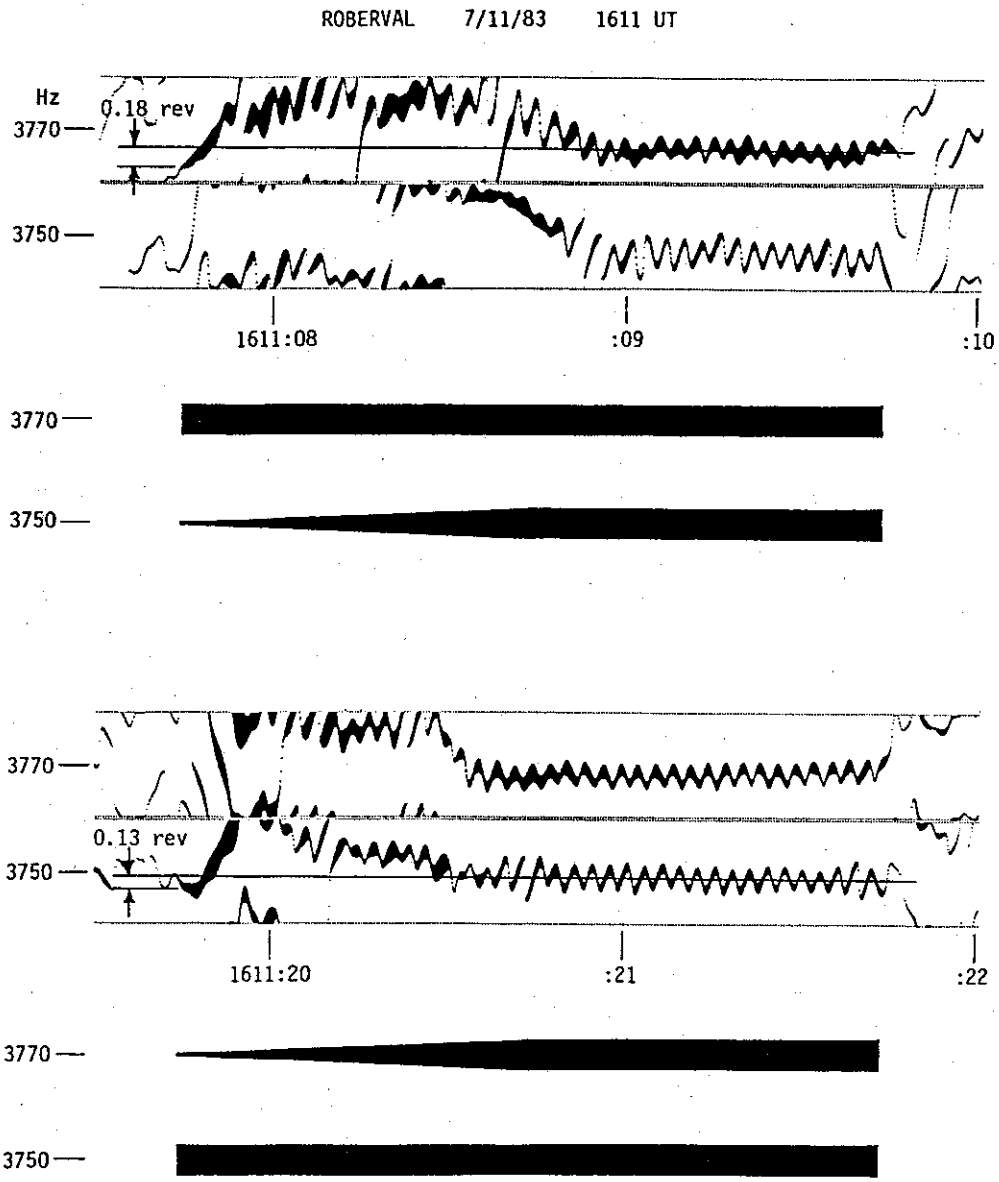

Figure 4.26. A closeup look at the first CBVA pulses in Fig.s 4.24 and 4.25. Rapid growth and phase advance begins on the stronger component, but is soon arrested as the other input tone is turned on. Note, in the steady-state growth-suppressed state at the end of the pulses, that the output phase is still slightly advanced compared to the input signal. The advance is about 0.18 rev in the first case, and 0.13 rev in the second.

beginning of the pulse, perhaps like those predicted by some of the theoretical models mentioned in Sec. 4.1. I'm not sure how this ambiguity can be resolved.

Summary of Growth Suppression by Multi-Component Signals.

1. Components in multiple-frequency signals mutually suppress each other's growth and triggering when their spacing is 10–50 Hz. Peak suppression occurs with 20 Hz spacing. Suppressed signals generate additional phase-coherent sidebands at multiples of their spacing.
2. Phase behavior can be used as well as the rates of growth and triggering to monitor mutual interactions between signal components. Relative phase can distinguish between those cases

where some growth occurs but components remain phase-coherent (the AM pulses with 50 Hz modulation in Fig. 4.21) and those where components behave independently.

3. The steady-state phase of each component in a two-tone signal is slightly advanced, say 0.15 rev, from the initial phase a one-tone signal has before growth begins. This may be indirect evidence of the linear amplification of two-tone signals.
4. If one component of a two-tone input signal is 10–20 dB lower in amplitude than the other, both components may show an additional phase advance of about 3/4 rev with respect to the equal-input case, and a slight increase in total amplitude. Suppression weakens, but the output signal is still phase-coherent.
5. The strongest output component from an unequal-amplitude two-tone input signal is usually the one immediately above the frequency of the larger input tone. This will be the first upper sideband if the upper input tone is the largest. If the input signal has a weak upper tone and strong lower tone, the upper tone may paradoxically be stronger at the output.
6. A well-suppressed pulse triggers no emissions. This may be because there is not enough phase advance accumulated through the interaction region, or it may be due to the decreased output amplitude. However, an unequal-amplitude (20 dB difference) two-tone signal, possibly because of the 3/4 rev phase advance allowed, can show termination triggering as strong as that of a single-tone signal.

4.5.2 Entrainment of Emissions by Idler Pulses

Figure 4.27 shows an *f-t* spectrogram and a gray-scale plot of the ULF75 transmission described in Section 4.1. The 0.5 s pulses at 4500 Hz show growth in amplitude with a total phase advance of 1.2–1.3 rev, 60 Hz sidebands, a BLI at the end of each pulse, and a faller. The interesting feature here is the behavior of the faller. From 1417 to 1418 in this record there are many examples of entrainment of the faller by one of the 50 ms idler pulses at 4100 Hz. The entrainment is always by the middle idler pulse of the five pulses at 4100 Hz, which starts 250 ms after the end of the 4500 Hz pulse. Spectrograms show entrainment as usually resulting just in enhancement of the amplitude of the idler pulse. The faller hits the front edge of the middle idler pulse and this pulse becomes much stronger than the two preceding or following it. The faller may stop once it hits the idler, or it may continue for, say, 100 Hz below the idler. However, in some cases, as with the pulses in Fig. 4.27, the middle idler pulse becomes so strong that it generates a second faller at its trailing edge. In this case the faller from above seems to hit the idler pulse, have some of its energy entrained for 50 ms, then be released again at the end of the idler pulse when it falls for as much as another 200–300 Hz.

The enlarged gray-scale plot in the middle of Figure 4.28 shows in more detail what happens when the last faller in Fig. 4.27 hits an idler pulse. The top plot shows the magnitude and phase of the 4100 Hz idler pulses themselves. Whenever entrainment occurs in this record, the middle (third) idler pulse at 4100 Hz is typically stronger (by 14 dB in Fig. 4.28) than the four pulses on either side of it, and from 0.1 to 0.4 rev advanced in phase with respect to them (0.25 rev in Fig. 4.28). Note in Fig. 4.28 that a small amount of growth occurs, as the peak amplitude of the entraining pulse is several dB larger than either the faller or the unaffected idler pulses (or their sum). For comparison, the bottom plot in Fig. 4.28 shows a gray-scale phase plot of a synthetic signal made up of a ramp falling at 3200 Hz/s to simulate the faller, and some 50 ms pulses at 4100 and 4000 Hz to simulate the idler. The idler pulses were made 10 dB lower in amplitude than the ramp to approximate the ratio of amplitudes seen in the real signal. The point to notice is that the two synthetic signals combine linearly, as expected. There is a little beating when the ramp crosses

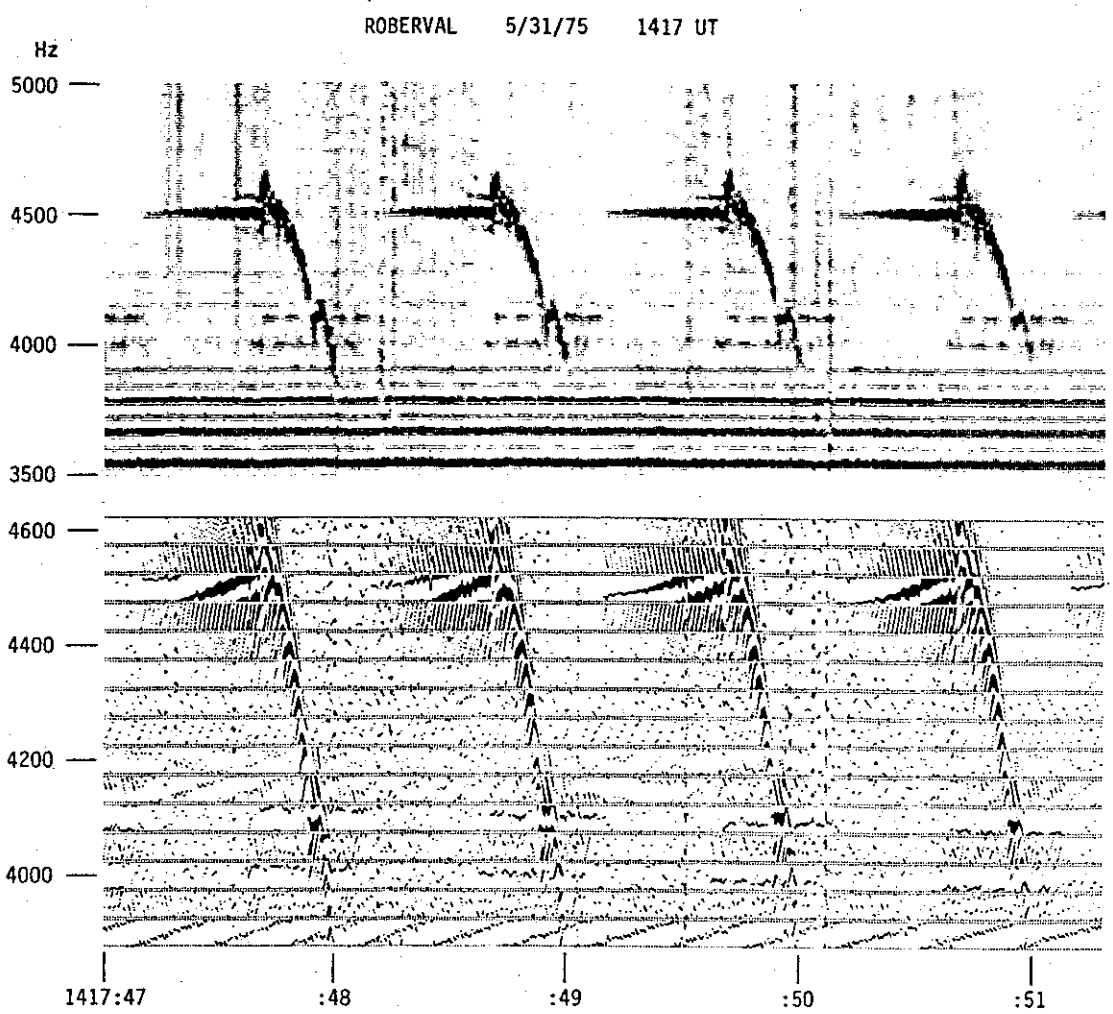

Figure 4.27. Half-second pulses at 4500 Hz from Siple showing entrainment of terminal fallers by 50 ms idler pulses. When the falling emission intersects the third of the five 4100 Hz idler pulses, it causes that pulse to be larger in amplitude and advanced in phase compared to its neighbors. A second faller is triggered at the end of the third idler pulse. (Gray-scale phase plot BW = 80 Hz, P_{span} = 1 rev.)

the middle 4100 Hz pulse, but there are no non-linear effects such as the entrainment and relatively constant phase advance, or the amplification, seen in the real signal.

As well as being amplified, the middle idler pulse also lasts longer than its neighbors. In Fig. 4.28 the entraining pulse is about 80 ms long compared to the 50 ms duration of the others. This is because the entraining pulse is so strong that it triggers a small faller and so lasts about 30 ms longer than the transmitted signal. If a faller is released at the end of the idler pulse in the ULF75 program, it appears to separate after a very small phase wrap-up (0.12 rev in Fig. 4.28). The magnitudes and phases of the two idler pulses following the entraining pulse are exactly the same as the two idler pulses which preceded it; the entrainment has no apparent lasting effect. In a few cases where a second faller is released from the end a 4100 Hz idler pulse, it is seen to fall and itself become entrained by the fourth 4000 Hz idler pulse, giving a small amplification and phase advance to that pulse. For comparison, the synthetic data in the bottom of Fig. 4.28 does not show a lengthening of the middle idler pulse, nor a second faller. The first faller passes through the idler

