A TRACKING PHASE VOCODER

AND ITS USE IN THE ANALYSIS OF ENSEMBLE SOUNDS

Thesis by

Mark Barry Dolson

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1983

(Submitted November 12, 1982)

ACKNOWLEDGEMENTS

In the course of this research, I have been influenced by more people than I can ever hope to properly acknowledge. Foremost among these have been Dr. John Pierce, without whom this work would never have been begun; and Dr. Edward Posner, without whom it would never have been completed. Also instrumental has been James Boyk, whose helpful advice has extended far beyond the merely musical. These three, I thank in particular for their encouragement and patience.

In addition, I have benefited greatly from discussions with numerous friends. Glenn Greene and Andy Pesthy provided countless valuable suggestions, while Jim Garvey, Jack Kaye, Paul Milenkovic, and Howard Yee each made important contributions as well. Rob Drezner also proved to be an inexhaustible source of valuable insights. To all of the above, I extend my deepest gratitude.

This research was made possible by the financial support of the Hughes Aircraft Company and by the personal support of the individuals there with whom I worked; crucial support was also provided by Walter Victor of the Jet Propulsion Laboratory. I am most grateful to all of these people. Lastly, I thank those musicians who gave so generously of their time in recording and listening sessions.

ABSTRACT

Additive analysis-synthesis using the phase vocoder is a powerful tool for the exploration of musical timbre. In this research, previous investigations of this subject are extended in two significant directions.

First, an improved analysis of the phase vocoder is developed to explain the errors introduced by undersampling and modification of the magnitude and phase-derivative signals. Two sources of error are identified. It is shown that the first of these involves crosstalk between adjacent frequency channels, and can be eliminated through the development of a tracking version of the phase vocoder. Alternatively, restrictions can be placed on the phase-derivative signal to preserve the absolute phase. The second source of error appears to be inherent in the phase vocoder formulation.

Secondly, the tracking phase vocoder is used to investigate differences between solo and ensemble sounds. A search is conducted for the minimal set of cues which will produce an ensemble sensation. It is shown that the primary requirement is that there be at least four to eight harmonics, each of which has a characteristic amplitude modulation proportional to its frequency. In addition, a number of issues related to the quality of the ensemble sensation and its efficient synthesis are examined.

A TRACKING PHASE VOCODER
AND ITS USE IN THE ANALYSIS OF ENSEMBLE SOUNDS

I. INTRODUCTION

II. ADDITIVE ANALYSIS-SYNTHESIS

III. THE PHASE VOCODER

IV. ADDITIVE ANALYSIS-SYNTHESIS OF ENSEMBLE SOUNDS

I. INTRODUCTION

The application of modern digital signal processing techniques to the analysis of musical sound has proven to be extremely fruitful. These methods have been especially useful in studying timbre - that aspect of musical sound which enables us to discriminate between two different instruments playing the same pitch at the same subjective loudness. Recent work at Stanford's Center for Computer Research in Music and Acoustics has greatly enhanced our understanding of this phenomenon and has led directly to several practical applications. This investigation significantly extends that work by improving the analytical techniques and applying them to more complex sounds. In particular, we consider the differences between a solo violin and a violin section playing in unison. We begin with a brief introduction to the field and an overview of the chapters to follow.

1.1 Background

The analysis of musical sound has a considerable history. The pioneering contributions of Helmholtz were first published in 1863. The following century, however, brought only sporadic progress. This was due almost entirely to the limitations of the available technololgy. In the 1960's, the widespread introduction of digital processing techniques essentially eliminated this restriction and revolutionized the field.

Today, investigations of musical