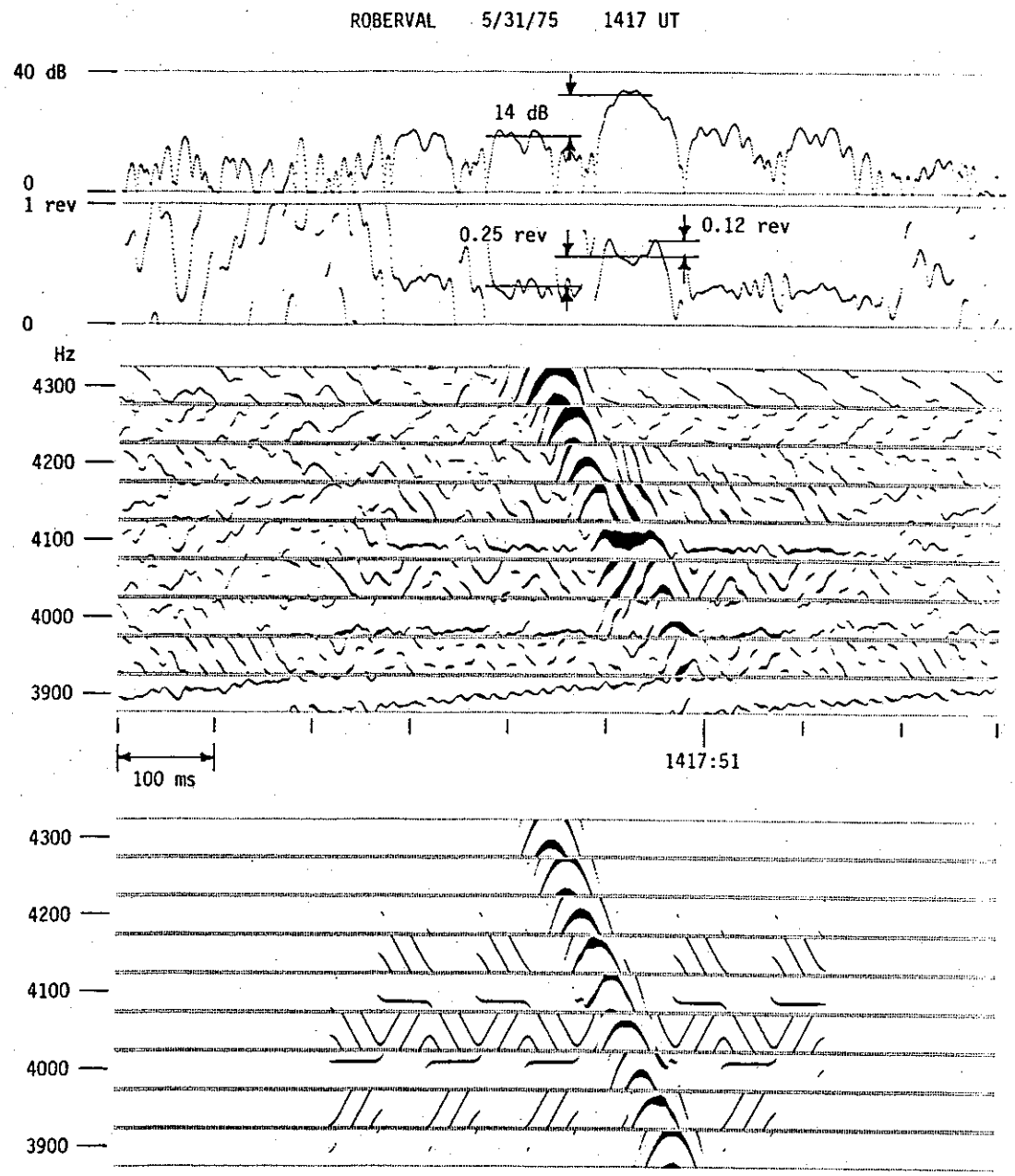

Figure 4.28. Details of the entrainment of the last faller in Fig. 4.27 (BW = 80 Hz, P_{span} = 1 rev in all plots). The magnitude-phase plot at top shows the 4100 Hz idler pulses. The middle pulse has been amplified by 14 dB and advanced in phase an average of 0.25 rev compared to its neighbors due to entrainment of the faller. The middle pulse continues for 80 ms; after the 50 ms input signal ends the phase wraps up an additional 0.12 rev as a faller is generated. The middle gray-scale plot shows the entrainment and the second faller graphically. For comparison, the bottom gray-scale plot shows synthetic signals—a ramp falling at −3200 Hz/s plus 4100/4000 Hz idler pulses at a level 10 dB lower. Intersecting signals combine linearly and no amplification, phase advance, pulse stretching, or second faller are seen.

pulses unaffected.

Discussion. According to the *Helliwell* [1967] model, rising and falling emissions are generated in an interaction region which is downstream or upstream, respectively, from the equator. *Helliwell and Katsufrakis* [1978] note that entrainment occurs too rapidly to be explained by the drift of this region from the location of the emission to the location where the entraining signal might resonate, and conclude that the entrained emission continues to be generated in the same place, but has its slope df/dt changed to that of the entraining signal. This will reduce the length of the interaction region, yet the phase-bunching now provided by the entraining input signal can maintain the amplitude of the original oscillation.

The slope of an emission is determined by the distance S from the equator to the center of the interaction region, as [*Helliwell* 1967, Eq. (15)]:

$$\frac{df}{dt} = S\frac{54cf_{Heq}^2}{r_{eq}^2 f_N}\frac{\lambda^{3/2}(1-\lambda)^{3/2}}{(1+2\lambda)^2}\left[1+\frac{(1-\lambda)}{3}\tan^2\alpha\right], \tag{4.4}$$

where $\lambda = f/f_H \approx f/f_{Heq}$. We can invert this, of course, and find the distance from the slope. The entrained emission in Fig. 4.28 has a slope of $df/dt = -3200$ Hz/s. In Sec. 4.1 we found the path of the Siple signals in this record to be at $L = 4.36$, with an equatorial electron density $N_{eq} = 260$ electrons/cm^3. The entraining pulse at 4100 Hz is at a relative frequency of $0.386 f_{Heq}$. Assuming that the dominant electrons have pitch angles of $\alpha = 45°$, the emission in Fig. 4.28 is generated by an interaction region at $S = -4420$ km, or 4420 km upstream (down-wave, or north) from the equator. This is well beyond the normal equatorial interaction region for constant-frequency signals.

The 4100 Hz idler pulses show no growth or phase advance, and presumably undergo no significant interactions with electrons near the equator. Yet as they cross the equator and approach the region where the emission is oscillating, these weak signals are able to control that interaction. There are two important features in Fig. 4.28. First, the entrained signal at the output of the emission interaction region is phase-locked to the 14 dB smaller controlling input signal, with a relatively stable phase advance of 0.25 rev. This is similar to the phase advance seen on normal growing signals after an equivalent amount of growth has taken place. Second, a termination faller occurs after a small phase wrap-up (0.12 rev) at the end of the controlling signal, just like fallers after short growing pulses. Both features show that growth occurs in the off-equatorial interaction region very much as it does on the equator, and point up the sensitivity of all interaction regions to weak signals on their downstream (wave input) side.

Finally, note in the spectrogram in Fig. 4.27 the slopes of the emissions triggered at the ends of the 4500 Hz pulses, and how they steepen with time. When each 4500 Hz pulse terminates, its interaction region is near the equator. It takes about 250 ms to drift upstream the 4420 km needed to reach a slope of -3200 Hz/s. However, each faller released at the end of the entraining idler pulse reaches this slope almost immediately; the interaction region does not have to drift. This means that entrainment indeed occurs at the location of the entrained emission. The small phase wrap-up when the final faller is released shows that even here triggering causes the emission to start with a positive frequency offset, however brief.

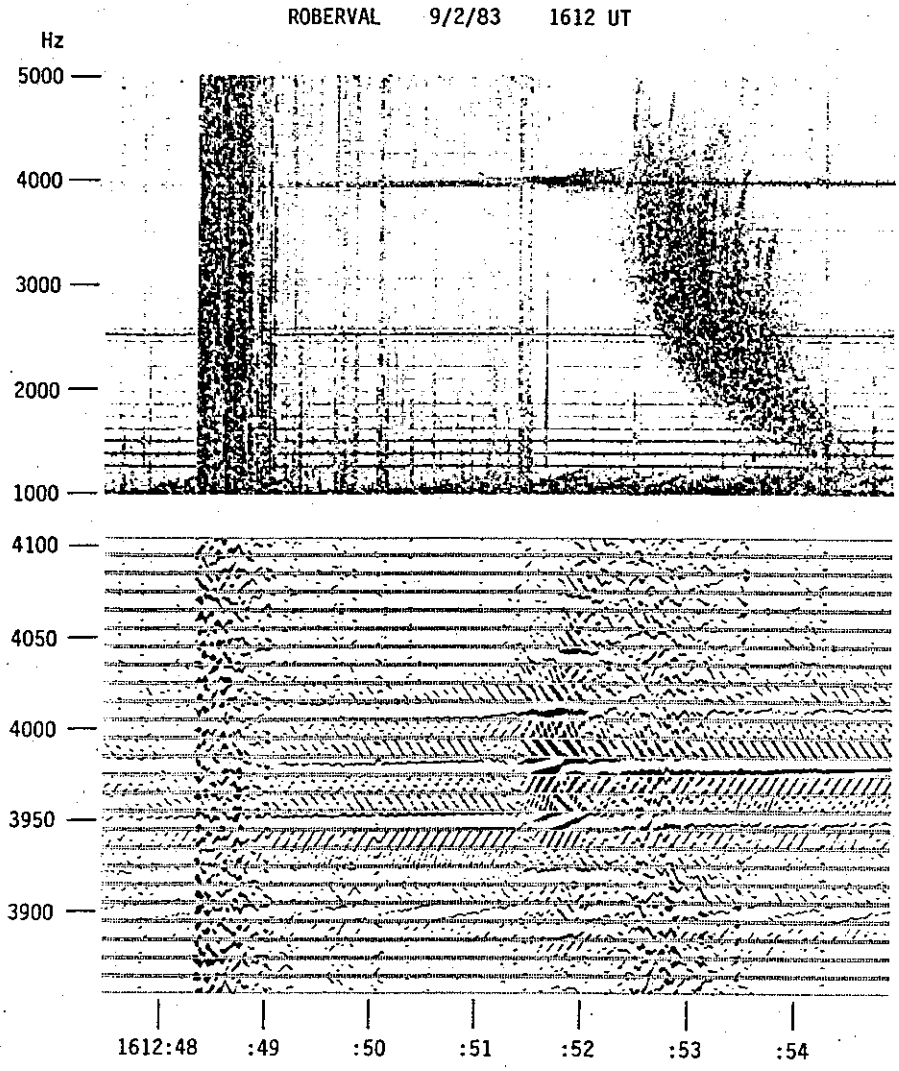

Figure 4.29. A two-tone LICO1 transmission from Siple showing a whistler precursor. Shortly before the whistler arrives, the two-tone components at 3950 and 3980 Hz show amplification and phase advance. Sideband effects are also enhanced. (Gray-scale phase plot BW = 20 Hz, $P_{span} = 1$ rev.)

4.5.3 Whistler Precursors on Transmitter Signals

Figures 4.29, 4.30, and 4.31 show examples of the two-tone LICO1 signal described in Section 3.4 (see also Fig. 2.2). A very interesting event occurs when the Siple signal meets a whistler. The Siple signal at 3950/3980 Hz is just at the nose frequency of the first whistler components, and passes through the tops of the whistlers. In several instances it shows considerable amplification about 1 s before the whistler—a *precursor*. In each case the amplification lasts for about 1 s. Good examples occur in this record at 1609:14, 1610:45, 1611:53, 1612:52, 1614:21, and 1615:30, and there are probably additional, though weaker, events. Fig.s 4.29–4.31 illustrate the last three cases.

The precursors cause the two transmitted tones and their sidebands to grow, usually with a noticeable phase advance. In most cases the phase advance amounts to, say, 1/3 rev. However, in the precursor in Fig. 4.29 the 3950 Hz tone advances 2 revs in 0.5 s (+4 Hz offset), the 3980 tone advances 1 rev, and the first upper sideband at 4010 Hz advances about 0.5 rev. After each

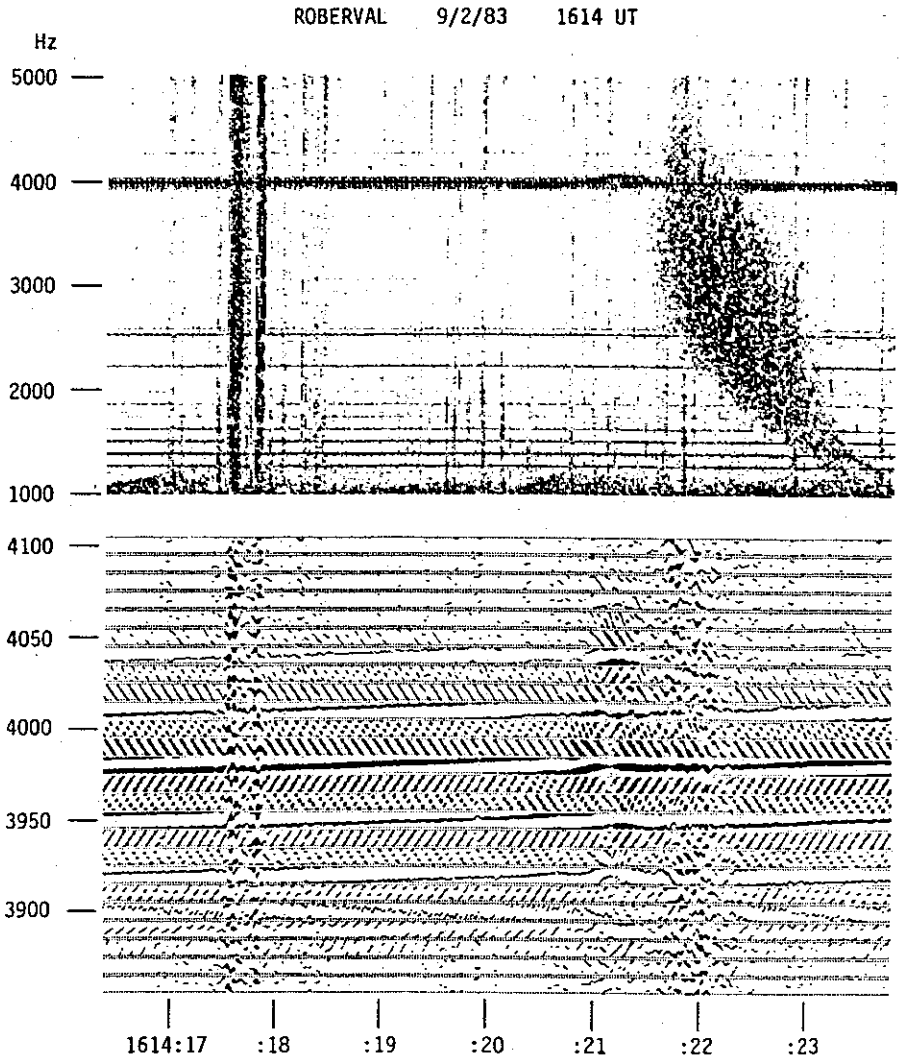

Figure 4.30. A second whistler precursor on a two-tone transmission, as in Fig. 4.29. The phase advance caused by the precursor is much smaller, only a fraction of a revolution. The upper transmitted signal at 3980 Hz shows a null in the middle of the precursor.

precursor, the transmitted tones and sidebands decrease in magnitude and phase and resume their previous behavior. The very slow drifts in phase (less than ±0.2 Hz) seen in Fig.s 4.29–31 are the micropulsation-related effects discussed in Section 3.4 and are not connected with the precursors.

As also mentioned in Sec. 3.4, the LICO1 signal has a group delay of $t_g = 2.1$ s. Whistler components with that delay (the first strong components in Fig. 4.29 with a two-hop time of 4.2 s) have nose frequencies close to $f_n = 4000$ Hz. Using these values, we found from *Park* [1972] a path L-value of $L = 4.32$ and a tube content of $N_T = 3.5 \times 10^{13}$ electrons/cm^2, or equatorial density $N_{eq} = 280$ electrons/cm^3, typical of magnetically quiet conditions.

There is a peculiar null in the magnitude of the upper Siple tone in the middle of each precursor. In Fig. 4.29 this appears as the tone undergoes a phase reversal as if from interference or fading. Shortly after the null a weak rising emission is triggered. In Fig. 4.30 the null occurs without any marked phase effect, though it seems to precede (trigger?) a burst of growth on the lower tone and the sidebands. And in Fig. 4.31 the null coincides with a short burst of signal at nearby frequencies,

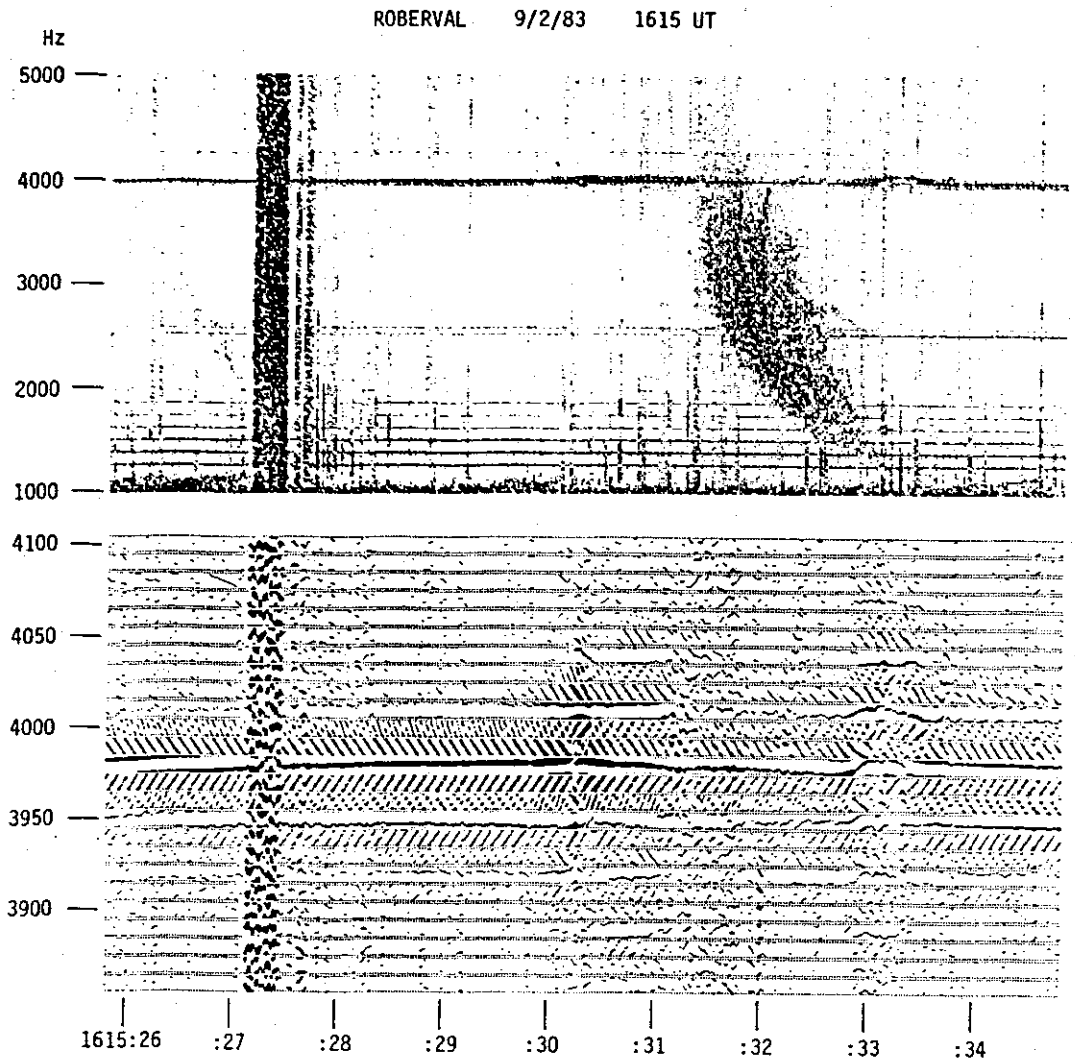

Figure 4.31. A third whistler precursor as in Fig. 4.29. This time there is a "postcursor" shortly after the whistler as well.

almost like a BLI.

There is also a certain variability in the timing of the precursors from one event to the next. In Fig. 4.29 the precursor starts 3.1 s after the spheric, 3.3 s in Fig. 4.30, and only 2.6 s in Fig. 4.31. The corresponding times from the spheric to the magnitude null are 3.5 s, 3.6 s, and 3.1 s, respectively. This variability has been noted before by *Helliwell* [1965, Fig. 7-53.].

In two cases in this record, at 1611:55 and 1615:33, a second amplification event lasting about 1 s also occurs *after* the whistler, a "postcursor." At 1611:55 the postcursor starts about 0.8 s after the first whistler component, or 2.0 s after the precursor. At 1615:33 (Fig. 4.31) the postcursor is about 1.4 s after the whistler, or 2.7 s after the precursor. The postcursors may be related to the precursors, but more likely they represent the entrainment of emissions from the tops of the later whistler components, and we will not discuss them further.

Discussion. The examples in Fig.s 4.29–4.31 show a definite correlation between whistlers and effects seen on the transmitted LICO1 signal. The effects are the momentary amplification of the LICO1 signal accompanied by a phase advance of 0.5 rev or so, and the enhancement of its various

30 Hz sidebands. We found in the preceeding sections that these effects are the characteristic features of growth due to cyclotron resonance with energetic electrons. The precursor behavior in Fig.s 4.29–4.31 indicates a sudden increase in growth activity. In fact, the growth is stong enough in Fig. 4.29 that the signal even triggers an emission. Whistler precursors have been observed before as spontaneous emissions occurring shortly before two-hop whistler echoes [*e.g., Helliwell,* 1965, Fig.s 7-52–7.54]. However, as far as I know, this is the first time precursors have been reported on a transmitter signal. This case is also unique in that the frequency of the precursor is quite high relative to the associated whistler—near the whistler nose frequency.

Various models have been developed to explain precursors. *Dowden* [1972] explains precursors as emissions triggered by hybrid whistlers (whistlers which propagate one hop sub-ionospherically before entering the magnetosphere). *Reeve and Boswell* [1976] propose that a strong whistler decays parametrically into a backward-moving whistler and a low-frequency ion-acoustic wave. The backward-moving whistler triggers the precursor. *Reeve and Rycroft* [1976] postulate that an unducted whistler is magnetospherically reflected at the end of its first hop and enters a duct midway through its second hop to trigger a precursor, slightly ahead of the two-hop ducted whistler. These three studies each assume that conditions are ripe for the growth of an emission, and the point of each model is to produce a triggering signal that has the right time delay with respect to the causative spheric and the two-hop whistler. However, to explain Fig.s 4.29-4.31 the problem is not to find a triggering signal but to find a mechanism that momentarily increases growth activity. None of these three models can do that. Neither are the LICO1 precursors due to the entrainment of a strong triggering signal by the LICO1 input, since the proposed triggering signals all enter up-wave of the LICO1 interaction region. (The precursors are different from the entrained fallers in Fig. 4.27 because the entrained signal, the fallers in that case, must be caused by growth which occurs down-wave of the entraining signal, the constant-frequency idler pulses. We could possibly have entrainment *of* the LICO1 interaction *by* a triggering signal, but that would be manifest as a change in the slope df/dt of the received LICO1 signal to that of the triggering signal, which is not what we see here.) In addition, the parametric decay model of *Reeve and Boswell* [1976] can be ruled out because it only works at the low frequency tail of a whistler, not at the nose frequency.

Park and Helliwell [1977] propose a mechanism that starts with a longitudinal resonance interaction between the outgoing whistler and co-streaming energetic electrons. This perturbs the electron energy distribution. The perturbation then gives rise through cyclotron resonance growth to a precursor moving in the opposite direction which arrives before the two-hop whistler echo. They also propose that power line harmonics provide a seed signal which is amplified to trigger the precursor. *Tkalcevic* [1982] studies the details of the longitudinal interaction, and supports the model of *Park and Helliwell* [1977].

Reitveld [1980] studies a particular narrowband type of precursor he calls a "monochromatic precursor start." This is a single-frequency signal which lasts for up to 200 ms and may then trigger the rising emission of typical precursors. He uses phase analysis (phasogram technique) to show that the frequencies of these signals are not related to harmonic frequencies of the power grid. He proposes that a monochromatic precursor may still be triggered from a power line harmonic, but be offset from it by up to 100 Hz, as many emissions are offset from their triggering signals.

The longitudinal resonance interaction of the *Park and Helliwell* [1977] model might explain the sudden increase in growth activity seen in Fig.s 3.29–3.31. To this extent, our observations support their model over earlier ones that merely produce triggering signals. However, as for the role of power line radiation in initiating precursors, our observations have little to say. The precursors we observed start (and remain, for the most part) at the frequency of the triggering signal. We did

not observe triggered emissions suddenly beginning up to 100 Hz above the transmitter signal. This may be because the two-tone LICO1 signal with 30 Hz separation suppresses growth and emission triggering. A power line harmonic, if there be such, would not be self-suppressed and might be more likely to trigger an emission, possibly somewhat offset in frequency. On the other hand, it seems to me as likely that an emission should arise spontaneously or be triggered by noise, as be triggered by an unseen power line signal. The real test of power line harmonics as triggers for precursors must be a statistical analysis of precursor starting frequencies. The only study so far, *Reitveld* [1980], found no particular correlation with power line harmonics, either at multiples of 50 or 60 Hz.

4.6 Magnetospheric Line Emissions and Power Line Radiation

Evidence for Power Line Radiation. The concept of power line radiation is that currents at harmonics of the fundamental power frequency (50 or 60 Hz, depending on country) radiate electromagnetic waves, some of which are coupled into the magnetosphere and produce various effects. Power line radiation (PLR) was first mentioned by *Helliwell et al.* [1975]. This paper started a sometimes spirited discussion of the evidence for and against PLR and its possible effects, which continued in the literature for the next few years. Reviews of PLR and bibliographic references can be found in *Park and Helliwell* [1978], *Helliwell* [1979a], *Park and Helliwell* [1981], and *Park et al.* [1983].

We have already seen signals caused by induction fields from local power lines, as in Fig. 3.4. These signals are narrowband ($\approx$ 1 Hz bandwidth), at mainly odd-numbered multiples of the power frequency, and extend at times to frequencies above 6 kHz. All harmonics are phase-coherent with the fundamental component. (Occasionally additional lines are created by rotating machinery; these lines have their own independent harmonic structure, as in Fig. 4.19.) However, the induction lines are due to local magnetic fields coupled to the loop antenna of the receiver. They demonstrate the presence of harmonic currents flowing near the receiver, but are not themselves radiated signals.

Little information is available about the power actually radiated by power line harmonic currents, though it is likely to be on the order of milliwatts or watts rather than kilowatts. *Barr* [1979] measured no significant (> 1 μV/m) harmonics above 1.5 kHz from New Zealand power lines, though he did see a radiated component at 300 Hz from the Benmore DC power line (≈ 10 μV/m at Stewart Island). *Yearby et al.* [1983], studying harmonic currents in the Newfoundland power system, estimate a radiated power of 0.05–0.5 μW per transmission line for components in the range 2.7–3.7 kHz. Radiated harmonics from the North American power grid have not themselves been seen, and all evidence for the existence of radiated power line harmonics is indirect, based upon their postulated effects as follows:

1. *Emission Cutoff, Entrainment, and Change of Slope. Helliwell et al.* [1975] present a record where emissions triggered by pulses from the Siple transmitter seem to change slope, become entrained, or terminate near the frequency of local induction lines. This is seen as evidence of wave-wave interactions between the emissions and magnetospheric signals with components at the frequencies of the induction lines; that is, PLR.

2. *Magnetospheric Line Emissions.* These are whistler-mode signals containing noisy (10–30 Hz bandwidth) lines roughly equally-spaced in frequency. *Helliwell et al.* [1975] show a case with a line separation of roughly 120 Hz, and claim that the frequencies of such lines are usually offset 20–30 Hz above the local induction lines. They interpret the lines as due to PLR components which are amplified during echoing and which may trigger narrowband emissions. They also note that lines are occasionally seen just below power lines frequencies instead of above

them; and that lines may drift in frequency, more often up than down, at rates as large as 50 Hz/min. They show cases of lines with spacings of only 20–30 Hz. *Park* [1976] states that lines are often offset 30–50 Hz above induction line frequencies, similar to the offset seen on emissions triggered by Siple pulses. He finds that magnetospheric lines are strongest during the recovery phase following a magnetic storm. *Park* [1977] notes that magnetospheric lines only occur during good whistler-mode echoing. *Park and Chang* [1978] simulate magnetospheric lines with a multi-component Siple transmission, and show that amplified lines can be seen at Roberval with radiated powers from Siple as low as 0.5 W. *Matthews and Yearby* [1981] find that magnetospheric lines at Halley Station, Antarctica, have a wide range of spacings from 50 to 90 Hz, with line bandwidths of 20–30 Hz. Line emissions have also been seen on satellites [*Koons et al.*, 1978; *Park and Helliwell*, 1981; *Bell et al.*, 1982].

3. *Geographic Concentration of Mid-Latitude Chorus. Luette et al.* [1977] report that VLF chorus in the 2–4 kHz band as observed by the OGO-3 satellite tends to be concentrated in longitude over four industrial areas in the northern hemisphere, presumably because chorus elements are triggered by PLR. *Thorne and Tsurutani* [1979] and *Tsurutani et al.* [1979a, 1979b] claim that the correlation with location found by *Luette et al.* [1977] was due to statistical errors caused by oversampling particular chorus events. They also state that chorus occurs outside the plasmapause, where triggering from PLR would not be expected (most triggering being at lower latitudes inside the plasmasphere). They present their own studies of ELF chorus on OGO-5 and ELF hiss on OGO-6 and find no correlation with longitude. The argument is continued by *Park and Helliwell* [1980], *Tsurutani and Thorne* [1980], *Russell* [1980], and *Luette et al.* [1980]. *Bullough and Kaiser* [1979] review studies of ELF/VLF signal amplitudes made with the Ariel 3 and 4 satellites. They find that signals at 3.2 kHz over North America in summer are stronger than anywhere else, which may be due to PLR but might also be correlated with thunderstorms. *Park et al.* [1981] show that there are two kinds of chorus, one which occurs outside the plasmapause and is correlated with energetic particle injection events, and the other which occurs inside the plasmasphere and may be controlled by PLR. If so, the OGO-3 and OGO-5/6 investigators may be observing and arguing about different phenomena.

4. *Initial Frequencies of Whistler Precursors and Chorus. Park and Helliwell* [1977] claim that whistler precursors seem to begin at the frequencies of power line harmonics. *Rietveld* [1980] does not find such a correlation. Precursors are discussed above in Sec. 4.5.3. *Luette et al.* [1979] report that in about 15% of the cases of chorus seen on OGO-3, individual chorus emissions have well-defined starting frequencies. These cluster around harmonics of 50 or 60 Hz, depending on geographic location.

5. *Weekend Effect. Park and Miller* [1979] report that the broadband amplitude of signals from 2 to 4 kHz received at Siple, mostly chorus, is about 30% lower at midday on Sundays compared to the average at the same time during the rest of the week. They interpret this as due to lower power demand, and lower PLR, in the conjugate region on Sundays. *Thorne and Tsurutani* [1981] claim that the signal at Siple was not chorus but hiss, had little connection to unducted signal levels in the magnetosphere, and that there is no Sunday effect seen in ELF chorus (outside the plasmapause) on satellites. *Park and Miller* [1981] rebut the arguments of *Thorne and Tsurutani* [1981]. I note that if there are two kinds of chorus as mentioned above in Item 3, the parties here are again comparing apples and oranges.

Helliwell et al. [1975] note the limitations of standard spectrum analysis when trying to verify PLR effects: "The primary problem was our inability to measure frequencies accurately enough. In most

though there is some low-level natural noise between 2100 and 2300 Hz. At 1305:33.3 a Mini-DIAG (actually MDIAS—Miniature-DIAgnostic, Synoptic) signal arrives from the Siple Transmitter. The transmission format is shown at the top of the figure in spectrogram form made from a recording of the transmitter antenna current. The format consists of a two-second pulse at 3000 Hz (including an amplitude ramp from -10 dBc to full power during the first second), five 200 ms pulses at 3480, 3240, 3000, 2760, and 2520 Hz, two descending ramps from 3480–2520 Hz (the second one at -6 dBc), and a two-tone pulse at 3480/3500 Hz (with components at -6 dBc).

All Mini-DIAG components at 3240 Hz and below show growth. They trigger intense emissions on at least three different paths. These emissions, predominately fallers, echo, grow, and generate new emissions. Activity in Fig. 4.33 is mostly in the band 2100–2600 Hz. The two-hop time for the strongest echoes at 2300 Hz is about 6.3 s.

There are several important features to note in Fig. 4.33. First, the strong growth and echoing activity is somewhat restricted in frequency. Note how quickly echoes above 2700 Hz die away. The later multi-component Siple pulse near 3000 Hz at 1306:03 has almost no echo at all. This is a common phenomenon—growth often occurs only over a limited bandwidth. Yet within the band 2100–2600 Hz, growth is strong enough to overcome any inherent attenuation and the amplitudes of echoing components remain roughly constant.

Second, echoing components tend to spread out in time. Note how the Mini-DIAG ramps and their emissions, which have only a few components in the first-hop signal, have a few more in the third-hop echo, and have become diffused in time by the fifth-hop echo. There are two factors at work here. One is the effect of emissions, which extend the trailing ends of input signals and make an echoing element last a little longer each time. The other is coupling between whistler-mode paths. This can cause a given element to echo on a different path and then couple back to its original path with a different delay than signals echoing continuously on that path. This twinning of echoing elements spreads and mixes them in time.

Third, as the Mini-DIAG elements echo and diffuse in time, they develop a line structure in frequency. This occurs both in the strong patches of activity echoing every 6.3 s, and in the intervals between where weaker signals, perhaps diffused cross-coupled echoes, exist. In other words, it is the Mini-DIAG transmission which creates (or becomes) the magnetospheric line emissions. This behavior has been seen before. *Helliwell and Katsufrakis* [1978, Fig. 5.6] show a case of multipath whistlers exciting trains of echoes that develop into lines. Two other examples caused by whistlers are found in *Park and Helliwell* [1981, Fig.s 4, 5].

The behavior of the magnetospheric lines over time can be seen in Figure 4.34. This is a compressed spectrogram which shows four minutes of the signals in Fig.s 4.32 and 4.33. Complex averaging with a time constant of $\tau_{avg} = 0.1$ s was used to decrease the bandwidth of the analysis filters.

During the first minute after the Mini-DIAG transmission, five lines with $\approx$ 45 Hz spacings develop from 2160 to 2340 Hz. These lines are apparent in the quiet space between the echoing emissions and are intensified in the echoes themselves. After the first minute, the lines between the echoes disappear but those within the echoes remain. As they echo, the emissions spread in frequency. At the end of Fig. 4.34, line structure extends from 1900 Hz to around 3000 Hz. Some of the lines above 2700 Hz may have been started by the later complex Siple pulses between 2750 and 3100 Hz.

During the first two minutes the echoes remain separated from each other by quiet intervals. Toward the end of Fig. 4.34 the echoes have spread out in time and begun to merge, though the 6.3 s periodicity is still clearly evident. These echoes continue throughout the rest of the record (to

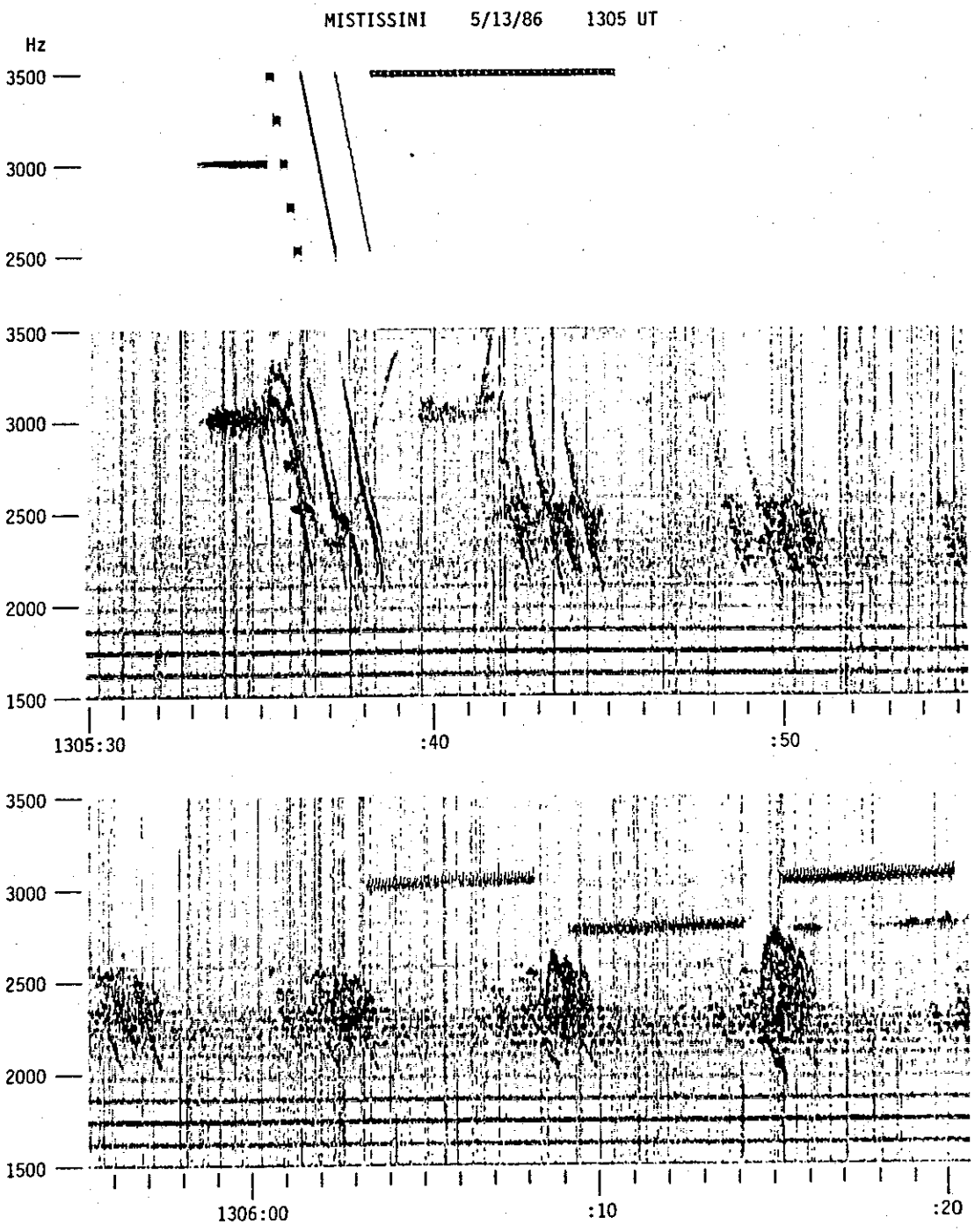

Figure 4.33. The development of the magnetospheric lines in Fig. 4.32. Pulses and decending ramps from a Siple Mini-DIAG transmission initiate a series of multipath echoes that spread in time, grow preferentially at discrete frequencies, and develop into lines.

1311), becoming stronger, more complex in form, and spreading slightly in frequency. By the end of the record the echoes look like typical mid-latitude chorus, except for their line structure.

Finally, notice that the lines in Fig. 4.34 gradually increase in frequency at about 25 Hz/min. This can best be seen by viewing the page edge-on. The frequencies of the local induction lines

A second plot was made tracking a magnetospheric line near 2610 Hz. This was done to check the mutual coherence of the lines, and see if there was any underlying structure such as a 60 Hz spacing otherwise corrupted by noise. No such structure was seen.

Summary. Magnetospheric line emissions do not seem to have any relation to power line harmonics, at least in the few cases studied. Magnetospheric lines do not fall on the frequencies of power line harmonics, nor do they have any constant offset from power line harmonics. The frequencies of a set of magnetospheric lines may drift slowly with time (usually upward), at rates of 25 Hz/minute or more. When lines form they do not start at power line harmonics, but develop over the course of several echoes during which time they may drift in frequency. When lines drift through the frequencies of power line harmonics they show no enhancement or entrainment at those frequencies.

The spacing of magnetospheric line emissions is not exactly 60 or 120 Hz (it's 45 Hz in Fig. 4.32, 63 Hz in Fig. 4.36), and may well change with time. The lines are noisy, with bandwidths of a few hertz, and their random variations in instantaneous frequency occur independently. Sometimes two or more sets of lines are intercalated in frequency. Intercalated lines are mutually frequency-locked, restricted in movement such that they never drift through one another, yet they have some freedom for independent fading and reappearance.

Generation Model of Magnetospheric Line Emissions. We saw above that there is no necessary connection between magnetospheric line emissions and power line harmonics. There may be such a connection in some cases, but we did not see it here. However, we do not need to invoke power line radiation into the magnetosphere to explain magnetospheric lines. I now describe a model which explains the behavior of magnetospheric lines, but is based only on magnetospheric processes:

We saw in Section 4.5.1 that a signal at a given frequency tends to suppress the growth of signals at nearby frequencies. This suppression is effective over a range approximately 10–100 Hz above and below the given signal, depending on conditions. Signals closer than 10 Hz are not suppressed, though they show modulation due to beating at their difference frequency. Signals separated by 100 Hz or more behave independently. Maximum suppression occurs for signals about 20 Hz apart, and becomes less at larger separations. A weak signal can suppress the temporal growth of another at least 20 dB stronger, though it may allow some steady-state amplification.

If we assume that the limit of suppression is, say, 60 Hz, then a signal (of whatever origin) in the magnetosphere will tend to grow only if there is no other signal within 60 Hz of itself. A small noise element that, by chance, is temporarily stronger than others nearby in frequency, will grow and suppress its neighbors, while allowing other noise elements 60 Hz or more removed to grow independently. Thus we may expect at any time that a sample of noise (at least if generated along a single path) will show a comb-like structure in frequency at any given time.

However, unless the state of the noise at a given time has some effect on noise that follows after it, a suppression mechanism that generates a comb structure in frequency will not be sufficient to generate a line structure that persists in time. That is, even though the spectrum at any given time shows a 60 Hz periodicity, combs will appear at different frequencies at different times. Such noise might appear quite random on a spectrogram.

If a noise comb generated at one time can control later noise generation, then we may expect to see spectral lines. We have two preconditions. First, we must have echoing so signal elements generated at one time come back to affect later events. Second, we need coupling between different echoing paths (even if amplification and suppression occurs on only one of them) so there will be a diffusion of signal elements in time. Echoing without cross-path coupling merely creates a two-hop

periodicity in the signal. With cross-path coupling, any signal whose frequency changes rapidly with time will suppress itself as components at different frequencies, following different delays on different paths, arrive together at the interaction region. The only signal structure that can persist under these conditions is one whose frequency changes very slowly with time. Thus the comb structure of amplification caused by frequency suppression (whatever its underlying mechanism) and the frequency stability caused by cross-path coupling of echoes together generate lines.

This model predicts that we will have magnetospheric lines only under the following conditions:

1. We must have multipath echoing.
2. We must have growth to make up for any path attenuation and keep signal amplitudes from decaying. However, growth need occur on only one of the echoing paths.
3. We must have suppression. A signal at a given frequency must be capable of suppressing the growth of nearby signals. The frequency range of suppression determines the line spacing of the emissions.
4. We must have cross-path coupling between echoing paths to provide dispersion in time.

The first condition, multipath echoing, probably always occurs when magnetospheric line emissions are seen. *Park* [1977] notes that line emissions (PLR, as he calls them) require good echoing. *Helliwell and Katsufrakis* [1978] present a case of multipath whistlers which echo and develop into lines. Multipath echoing was certainly the case in Figure 4.33 above. The second condition, growth, almost certainly occurs with line emissions as well. In many cases growth can be seen directly as chorus elements change form and trigger emissions with each echo. The third condition, suppression, has not been actively measured during line emission events. However, as far as is known, suppression always accompanies whistler-mode growth, so we may assume its presence when we see growth. The last condition, cross-path coupling, is least understood. *Smith and Carpenter* [1982] think it may be quite common, at least in the case of whistlers.

Using this model we can also account for intercalated sets of lines. The line sets are generated in separate regions, each with its own frequency-suppressed growth interaction and cross-path coupled echoes. If signals from one region are only weakly coupled into a second, then lines generated in the second region will not be completely suppressed by the first. However, they will be synchronized in frequency. Since suppression diminishes with increasing frequency separation, lines in the second region will be forced to be as far away in frequency from the first as possible, where suppression is a minimum. That is, lines in the second spectrum will lie midway between lines in the first. Depending on the strength of the coupling between the regions we may see some independent frequency drift between the two sets of lines, as seems to be the case. Triple sets of intercalated lines, as seen at 1234 in Fig. 4.35, must be due to three generation regions. A test of this model will be to check the direction of arrival of different sets of lines, or their strengths at different receivers, and show that different line sets are generated in different regions of the magnetosphere.

Future Search for Power Line Radiation. As the reader may have gathered, the evidence for effects caused by PLR is open to interpretation and is not universally accepted. I wonder whether or not it really "has been clearly demonstrated that PLR can trigger emissions that strongly interact with trapped energetic particles in the magnetosphere" [*Park and Helliwell*, 1980]. The most serious objection in my mind is that PLR itself has not been observed. Radiated signals surely exist at some level (since we do see harmonic currents on the ground); but if they are so weak they cannot be seen, the chain of evidence for proposed effects is weakened as well.

There are several problems in trying to observe PLR directly. It is very difficult, if not impossible, to observe with ground-level receivers in North America because we cannot distinguish between radiated signals such as PLR and the induction fields of local power line currents. If PLR is weak, it may well be below the detection threshold of satellite VLF receivers, and may still be susceptible to local power-line interference during telemetry reception and recording. Yet perhaps we could observe it in the Antarctic. Power generators at Antarctic stations are not as well regulated in frequency as the North American power grid, and the instantaneous frequency of local interference in Antarctic VLF recordings should be different from North-American PLR. With phase analysis it may be possible to separate whistler-mode PLR from local interference. At Siple Station the generators have occasionally been run at 58 Hz, making the task even easier. Future investigators must remember, however, that the station and transmitter generators are often run independently, and there may be two sets of local interference lines in Siple recordings.

4.7 Summary of the Characteristics of Whistler-Mode Growth

The following list summarizes various features seen during cyclotron-resonant growth, as observed mostly with man-made signals from the Siple Station VLF transmitter. Many of these features were discovered by earlier investigators using *f-t* spectrograms or narrowband filters to measure signal magnitudes, and are described in the literature. Some, those involving observations of the phase or instantaneous frequency of signals, are the products of phase analysis and are listed here for the first time. *These new contributions are shown in italic type.*

1. **Exponential Growth in Magnitude.** The amplitude of a growing pulse typically increases exponentially with time, at least initially. Growth rates up to 250 dB/s have been observed [*Stiles and Helliwell,* 1977]. Temporal growth usually continues for 20 to 35 dB (if the input signal lasts long enough), at which point amplitude saturation will occur.

All of the signals shown in this chapter have growth rates well below 250 dB/s. In fact, the most interesting signals seem to have growth rates on the order of 40 dB/s. This may be a selection effect. Rapidly-growing signals tend to more turbulent and chaotic behavior, and to occur in times of multipath propagation, both of which characteristics make them harder to interpret.

Growth rates are often only approximately exponential. The growth rate is often highest at the very beginning of a pulse, and may decrease well before saturation is approached. Pulses may show several episodes of faster and slower growth before saturation is reached.

2. **Advance in Relative Phase.** *The relative phase of a growing signal increases with time, indicating that the received signal is higher in frequency than when transmitted. Offset frequencies* $\Delta f = d\phi/dt$ *are typically in the range 1-8 Hz. Offset frequencies due to growth are an order of magnitude greater than the Doppler shifts due to duct motion, and change much more rapidly.*

Phase advance almost invariably accompanies growth. More than a momentary retardation in phase on a growing signal is never seen. (Phase lag may accompany diminishing signals, but this is a relatively uncommon phenomenon.) However, when growth rates are very low, the phase may not advance noticeably until the signal magnitude has grown by 10 dB or so. Phase advance is often parabolic with time (proportional to t^2*), especially at the beginning of a pulse, indicating a linearly-increasing frequency offset. Many pulses show a phase which increases only linearly with time toward the end (or until triggering occurs), evidence of a limiting value in offset frequency. Phase advance continues even after the amplitude has saturated.*

3. **Initial Frequency Offset.** *Sometimes the phase advance of growth seems to occur at a non-zero rate even at the beginning of a pulse, showing that the initial output frequency is above that of the*

input signal. For example, Fig. 4.3 shows a signal with an initial frequency offset of 1.1 Hz. Some of the largest initial offsets that have been observed (4.5 Hz) are shown in Fig. 4.4. Unfortunately, the amplitude at the beginning of a growing signal is weak and its phase plot is correspondingly noisy. It is difficult to say exactly how early a measurable frequency offset can occur. Some pulses show offsets within the first 30 ms, but whether the frequency offset is ever non-zero right at the start is unknown.

4. **Saturation.** After a pulse has grown by 20 to 35 dB, growth slows, stops, and the total signal amplitude remains at a roughly constant level.

5. **Magnitude Ripples at Saturation.** As a pulse approaches saturation and growth slows, regularly-spaced ripples in magnitude often develop, indicating the appearance of sidebands. *Associated ripples in phase also occur but are often less pronounced.*

6. **Band-Limited Impulse.** At the end of a pulse an amplitude transient often occurs, known as a band-limited impulse (BLI). Energy appears briefly in a short spheric-like event at frequencies up to 100 Hz above and below the pulse frequency. The BLI is usually asymmetric, with more energy above than below. A BLI may also appear at the beginning of a pre-termination emission, though such emissions are often so noisy that it is hard to distinguish an initial BLI from impulsive noise of the emission itself.

When a termination BLI is especially strong, it may appear as a brief interval of signal at a well-defined frequency above that of the growing pulse. *The transition in frequency from the pulse to the BLI may take place instantaneously (< 2.5 ms).*

7. **Termination Triggered Emission.** Following the BLI at the end of a transmitted pulse, the output signal is often found to continue in a self-sustaining magnetospheric oscillation or emission. The emission typically starts 30–100 Hz above the frequency of the input signal and may last for a second or two. Emissions are "risers" or "fallers" as they rise or fall in frequency after the end of the input signal. Termination emissions are usually fallers. Even a pulse exhibiting only very weak growth can trigger a faller. *Risers are only seen when* growth and *phase advance are well along by the end of the pulse. Even when a BLI is absent, the transition from the frequency of the growing signal to the offset frequency of the emission may be almost instantaneous, taking as little as a few milliseconds.*

The amplitude of a termination emission immediately after the end of the pulse is approximately the same as that of the growing pulse just before the end, though the emission may grow or decay rapidly from then on.

8. **Pre-Termination Triggered Emission.** *If the phase of a growing pulse has time to advance about two or three revolutions before the end of the input signal, a pre-termination emission will be generated.* These emissions are always risers; fallers are never generated in the middle of a growing signal.

Immediately after the emission is triggered, there may sometimes be a brief interval (say, 100 ms) during which the signal at the input frequency is suppressed below its initial value. It is not known how common this effect is. It is possible that it always occurs but is observed only when not masked by signals on other paths.

After the emission separates and any post-triggering suppression has passed, conditions are reset such that the signal begins to grow and *advance in phase* all over again, almost as if the input had just been turned on. *Subsequent periods of growth start from the same phase* and approximately the same amplitude *as that of the initial signal.*

9. **Spontaneous Sidebands.** The magnitude *and phase ripples* at saturation indicate sidebands—some of the pulse energy appears at one or more nearby frequencies from 10 (possibly less) to 100 Hz both above and below the transmitted signal. *When multiple sidebands are seen they are usually phase-coherent with each other—they are offset from the input signal by multiples of some common frequency and are harmonically related. When one sideband dominates, others at subharmonic frequencies may be seen.* Sidebands often occur at frequencies very close to 60 Hz, but whether this is due to magnetospheric signals such as power line radiation or to low-level hum modulation at the transmitter is undecided.

10. **Two-Tone Sidebands.** Sidebands are created when two equal-amplitude signals are transmitted with a separation from 10 to 50 Hz. That is, the received signal contains not only the two input components but equally-spaced sidebands offset by multiples of the input separation. *Each beat in a two-tone signal may act as a miniature pulse, showing growth and phase advance, and triggering a brief termination emission. However, the growing pulses remain phase-locked to the input such that coherent sidebands are generated.* Sidebands occur both above and below the input signals, though the ones above are usually stronger. They may be seen as much as 100 Hz from the input frequencies.

If two input tones are of unequal amplitude, the strongest sideband is usually the first component above the stronger input tone. If a two-tone input has a strong lower tone and weak upper one, the upper input frequency may paradoxically be the strongest component at the receiver. The relative amplitudes of different sidebands may change with time; *preliminary evidence suggests there may be small phase changes associated with this.*

11. **Suppression by Nearby Signals.** Two-tone signals with separations from 10 to 50 Hz suppress each other's growth and emission triggering. These are also the separations where two-tone sidebands are generated. Signals separated by less than 10 Hz behave as single tones, though with growth and emissions modulated at the beat frequency. The greatest suppression occurs with a separation of 20 Hz. *At slightly larger separations, some amplification may occur though components remain phase coherent. At separations of 100 Hz or more, two input signals behave independently, showing separate* growth and *phase advance.*

Completely suppressed signals show a slight (fractional-revolution) phase advance over the initial phase of a single-tone signal, evidence of amplification. As the amplitude of one component of a two-tone signal is decreased 20 dB with respect to the other, suppression weakens, all output components advance in phase by about 0.75 rev and increase slightly in magnitude, but they still remain phase-coherent.

12. **Change in Growth Activity with Time.** Growth rates, frequency offsets, and the strength of sidebands and emissions may change from pulse to pulse over the course of a few seconds or tens of seconds. Pre-termination emissions (always risers) and termination risers are common when activity is high, while termination fallers are more common at lower levels of activity. When activity is very high, signals may become very noisy and turbulent, probably because of an increase in the number of paths supporting growth.

When activity changes over the course of a few minutes such that both growing and non-growing pulses can be seen at a given frequency, the growing signals seem to start at the amplitude of the non-growing signals.

5. SUMMARY AND RECOMMENDATIONS

5.1 Summary

This thesis describes a new data analysis system which has been developed to measure the phase of recorded VLF signals, primarily whistler-mode signals from the VLF transmitter at Siple Station, Antarctica. Phase information has not been generally available in the past because ordinary spectrum analyzers cannot correct for timing and frequency errors in tape-recorded data. Knowing the phases of signal components as well as their magnitudes doubles the amount of spectral information. Phase information is not always easy to interpret; but when it is, as with coherent signals from VLF transmitters, we gain an entirely new perspective from which to view signal behavior. Some signal characteristics, such as the radial motions of whistler-mode ducts caused by electric fields in the magnetosphere, have been studied before but can now be measured more accurately with phase analysis. Some characteristics, such as the behavior of triggered emissions at the moment of separation, can be seen in greater detail than was possible with older methods. Many characteristics are seen here for the first time, such as the initial frequency offset of a growing pulse, the phase coherence of suppression, and the phase-locking of entrainment.

Method. In Chapter 2 we described the new analysis system and the algorithms used to process tape-recorded data, with an emphasis on those used to reconstruct signal phase information. A constant-frequency pilot tone is recorded along with the data in the field. We play back the field tape and digitize it in the laboratory. The phase of the pilot tone gives the data time at each sample, and the frequency of the pilot tone tells us the data rate (relative tape speed) at that point. We calculate the windowed discrete Fourier transform (DFT) of overlapping segments of sampled data. We correct for frequency shifts in the sampled data due to rate errors by linear interpolation between points in the DFT spectrum. This gives a new spectrum whose synthesized filters are at frequencies as originally recorded, ensuring that signal components are properly filtered. Linear interpolation has a small effect on the magnitude response of the synthesized filters, but has no effect on their phase response. Finally, if $\phi(t)$ is the phase at data time t of the output of a spectral filter whose passband is centered at frequency f_0, we calculate the relative phase of the filtered signal as $\phi_{rel}(t) = \phi(t) - 2\pi f_0 t$, the phase of the signal relative to a reference oscillator at frequency f_0. This corrects for tape timing errors and puts the phase information in an easily interpreted form.

Non-Growing Signals. In Chapter 3 we examined sub-ionospheric signals from VLF communications transmitters. Phase information allows us to identify the modulation and stability of different transmitter signals. A presumed Soviet transmitter at 19550/19600 Hz showed curiously poor phase stability.

We examined the phase of the sub-ionospheric signal from the Siple transmitter as received at South Pole Station and saw phase changes due to Trimpi events—the precipitation of energetic electrons into the ionosphere caused by whistlers. However, noise in the reconstructed phase of the best of signals is still around 1 μs rms. This is too much allow us to detect Trimpi events on lower-latitude paths, where perturbations are typically 1 μs or less. This noise is partially due to errors during analysis, and may be improved in future systems.

Whistler-mode signals from the Siple transmitter are sometimes received at the conjugate station in Roberval, Quebec, without any of the distortions caused by cyclotron-resonant growth. These

signals have the fidelity of sub-ionospheric signals, but reveal changes in the phase length of their magnetospheric path. Calculations show that some of this change is due to plasma flux between the ionosphere and magnetosphere, but that most must be from radial motion of the whistler-mode ducts caused by east-west electric fields mapped up into the magnetosphere. Phase measurements can supplement previous methods of measuring duct motion based on Doppler frequency shift or whistlers.

A continuous non-growing whistler-mode signal was studied, the eleven-minute two-tone LICO1 transmission of 9/2/83. Signal phase was measured every 0.25 s, corrected for errors due to fading, and compared to the H and D components of the earth's magnetic field measured near the receiver. Spectrograms of the phase delay and magnetic field showed strikingly similar features, with phase features preceeding magnetic features by 20 to 30 seconds. A cross-correlation plot between phase delay and D showed a bipolar signature which was interpreted as caused by relatively brief Pc 3 micropulsations traveling as Alfvén waves from equator to ground. Similar correlations have been seen before between Doppler shift and resonant-line micropulsations, but this is the first example where resolution has been good enough to see transient events, disturbances which occupy only a fraction of the field line length at any given time.

Growing Signals. The most interesting features of whistler-mode signals are those caused by cyclotron-resonance interactions between the wave and energetic electrons streaming in the opposite direction. These interactions cause temporal growth, the triggering of emissions, sidebands on two-tone signals, and other related phenomena. In Chapter 4, phase analysis uncovered several new growth-associated features, including the following:

1. The relative phase of a growing pulse advances with time, meaning that the frequency of the received signal is higher than when transmitted. The phase advance is often parabolic with time, indicating a linear increase in offset frequency. This phase behavior is a major constraint on models of wave-particle interactions. Sometimes growing pulses show a frequency offset of several hertz even at their beginning, a completely unexpected feature.
2. Termination emissions, those which occur at the end of a growing pulse, are found to begin at a frequency above that of the pulse, itself a few hertz above the input signal. This is true of fallers, emissions which fall in frequency, as well as risers. The change from the frequency of the growing pulse to that of the emission may occur very rapidly, in a few milliseconds or less. Phase analysis has let us see instantaneous frequency changes that are unobservable by other methods.
3. If the phase of a growing wave has time to advance sufficiently, a pre-termination emission, always a riser, will be triggered. Three revolutions seems to be the maximum phase wind-up which can be tolerated in the interaction region before this instability sets in. After each pre-termination emission separates, the signal is found to start regrowing with the same phase and approximately the same amplitude it had at the beginning; that is, with the presumed phase and amplitude of the input signal.
4. The beats between components in two-tone signals with separations of 10 to 50 Hz are found to behave like a series of isolated pulses, each pulse showing an increase in magnitude and a corresponding advance in relative phase. However, the initial phase of each growing beat remains locked to the input signal, such that sidebands in the output are harmonically-related and phase-coherent.
5. Multiple sidebands forming about single-frequency signals are often phase-coherent, with components at harmonics, and sometimes sub-harmonics, of the predominant offset frequency.

There seem to be frequent cases of such sidebands with offset frequencies very close to 60 Hz, the frequency of the North American power grid. However, whether these are due to the radiation of power line harmonics into the magnetosphere or to low-level hum modulation at the transmitter is still undecided.

6. Mutual growth suppression occurs when signals at nearby frequencies decrease each other's temporal growth and emission triggering rates. Suppression can now be monitored by its phase effects as well. The coherence bandwidth, the range of signal frequencies over which waves can resonate with the same energetic electrons, is found to be about 50 Hz, similar to previous estimates. The steady-state phase of each component in a suppressed two-tone signal is slightly advanced from the initial phase of a one-tone signal before growth, indirect evidence of the linear amplification of the two-tone signal. When one component in a two-tone signal is reduced by 20 dB, suppression weakens and all received components may show a constant phase advance of about 3/4 rev, yet remain phase-coherent.

7. Entrainment occurs when a weak whistler-mode signal changes the df/dt characteristic of a free-running emission and captures it. Examination of the entrainment of falling emissions by weak 50 ms idler pulses shows that the pulses gain control of the emission interaction region where it exists, well down-wave of the equator. Yet even here we see phase advance accompanying growth, and fallers with termination phase wrap-up, just as at the equator.

8. Whistler precursors are observed on a transmitted signal, the first time this has been seen. Phase analysis demonstrates the precursors are caused by a rapid increase in growth activity, and not merely the presence of a triggering signal as postulated by several previous studies.

9. Magnetospheric line emissions are found to be unconnected with possible power line signals. They are explained as due to natural magnetospheric growth, combined with the effects of suppression, multipath propagation, and coupling between different paths.

The phenomena described above must be only a sampling of those that phase analysis can uncover; I think we have just scratched the surface. I have analyzed very few records in which there was not something new and interesting. One of the purposes of this report has been to present as many examples of phase behavior as possible, since so little has been available. The other is to stimulate a general interest in phase analysis, and spur some reader to carry the process a little further.

5.2 Improvements to Field Station Equipment

Digital Recording in the Field. The biggest improvement we can make to data analysis will be to record broadband data digitally in the field. This will eliminate the tape wow and flutter problems that make phase analysis so difficult. In fact, future researchers will probably regard the techniques I've developed to cope with tape timing errors as being of historical interest only. In the past, digital recording entailed such a great penalty in tape cost over analog recording that it was prohibitive to use it except in limited cases. For instance, the 800 bpi digital tape in the digital analysis system holds 6.5 minutes of sampled data per reel, compared to 90 minutes on an analog tape for approximately the same cost. However, digital recording techniques have been advancing rapidly, and it will soon be feasible, if not cheaper and easier, to record digitally. Analog decks are also becoming harder to maintain as they reach obsolescence.

We have done some digital recording on an experimental basis already, using a PCM converter together with a video cassette tape recorder. This has reduced wow and flutter problems considerably, though it has not eliminated them. There is still some low-frequency wow with phase excursions around 30 μs due apparently to variations in tape speed. (The average speed is controlled quite accurately through a servo loop, but there are small variations about this average. On playback, data samples are converted as they are read instead of being buffered and converted at a constant rate.) This system, using 16-bit analog-to-digital conversion, has also given us recordings with the highest fidelity and largest dynamic range yet achieved.

I envision that systems in the near future will use DAT (Digital Audio Tape) machines. These record high-fidelity sampled data using helical-scan heads like the video cassette recorders, but have smaller tape cartridges. (They are not marketed in the U.S. at the moment because of unresolved concerns about copyright infringement, but should be available soon.) Optical disk technology is also advancing rapidly, and this may be an alternative recording medium. Both DAT and optical disks will encourage the development of desktop analysis systems since they will eliminate the tedious step of sampling and digitizing the analog tape.

Instrument to Phase-Track Two-Tone LICO1 Signals. The LICO1 signal analyzed in Sec. 3.4 showed how whistler-mode phase delay can be used to monitor magnetic field line disturbances at the equator as well as on the ground, where magnetic observations are now made. (Satellites can also measure the magnetic field at altitude, of course, but at $L = 4$ they don't allow lengthy observation at any one place.) This technique promises lots of exciting results. However, extracting the phase delay information in the LICO1 signal from the broadband sampled data was a great deal of work. Data sampled at 25600 samples/sec produced four measurements of phase delay every second. This is clearly an inefficient process.

This is one case where a special-purpose instrument in the field is better than a general-purpose analysis system. We need to develop an instrument that can track the phase of LICO1-like signals, correct for differential fading as in Eq. (3.12), and produce a low-rate output that can be sampled and recorded along with the magnetometer data at field stations. If the instrument also produced some measure of signal bandwidth, which increases with growth activity, we could monitor that as well. This need not be a complicated machine, and could be restricted to a particular carrier frequency and tone separation. Unless custom-made narrowband crystal filters can be used, sampling and processing by one of the new digital signal processing (DSP) chips (such as the TI TMS320 series) is probably the best approach.

We are currently developing a new generation of phase-measuring receivers to monitor VLF communications and navigation transmitters to detect Trimpi events. Measuring the phase effects

of Trimpi events is not much different from measuring the phase of whistler-mode LICO1 signals, and a similar design might work in both cases.

5.3 Improvements to the Analysis Algorithms

Using Proportional Clipping During Spectrum Normalizing. One of the easiest improvements to the analysis algorithms will be to introduce proportional clipping during the normalizing procedure in Sec. 2.5.7. Normalizing is performed on points in the interpolated spectrum to compensate for halving (scaling) during the FFT, filter gain changes with transform size and window order, and in order to set the overall processing gain. Normalizing is done independently on the real and imaginary values at each point. During normalizing, each point is clipped so its value cannot exceed ±32767 to prevent arithmetic overflow. Unfortunately, clipping real and imaginary points separately introduces phase distortion as well as amplitude distortion. Phase angles tend to concentrate around 45°, 135°, 225°, or 315° (those angles whose tangents are ±1).

We should really use proportional clipping, where the stronger component is clipped to ±32767 and the other component is reduced by the same factor so their ratio (and phase angle) remains constant. I think this will reduce the phase noise introduced by spherics, particularly when averaging.

Development of a Better Phase Estimator in the Pilot Tone Tracker. Some of the phase noise seen on a given signal is due to noise within the passband of the analysis filter for that component and cannot be avoided. Some of it, however, is due to noise in the filter used to track the pilot tone, since errors in measuring pilot tone phase become errors in data time, and hence errors in the relative phase of all signals. This is especially aggravating when the desired signal is strong and clean, yet the pilot tone is embedded in heavy spheric activity, or was recorded at too low a level. Most of the phase errors in this case will be pilot tone errors.

Yet, as we saw in Fig. 2.4, the phase of a typical pilot tone has a simple structure, with most of the variations (and perhaps almost all, if spherics did not interfere with our measurements) due to two periodic components: the rotations of the capstan and the supply idler. I think phase noise would improve markedly in many cases if we used an adaptive filter to estimate the pilot tone phase. Since the phases of only two periodic components at known frequencies (which are, however, different in different tapes) are involved, the filter would not have to be very complicated.

Equalizing the Anti-Aliasing Filter. A small source of phase error is due to distortion in the anti-aliasing filter as given by Eq. (2.10). Because the phase delay of the filter is not constant with frequency, tape speed variations lead to differential phase variations between the pilot tone and signal components. These errors are probably small in most cases, on the order of 1 μs or less. However, if we implement an adaptive filter to measure the pilot tone phase as proposed above, we may decrease other processing errors to the point where filter distortion becomes important, especially when looking for small phase changes as in Trimpi events. In this case it will be necessary to reduce the distortion caused by the filter. I think a very simple digital all-pass phase equalizer could decrease distortion to insignificant levels.

Ideal Frequency Interpolation for Synthesized Filters. As shown in Sec. 2.5.5, we use linear interpolation between points in the DFT spectrum to correct for frequency shifts due to tape speed errors, and also to place analysis filters at desirable frequencies. Given the discrete windowed spectrum $\{S_k\}$ and the current data rate $\bar{r}$, we calculate the interpolated spectral point U_j at data frequency f as

$$U_j = (1-p)S_k + pS_{k+1} \tag{2.43}$$

where k is an integer, and p an interpolation offset in the range $0 \le p < 1$, such that $f = (k+p)NT/\bar{r}$. Linear interpolation works well enough but has the disadvantage of changing the passband shape of the analysis filters slightly, depending on the value of p. In some future system where data are sampled in the field, we won't have to worry about unshifting the spectrum to correct for rate errors. However, it will still be useful to place analysis filters at arbitrary frequencies to match particular signal components. Is there some way of interpolating that preserves the filter shape?

The answer may be yes. We can calculate the value of the discrete windowed spectrum at an arbitrary frequency $f = (k+p)NT/\bar{r}$ as*

$$U_j = \sum_{m=0}^{N-1} S_m e^{-j2\pi(k+p-m)(1-1/N)} \frac{\sin[\pi(k+p-m)]}{N \sin[\pi(k+p-m)/N]}. \tag{5.1}$$

The value U_j now represents the output of a filter at frequency f whose passband is exactly the same as the filters in the original spectrum $\{S_k\}$ (including the effects of passband aliasing). The shape and center-frequency gain are now independent of interpolation offset p.

Equation (5.1) looks more complicated that it is. Remember that windowing reduces the sidelobe response of the DFT filters. Using the usual *3rd*-order window, we see in Fig. 2.6 that the main lobe of the filter is about $6/NT$ (six bins) wide, and all responses outside this lobe are at -72 dB or less. We only need to use the six terms $S_{k-2}, \ldots, S_{k+3}$ in Eq. (5.1) to capture all the signal in the main-lobe responses of the windowed filters. The error in interpolated filter response from omitting the other S_m terms is proportional to the signal in the sidelobes of the other windowed filters, and is too small to matter. In fact, using only four central terms might be more than adequate. The exponential phase and sine factors in Eq. (5.1) would have to be approximated in some fashion to simplify the calculation, of course.

Narrow-Band Analysis Using Translation and Decimation. One of the limitations of the present digital analysis system is that the FFT routine can transform at most $N = 2048$ real data points. With the usual 10.6 kHz sampled data, this means the minimum analysis filter bandwidth is 20 Hz with the *3rd*-order window ($1.6/NT$). We cannot process data with narrower filter bandwidths without reducing the sampling rate and the total analyzed bandwidth. We could play back the tape at, say, twice normal speed, digitizing only the bottom 5.3 kHz of the data, and then use 10 Hz filters. But this approach only works if both the signals being analyzed and the pilot tone are in the sampled bandwidth; that is, below 5.3 kHz. We can also fudge a bit by using complex averaging, but narrowing the filters in this way doesn't give a very sharp passband, and makes the filter impulse response distinctly asymmetrical in time (see Sec. 2.5.7).

The maximum transform size was fixed at $N = 2048$ because of the limited memory available in the system computer. It was not feasible to use larger data arrays and still fit everything into core at once. Memory is cheap in modern computers, and the next analysis system will surely allow much larger transforms and narrower filters. However, this is only a partial solution. In most cases when narrow filters would be useful, only a very small part of the total digitized bandwidth is going to be analyzed. Instead of synthesizing filters throughout the input bandwidth, it may be better to filter the sampled signal to reduce its bandwidth (by translation and low-pass filtering), decimate

* To show this, evaluate the discrete windowed spectrum $U_j = S_D(0, f)$ in Eq. (2.15) as a sum of terms $x_n w_n$, remembering that the window samples w_n are zero except for $n = 0, 1, \ldots, N-1$. Expand $x_n w_n$ as a sum in S_k from Eq. (2.41). Reverse the order of summation, and evaluate the sum of exponentials using Eq. (A.3).

it (reduce the number of samples), and then proceed with spectrum analysis. This would produce a zoom function, much like that used on the SD-350 analyzer. Filtering and decimation is a fairly straightforward process, but special care must be taken to measure and preserve the pilot tone rate and time information for later frequency and phase corrections. The time error could be corrected during translation, but frequency magnification to correct the rate error would still have to be done by spectral interpolation.

Chirp Z-Transform Algorithm for Resampling. An alternative method to correct for rate and time errors is through resampling—interpolating between data samples to construct a new sequence whose samples are uniformly spaced in data time rather than in laboratory time. This new sequence represents the samples that would have been taken if we had digitized signals in the field. All tape timing errors are removed, and we can filter, translate, and otherwise process the data without worrying about frequency and phase errors.

Resampling would also useful when very long data segments are analyzed (in order to make very narrow filters). Spectral filter interpolation to unshift data frequencies caused by rate errors only takes out the average error in a given segment. If the segment contains one or more cycles of tape flutter, there will be some residual frequency modulation in the data that will not be removed and will appear as low-level signal modulation at the flutter frequency. Resampling would allow us to reconstruct short segments of data, removing rate errors that change from one segment to another, and then concatenate these segments to make a larger data sequence free from flutter modulation.

Assume we have an original waveform $x(t)$. We record it, play it back, and sample it, hoping to obtain a series of samples $x_n = x(nT)$. However, because of rate and time errors we actually obtain a series of samples of the form

$$y_m = x(mrT - \xi). \tag{5.2}$$

We will assume that the data rate r and the time error ξ are approximately constant for some interval around $t = 0$, though they really vary from segment to segment. To reconstruct the desired samples x_n around $n = 0$ we interpolate as follows [*e.g.*, *Papoulis*, 1962, Eq. (3-77)]:

$$x_n = x(nT) = \sum_{m=-\infty}^{+\infty} y_m \frac{\sin[\pi(n/r + \xi/rT - m)]}{\pi(n/r + \xi/rT - m)}. \tag{5.3}$$

In practice, we would approximate x_n by

$$\hat{x}_n = \frac{1}{N} \sum_{k=-N/2}^{N/2} H_k Y_k A^k B^{nk} \tag{5.4}$$

where

$$H_k = \begin{cases} 1, & \text{if } -N/2 < k < N/2, \\ 1/2, & \text{if } k = \pm N/2, \end{cases}$$

$$Y_k = \sum_{m=0}^{N-1} y_m e^{-j2\pi mk/N},$$

$$A = e^{j2\pi\xi/NrT},$$

$$B = e^{j2\pi/Nr}.$$

This gives

$$\hat{x}_n = \sum_{m=0}^{N-1} y_m \frac{\sin[\pi(n/r + \xi/rT - m)]}{N \tan[\pi(n/r + \xi/rT - m)/N]} \tag{5.5}$$

$$\approx x_n.$$

The value $\hat{x}_n$ only approximates x_n because of, first, the use of a finite sum; and second, the approximation $\sin(x)/x \approx \sin(x)/N \tan(x/N)$. We must choose N as large as possible (yet smaller than any flutter period to be removed) to meet the first condition, and use only $\hat{x}_n$ for N near the center of the interval $[0, N-1]$ to meet the second. Equation (5.4) is best calculated using the chirp z-transform algorithm, as shown in Appendix C. This involves calculating three forward and one inverse Fourier transforms, which can be done using the FFT algorithm.

Resampling using the chirp z-transform was actually my first approach to correcting tape timing errors, before I settled on spectral interpolation. Unfortunately, it involved too much computation, given the power of the Eclipse S/230 computer, and took too long. With the advent of DSP chips that can perform very fast Fourier transforms, resampling may now be a viable approach.

The Modified Moving-Window Method of Spectrum Analysis. The standard method of spectrum analysis, where signal magnitudes are shown as functions of frequency and time by means of spectrograms, is known as the moving-window (or hopping-window) method. There is an extension of this method called the *modified* moving-window method, developed by *Kodera et al.* [1976, 1978]. The modified method improves the spectrogram by using signal phases to refine the positions in the *f-t* plane where specific signal elements are to be plotted. This method has not been used at Stanford in VLF studies, but might be valuable in some future system.

The standard moving-window spectrogram gives us a fuzzy picture of the underlying signal structure. For instance, in the spectrogram of an impulse we don't see an infinitely sharp vertical line but rather a vertical band, faint at the edges and darker in the middle, with a width equal to the duration of the weighting function $w(t)$. Similarly, the spectrogram of a constant-frequency tone of infinite duration is not a sharp horizontal line but a fuzzy line, densest at the proper frequency to be sure, but whose width reflects the passband width of the synthesized filters. (We're ignoring the additional uncertainties involved in sampling the *f-t* plane only on a lattice of points. For this discussion we assume the spectrogram is the product of an unlimited number of filters spaced arbitrarily closely in frequency whose outputs are measured and plotted arbitrarily often.) The limit to resolution is the uncertainty principle $\Delta t \cdot \Delta f \approx 1$; the risetime and bandwidth of an analysis filter define the size and shape of the area of achievable resolution in the *f-t* plane.

We saw in Section 1.3 that we could use signal phase information to measure the instantaneous frequency $\tilde{f}$ of a signal with an accuracy limited only by the signal-to-noise ratio, as long as we knew there were no competing signal components at nearby frequencies. We used this technique with VLF signals, plotting their phases to find their instantaneous frequencies. The modified moving-window method also uses this information, as follows: We look at the change in phase with time of the output of a filter, say at center frequency f_0. Suppose we measure an instantaneous frequency of f_a. With the modified method, instead of plotting the signal as a patch of gray proportional to its magnitude at frequency f_0 in the *f-t* plane, we plot it at the measured frequency f_a. Instead of plotting only on a lattice of points we can plot at the exact frequencies measured.

The modified moving-window method goes a step further and measures the *instantaneous time* of a signal component as

$$\tilde{t} = \frac{1}{2\pi} \frac{\partial \phi(t, f)}{\partial f} \tag{5.6}$$

in analogy to Eq. (1.8). This makes sense if we remember the discussion in Sec. 1.2 about how the time of occurrence of a signal element is translated into a winding up of the spectral phase as a function of frequency. Here we measure the helicity of the spectrum to find the time. (The calculation is actually made as a finite difference approximation between the outputs of adjacent

filters, like Eq. (1.9).) Now we can plot a signal element at both its precise time *and* frequency in the *f-t* plane, which are not necessarily the center time of the window or the center frequency of the filter.

The modified method seems promising for VLF studies. I think it will be especially useful when we need to know both the time and frequency of signal elements as accurately as possible, such as when using whistlers to measure the latitude and electron content of whistler-mode paths. It will not have as much to say in those cases where phase *per se,* rather than frequency, is important, as when identifying coherent signals.

Other Output Formats. The plotting routines need to be expanded to allow other types of output formats. Overlapping-trace A-scans would be a useful format, for instance. This would be similar to the current A-scan format, except that the step size from one trace to the next would be less than the maximum allowed displacement of the trace. This may cause successive traces to overlap at frequencies where signal intensity is changing rapidly. These regions can be left as is, or masking may be used to delete hidden lines. If the plots are made with frequency increasing to the right and time increasing downward, the format is called a waterfall display, and is very popular in vibration analysis circles. An overlapping-trace A-scan format was not implemented in the current system because calculating the output requires saving the last several spectra, and there was not enough memory available to do this. In a future system this will not be a problem.

A similar overlapping format would be useful with magnitude plots. In this case, the magnitudes of a set of filters at different frequencies are plotted in a series of traces as continuous functions of time, the converse of the A-scan plot. If the deflection of each trace was greater than their spacing in frequency, strong signals could overlap weaker ones from adjacent filters. Again, masking can be used to blank out hidden lines. The direction of deflection can also be tilted from the frequency axis to give the plot a three-dimensional look.

Ability to Analyze More Complicated Signals. One of the limitations of the present approach to phase analysis is that we can only easily interpret the phase of constant-frequency or slowly-varying signals. There is a need to be able to plot the relative phase of frequency ramps, say. In principle this should not be difficult. The rate of change df/dt of a transmitted ramp is known, and it is only necessary to use a phase reference signal with the same frequency characteristic, and interpolate analysis filters which change in frequency with time. Plotting the results will be a little more difficult. We could have a frequency scale which rotates through a range of values at the same rate as the ramp, in which case the phase of a ramp would plot as a horizontal line whose vertical position depends on the starting time of the ramp. Or, perhaps more sensibly, we could plot in a series of diagonal bands (with the df/dt slope of the ramp), so the phase of a coherent ramp would be a straight diagonal line. The phase of a linearly-propagating ramp should be a very sensitive measure of dispersion. We may also be able to separate close multipath components through phase ramps.

The next step after this will be to design a procedure to analyze the phase of whistlers. This might be similar to the phase-ramp technique, but use a reference frequency (and sweeping filters) whose rate of change follows a given whistler dispersion. This would allow us to make very accurate measurements of nose frequency and group (or phase) delay, and might show second-order dispersion effects caused by the length of the sub-ionospheric path or magnetospheric ions.

Another step, and I speculate a bit here, will be to develop an algorithm which changes the *f-t* plane into a nose-frequency group-delay or f_n-t_n plane. This algorithm would take a digitized waveform, and by calculating the autocorrelation of components at a given frequency *vs.* time

delay, and the dispersion in these correlations, could generate a plot where density at a given point is proportional to the signal intensity on a given whistler-mode path. I can imagine, at least conceptually, how this might be accomplished by a brute-force attack, but the important question is whether or not there is a fast algorithm for it, like the FFT algorithm which facilitates the calculation of *f-t* spectrograms. If such an algorithm could be devised (or even a slow one) we might answer some basic questions about the origin of chorus and hiss, and use them and other noisy signals for cold-plasma diagnostics much as we use whistlers. That is, since chorus contains echoing whistler-mode elements, we should be able to correlate successive echoes and determine the path(s) of the signal.

5.4 The Next Generation of Analysis Systems

Personal Workstation for Ease of Access. The current digital analysis system occupies two equipment racks, with a separate printer and terminal, in its own room in the ERL building at Stanford. Next door is the room containing our analog spectrum analyzers, which take up a dozen racks. To use either of these systems involves working in a building separate from the one where most of the VLF Group have their offices. It requires many hours of hands-on training. The actual analysis is tedious. To produce hard-copy plots from the analog analyzers requires wet-chemical processing of photographic paper or film, and must be left to the system operator. Faculty members rarely have the time to spectrum analyze data, even though they may be very interested in the results. There are several steps we can take to improve this situation.

Nearly everyone in the VLF group has a terminal in his office and regularly uses a computer, either a personal computer or one of the Space, Telecommunications, and Radioscience Laboratory's mainframe computers. The next logical step for data processing is to provide everyone who wants one his own spectrum analysis system. This should be a desk-top machine, a personal workstation. I think such a machine can be developed around a personal computer. The same system could also perform the other tasks personal computers are now used for, such as word processing.

Use of Pre-Digitized Data and Denser Media. One requirement for the personal analysis system will be a cheap and compact way to store the signals to be analyzed. Analog tape recorders are so large and expensive that it is not practical to have one for each workstation. The answer will be to have a central facility (perhaps operated in conjunction with our analog analyzers) where selected intervals of analog tapes are sampled and digitized. The sampled signals can be recorded on magnetic tape cartridges, DAT tapes, or optical disks. Each workstation will have a cartridge, DAT, or optical disk player, which are much smaller and (when they become consumer items in the near future) cheaper than analog recorders. Later on, when we record sampled data directly in the field, the digitizing problem will take care of itself and each investigator can have direct access to field recordings.

It will be important to use a medium capable of storing more data than the digital tape in the current system. A minimum requirement might be 100 megabytes per recording. This represents a bit over 40 minutes of 10.6 kHz data as presently used (12-bit samples at a rate of 25600 per second).

Hardware Tailored for Signal Analysis—DSP Chips. Another requirement for a workstation analysis system is a real-time analysis rate of at least a few kilohertz. That is, one second of, say, 3 kHz data should not take more than one second to process and display. Anything much slower becomes too tedious to use. Real-time analysis to 10 kHz would be a good goal.

This processing speed cannot be attained with current personal computers by themselves. For example, the IBM PC/AT is only slightly more powerful than the Data General Eclipse S/230

used in the present digital analysis system, whose real-time rate is about 1 kHz. However, there are several companies making boards that plug into personal computers, which use digital signal processing chips such as the Texas Instruments TMS320 series to perform filtering, calculate Fourier transforms, and so on. Using one of these boards we should get a real-time rate approaching 10 kHz, even for phase analysis.

Color Graphics for Increased Dynamic Range of Spectrograms. All the spectrum analysis we performed up to 1986 was in black and white. We are just starting to make color spectrograms. Color brings certain advantages, and color graphics should be part of any future system. One of the advantages of color, particularly in spectrograms, is that it allows more information to be displayed in a given area.

Black and white spectrograms rely on intensity modulation to show signal magnitudes as functions of time and frequency in the *f-t* plane. The magnitude output of an analog analyzer takes on continuous values over a range of about 50 dB. When this output is recorded on photographic film or paper, however, the range is reduced to around 20 dB. Any stronger signal appears uniformly black, and any weaker one is unseen. The magnitude of a signal in the present digital analysis system takes on discrete values over a somewhat larger range, perhaps up to 80 dB. However, the problem of hard copy is even worse than with the analog units. The *f-t* plane is divided up into fixed-size pixels, so many plot nibs high by so many wide. To show varying shades of gray involves turning on varying numbers of nibs in each pixel. Many of the figures in previous chapters show spectrograms with 5×3-size pixels, each containing 15 nibs. Only 16 different shades of gray can be shown (from white to solid black). Even this range is not uniformly divided—the change in intensity from 1 to 2 nibs turned on looks much larger than the change from 14 to 15.

Color overcomes these problems. Given a palette of sufficient size (that is, a sufficient number of available colors) we can display a very large dynamic range with fine resolution. Also, the color scale may look more linear than the gray-scale of varying-nib pixels. That is, small color changes at one end of the spectrum can be made to seem the same size as small changes at the other end. The choice of color graphics for an analysis system involves two variables, the resolution of the screen (the total number of viewable pixels) and the number of different colors available for each pixel.

Resolution need be only moderate to make good *f-t* spectrograms. The spectrograms in Fig. 4.34, for instance, plotted the outputs of 273 filters at 632 increments of time (using 4×3 pixels). The common IBM EGA color graphics standard uses a screen 350 pixels high by 640 wide, and would be more than adequate in this regard. Higher-resolution displays are also available. The other factor, the number of different colors, is another matter. The EGA standard allows 16 colors to be displayed at one time. This would be the absolute minimum palette useful for spectrograms. Other systems are available with 256 colors, and would be much better. Even larger palettes can be found, but more than 256 is probably unnecessary. Of course, besides a color graphics display, we need some way of making hard-copy output in color.

One other idea is worth mentioning. Variable-density black and white graphics can display a one-dimensional value in any given pixel—intensity. Color graphics can display a three-dimensional value. Our perception of color has three dimensions since color can vary in hue (wavelength or tint), saturation (bandwidth or purity), and intensity (brightness) at the same time. This presents the possibility of encoding both relative phase *and* magnitude in a spectrogram, say phase by hue and magnitude by intensity.

APPENDIX A.

DFT OF A COSINE WAVE

Section 2.5.3 uses an explicit expression for the discrete Fourier transform of a sequence $\{x_n\}$ representing samples of a cosine function. The expression is derived as follows:
Let

$$x_n = x(nT) = A\cos[\phi_0 + 2\pi f_0(nT - t_0)], \qquad \text{where } f_0 = \frac{k+q}{NT}, \quad \text{and } t_0 = \frac{NT}{2}. \tag{A.1}$$

We have

$$\begin{aligned} X_k &= \sum_{n=0}^{N-1} x_n e^{-j2\pi nk/N} = \sum_{n=0}^{N-1} A\cos[\phi_0 + 2\pi(k+q)(n/N - 1/2)]e^{-j2\pi nk/N} \\ &= \frac{A}{2}\sum_{n=0}^{N-1} e^{j[\phi_0 + 2\pi(k+q)(n/N-1/2)]}e^{-j2\pi nk/N} + \frac{A}{2}\sum_{n=0}^{N-1} e^{j[-\phi_0 - 2\pi(k+q)(n/N-1/2)]}e^{-j2\pi nk/N} \\ &= \frac{A}{2}\sum_{n=0}^{N-1} e^{j[\phi_0 - \pi k - \pi q + 2\pi nq/N]} + \frac{A}{2}\sum_{n=0}^{N-1} e^{j[-\phi_0 + \pi k + \pi q - 2\pi n(2k+q)/N]} \\ &= \frac{A}{2}e^{j[\phi_0 - \pi k - \pi q]}\sum_{n=0}^{N-1} e^{j2\pi nq/N} + \frac{A}{2}e^{j[-\phi_0 + \pi k + \pi q]}\sum_{n=0}^{N-1} e^{j2\pi n(2k+q)/N}. \end{aligned} \tag{A.2}$$

Now the sum of a geometric progression is given by

$$\sum_{n=0}^{N-1} z^n = \frac{z^N - 1}{z - 1}$$

from which we find

$$\sum_{n=0}^{N-1} e^{j2\pi na/N} = \frac{\sin(\pi a)}{\sin(\pi a/N)}e^{j\pi a(1-1/N)}. \tag{A.3}$$

Finally, using Eq. (A.3) in (A.2) we have the desired result

$$\begin{aligned} X_k &= \frac{A}{2}\frac{\sin(\pi q)}{\sin(\pi q/N)}e^{j[\phi_0 - \pi k - \pi q/N]} + \frac{A}{2}\frac{\sin(\pi q)}{\sin[\pi(2k+q)/N]}e^{j[-\phi_0 - \pi k + 2\pi k/N + \pi q/N]} \\ &= \frac{AN}{2}\frac{\sin(\pi q)}{N\sin(\pi q/N)}e^{j[\phi_0 - \pi k]}\left[e^{-j\pi q/N} + \frac{\sin(\pi q/N)}{\sin[\pi(2k+q)/N]}e^{j[-2\phi_0 + 2\pi k/N + \pi q/N]}\right]. \end{aligned} \tag{A.4}$$

We can also derive Eq. (A.4) from Eq. (2.16), evaluating $S_D(mT, f)$ at $mT = 0$ and $f = f_k = k/NT$, where the Fourier transform of $x(t)$ is given by

$$X(f) = \int_{-\infty}^{+\infty} x(t)e^{-j2\pi ft}\,dt = \frac{A}{2}[\delta(f + f_0) + \delta(f - f_0)]e^{j[\phi_0 f/f_0 - 2\pi f t_0]} \tag{A.5}$$

and where

$$W_D(f) = \frac{\sin(\pi fNT)}{\sin(\pi fT)}e^{-j\pi f(N-1)T} \tag{A.6}$$

is the discrete Fourier transform of the weighting sequence w_n given by Eq. (2.19).

APPENDIX B.

RATIO OF WHISTLER-MODE PHASE AND GROUP DELAYS

The refractive index μ, plasma frequency f_N, gyrofrequency f_H, and phase delay t_p are given by Equations (3.5)—(3.9), which are repeated here for convenience:

$$\mu = \left[1 + \frac{f_N^2}{f(f_H - f)}\right]^{1/2} \approx \frac{f_N}{f^{1/2}(f_H - f)^{1/2}} \tag{3.5}$$

$$f_N = \frac{e}{2\pi}\left[\frac{N}{\epsilon_0 m_e}\right]^{1/2} = 8.98 N^{1/2}\ [\text{Hz-m}^{3/2}] \tag{3.6}$$

$$f_H = \frac{Be}{2\pi m_e} = 2.80 \times 10^{10} B\ [\text{Hz/T}] \tag{3.7}$$

$$t_p = \int_S \frac{1}{v_p}\, ds = \frac{1}{c}\int_S \mu\, ds \approx \frac{1}{c}\int_S \frac{f_N}{f^{1/2}(f_H - f)^{1/2}}\, ds \tag{3.8}$$

The phase delay is the time it would take a given wavefront to propagate from the signal source to the receiver. For a continuous signal at a constant frequency it is the number of wavefronts of signal between the source and the receiver divided by the frequency. While we can see changes in phase delay at the receiver (as changes in relative signal phase), the total delay is not usually an observable quantity.

The group delay t_g is the time it takes signal energy to propagate from source to receiver. This is the time delay between the transmission and reception of a short pulse, for example, and *is* an observable quantity. The group refractive index μ_g is given by

$$\mu_g = \frac{d}{df}(f\mu) \approx \frac{f_H f_N}{2f^{1/2}(f_H - f)^{3/2}} \tag{B.1}$$

and the group delay t_g is then

$$t_g = \int_S \frac{1}{v_g}\, ds = \frac{1}{c}\int_S \mu_g\, ds \approx \frac{1}{2c}\int_S \frac{f_H f_N}{f^{1/2}(f_H - f)^{3/2}}\, ds, \tag{B.2}$$

where v_g is the *group velocity*. The phase velocity can be expressed in terms of the group velocity as

$$v_p = \frac{f_H}{2(f_H - f)} v_g. \tag{B.3}$$

Note that for frequencies well below the gyrofrequency the phase velocity is half the group velocity.* For signals well below the equatorial gyrofrequency f_{Heq} we will have $v_p = v_g/2$ everywhere in the

* Different waves show different types of dispersion. At the beach one may be disappointed to see a large wave seem to diminish as it approaches the shore, and then be surprised by the unexpected size of the one following it. Surface waves in water have a phase velocity which decreases with frequency and is greater than the group velocity. Individual wave crests move faster than (and peter out in front of) an approaching wave packet.

magnetosphere (since f_H is lowest at the equator), and we expect that the phase delay t_p will be twice the group delay t_g. At higher frequencies the difference is not as great. When the signal frequency is half the equatorial gyrofrequency (signals higher than this will not propagate in a whistler-mode duct) the ratio of phase velocity to group velocity will be as large as possible everywhere, and the phase delay will be closest to the group delay.

If we knew the relationship between t_p and t_g for an actual whistler-mode path then we might be able to estimate the phase delay even if it isn't directly observable. This Appendix will present the results of calculations of t_p and t_g derived using a standard model of the magnetosphere. We will find that the ratio of phase to group delay, t_p/t_g, for a signal at a frequency f, is a function only of the ratio f/f_{Heq} and is independent of the particular latitude or tube content of the path, at least for typical conditions inside the plasmapause.

The calculations presented here are a small extension of the work by *Park* [1972], who calculated the group delay t_g as a function of frequency for various path latitudes and electron densities in order to determine L and N_T from the nose whistler parameters f_n (nose frequency) and t_n (nose group delay). *Park* adopted a model for the electron density N as a function of distance along the path, and then integrated Eq. (B.2) numerically to find t_g. I have repeated his calculations and also integrated Eq. (3.8) to find t_p as well.

A dipole magnetic field is assumed, and the path S is assumed to follow a field line with specified McIlwain parameter $L = r_{eq}/r_0$, where r_{eq} is the radial distance to the top of the field line at the equator, and where $r_0 = 6370$ km is the radius of the earth. Delay times are calculated only for the magnetospheric part of the path, which is assumed to be from an altitude of 1000 km, or radial distance $r_1 = 7370$ km, in one hemisphere to a similar altitude in the opposite hemisphere. (Neglecting propagation through the ionosphere introduces a relatively small error). The integrations of Eq.s (3.8) and (B.2) are carried out in (r, ϕ) coördinates, where r is the radius and ϕ the latitude of a point along the path S. The equations can be written as

$$t_p = \frac{2}{c} \int_0^{\phi_1} \frac{f_N}{f^{1/2}(f_H - f)^{1/2}} \frac{ds}{d\phi}\, d\phi \tag{B.4}$$

and

$$t_g = \frac{1}{c} \int_0^{\phi_1} \frac{f_H f_N}{f^{1/2}(f_H - f)^{3/2}} \frac{ds}{d\phi}\, d\phi, \tag{B.5}$$

where

$$\phi_1 = \arccos\left(\sqrt{r_1/r_0 L}\right) \tag{B.6}$$

is the latitude of the end of the path at 1000 km altitude, and where

$$\frac{ds}{d\phi} = r_0 L \cos\phi\,(1 + 3\sin^2\phi)^{1/2}. \tag{B.7}$$

Note that we have used the approximation for large μ shown in Eq. (3.5). The gyrofrequency f_H at a point along the path is given by

$$f_H = f_{Heq}(r_0/r)^3(1 + 3\sin^2\phi)^{1/2}, \tag{B.8}$$

where

$$f_{Heq} = 8.736 \times 10^5/L^3 \text{ [Hz]} \tag{B.9}$$

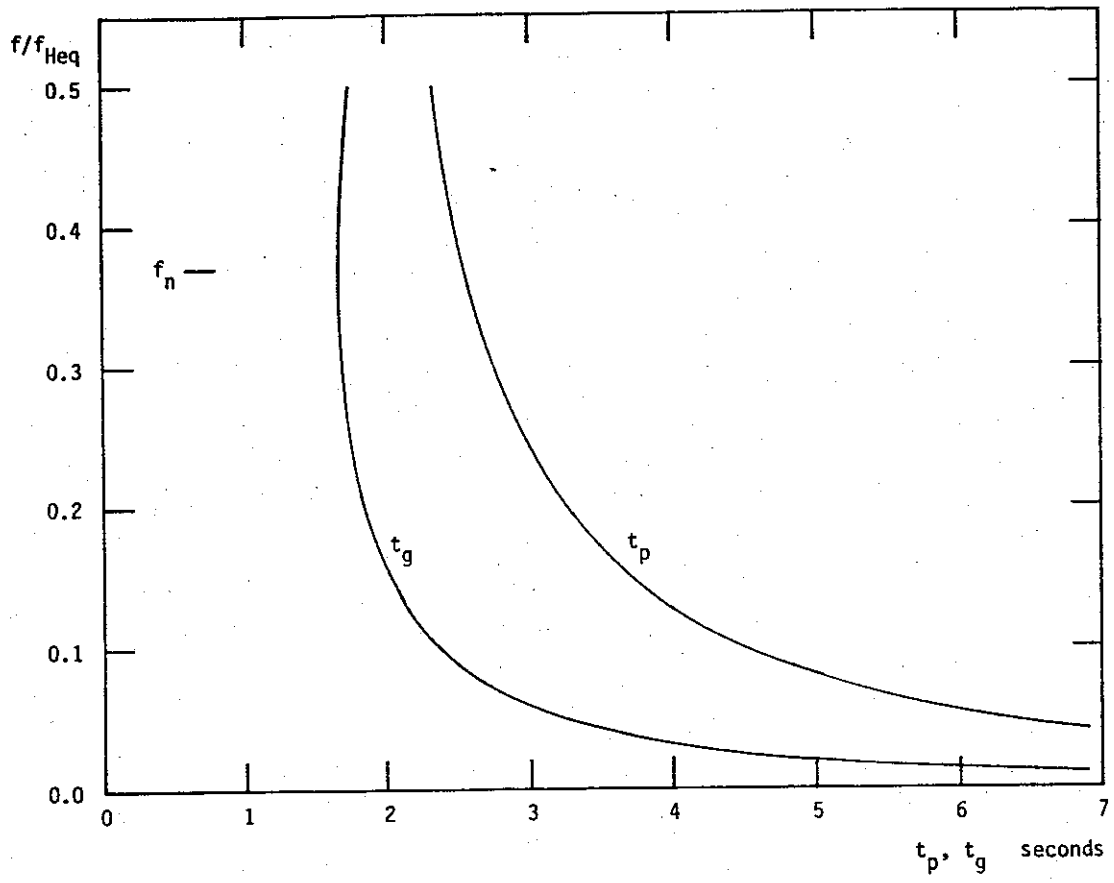

Figure B.1. Whistler-mode group delay t_g and phase delay t_p for a path at $L = 4$ (60° geomagnetic latitude) assuming an equatorial electron density of $N_{eq} = 300$ electrons/cm^2. The group delay (signal travel time) shows the typical whistler curve. The minimum group delay occurs at the whistler nose frequency $f_n = 0.37 f_{Heq}$, or at 5050 Hz for this path. The phase delay (wavefront travel time) is longer than the group delay at all frequencies up to half the equatorial gyrofrequency, the maximum frequency for a ducted signal.

is the equatorial gyrofrequency for the path, and where

$$r = r_0 L \cos^2 \phi \tag{B.10}$$

is the radius of a point on the path, given its latitude ϕ.

The plasma frequency f_N is found from the local electron density N using Eq. (3.6). The electron density model I have used is *Park's* DE-1 diffusive equilibrium model which was adapted from *Angerami* [1966]. This assumes a plasma with a uniform temperature $T = 1600$ °K and a composition at 1000 km altitude consisting of 90% O^+, 8% H^+, and 2% He^+. An equatorial electron density N_{eq} is specified, and the electron density along the path is then found from

$$N = N_{eq} \left[\frac{\sum_{i=1}^{3} \xi_i \exp(-z/H_i)}{\sum_{i=1}^{3} \xi_i \exp(-z_{eq}/H_i)} \right]^{1/2}, \tag{B.11}$$

where ξ_i is the relative concentration and H_i is the scale height of a given ion, and where

$$z = r_1 - \frac{r_1^2}{r} - \frac{\Omega^2}{2g_1}(r^2 \cos^2 \phi - r_1^2 \cos^2 \phi_1) \tag{B.12}$$

Carlson, C. R., *Simulation and Modeling of Whistler Mode Wave Growth Through Cyclotron Resonance with Energetic Electrons in the Magnetosphere,* PhD Thesis, Stanford University, Stanford, CA, 1987.

Carpenter, D. L., Ducted whistler-mode propagation in the magnetosphere; a half-gyrofrequency upper intensity cutoff and some associated wave growth phenomena, *J. Geophys. Res.,* **73**, 2919–2928, 1968.

Carpenter, D. L., U. S. Inan, E. W. Paschal, and A. J. Smith, A new VLF method for studying burst precipitation near the plasmapause, *J. Geophys. Res.,* **90**, 4383–4388, 1985. See also *J. Geophys. Res.,* **90**, 12340–12342, 1985, for improved reproductions of the figures in this article.

Chang, D. C. D., *VLF Wave-Wave Interaction Experiments in the Magnetosphere,* Tech. Rept. No. 3458-1, Radioscience Lab., Stanford Electronics Labs., Stanford University, Stanford, CA, 1978.

Chang, D. C. D., R. A. Helliwell, and T. F. Bell, Side-band mutual interactions in the magnetosphere, *J. Geophys. Res.,* **85**, 1703–1712, 1980.

Childers, D. G., ed., *Modern Spectrum Analysis,* Institute of Electrical and Electronics Engineers, Inc., New York, 1978.

Chilton, C. J., F. K. Steele, and R. B. Norton, Very-low-frequency phase observations of solar flare ionization in the *D* region of the ionosphere, *J. Geophys. Res.,* **68**, 5421–5435, 1963.

Cooley, J. W., and J. W. Tukey, An algorithm for the machine calculation of complex Fourier series, *Math. Comp.,* **19**, 297–301, 1965. Reprinted in *Rabiner and Rader* [1972].

Cooley, J. W., Lewis, P. A. W., and P. D. Welch, Historical notes on the fast Fourier transform, *IEEE Trans. Audio Electroacoust.,* **AU-15**, 76–79, 1967. Reprinted in *Rabiner and Rader* [1972].

Cooley, J. W., Lewis, P. A. W., and P. D. Welch, The fast Fourier transform algorithm: Programming considerations in the calculation of sine, cosine, and Laplace transforms, *J. Sound Vib.,* **12**, 315–337, 1970. Reprinted in *Rabiner and Rader* [1972].

Cousins, M. D., *A Computer Program for Dynamic Spectrum Analysis Applied to VLF Radio Phenomena with Examples and Comparisons to Rayspan Results,* Tech. Rept. No. 3432-1, Radioscience Lab., Stanford Electronics Labs., Stanford University, Stanford, CA, 1971.

de Buda, R., Coherent demodulation of frequency-shift keying with low deviation ratio, *IEEE Trans. Communications,* **COM-20**, 429–435, 1972.

Doelz, M. L, and E. T. Heald, Minimum-shift data communication system, US Patent 2977417, 1961.

Dowden, R. L., Trigger delay in whistler precursors, *J. Geophys. Res.,* **77**, 695–699, 1972.

Dowden, R. L, A. D. McKay, L. E. S. Amon, H. C. Koons, and M. H. Dazey, Linear and nonlinear amplification in the magnetosphere during a 6.6 kHz transmission, *J. Geophys. Res.,* **83**, 169–181, 1978.

Ferrell, O. P., *Guide to RTTY Frequencies,* 2nd ed., Gilfer Associates, Park Ridge, NJ, 1983.

Gallet, R. M., and R. A. Helliwell, Origin of very low frequency emissions, *J. Res. NBS,* **63D**, 21–27, 1959.

Gold, B., and C. M. Rader, *Digital Processing of Signals*, McGraw-Hill, Inc., New York, 1969.

Harris, F. J., On the use of windows for harmonic analysis with the discrete Fourier transform, *Proc. IEEE*, **66**, 51–83, 1978.

Hart, J. F., E. W. Cheney, C. L. Lawson, H. J. Maehly, C. K. Mesztenyi, J. R. Rice, H. G. Thacher, Jr., and C. Witzgall, *Computer Approximations*, Robert E. Krieger Publishing Company, New York, 1978.

Helliwell, R. A., *Whistlers and Related Ionospheric Phenomena*, Stanford University Press, Stanford, CA, 1965.

Helliwell, R. A., A theory of discrete VLF emissions from the magnetosphere, *J. Geophys. Res.*, **72**, 4773–4790, 1967.

Helliwell, R. A., Controlled VLF wave injection experiments in the magnetosphere, *Space Sci. Rev.*, **15**, 781–802, 1974.

Helliwell, R. A., Coherent VLF waves in the magnetosphere, *Phil. Trans. R. Soc. Lond.*, A. **280**, 137–149, 1975.

Helliwell, R. A., Effects of power line radiation into the magnetosphere, in *Wave Instabilities in Space Plasmas*, edited by P. J. Palmadesso and K. Papadopoulos, pp. 163–190, D. Reidel, Dordrecht, Holland, 27–36, 1979a.

Helliwell, R. A., Siple Station experiments on wave-particle interactions in the magnetosphere, in *Wave Instabilities in Space Plasmas*, edited by P. J. Palmadesso and K. Papadopoulos, pp. 163–190, D. Reidel, Dordrecht, Holland, 191–203, 1979b.

Helliwell, R. A., VLF wave injections from the ground, in *Active Experiments in Space, Symposium at Alpbach 24–28 May, 1983*, ESA SP-195, 3–9, 1983a.

Helliwell, R. A., Controlled stimulation of VLF emissions from Siple Station, Antarctica, *Radio Sci.*, **18**, 801–814, 1983b.

Helliwell, R. A., and E. Gehrels, Observations of magneto-ionic duct propagation using man-made signals of very low frequency, *Proc. IRE*, **46**, 785–787, 1958.

Helliwell, R. A., and U. S. Inan, VLF wave growth and discrete emission triggering in the magnetosphere: A feedback model, *J. Geophys. Res.*, **87**, 3537–3550, 1982.

Helliwell, R. A., and J. P. Katsufrakis, VLF wave injection into the magnetosphere from Siple Station, Antarctica, *J. Geophys. Res.*, **79**, 2511–2518, 1974.

Helliwell, R. A., and J. P. Katsufrakis, Controlled wave-particle interaction experiments, in *Upper Atmosphere Research in Antarctica, Antarctic Res. Ser.* **29**, edited by L. J. Lanzerotti and C. G. Park, pp. 100–129, AGU, Washington, DC, 1978.

Helliwell, R. A., J. Katsufrakis, M. Trimpi, and N. Brice, Artificially stimulated very-low-frequency radiation from the ionosphere, *J. Geophys. Res.*, **69**, 2391–2394, 1964.

Helliwell, R. A., J. P. Katsufrakis, and M. L. Trimpi, Whistler-induced amplitude perturbation in VLF propagation, *J. Geophys. Res.*, **78**, 4679–4688, 1973.

Tkalcevic, S., *Nonlinear Longitudinal Resonance Interaction of Energetic Charged Particles and VLF Waves in the Magnetosphere*, Tech. Rept. No. E4-2311, Radioscience Lab., Stanford Electronics Labs., Stanford University, Stanford, CA, 1982.

Tsurutani, B. T., and R. M. Thorne, letter, *Science*, **207**, 716, 1980.

Tsurutani, B. T., S. R. Church, and R. M. Thorne, A search for geographic control of the occurrence of magnetospheric ELF emissions, *J. Geophys. Res.*, **84**, 4116–4124, 1979a.

Tsurutani, B. T., E. J. Smith, S. R. Church, R. M. Thorne, and R. E. Holzer, Does ELF chorus show evidence of power line stimulation?, in *Wave Instabilities in Space Plasmas*, edited by P. J. Palmadesso and K. Papadopoulos, pp. 163–190, D. Reidel, Dordrecht, Holland, 51–54, 1979b.

Welch, P. D., The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms, *IEEE Trans. Audio Electroacoust.*, **AU-15**, 70–73, 1967. Reprinted in *Rabiner and Rader* [1972] and in *Childers* [1978].

Welch, P. D., A fixed-point fast Fourier transform error analysis, *IEEE Trans. Audio Electroacoust.*, **AU-17**, 151–157, 1969. Reprinted in *Rabiner and Rader* [1972].

Yearby, K. H., A. J. Smith, T. R. Kaiser, and K. Bullough, Power line harmonic radiation in Newfoundland, *J. Atmos. Terr. Phys.*, **45**, 409–419, 1983.

Zverev, A. I., *Handbook of Filter Synthesis*, John Wiley and Sons, Inc., New York, 1967.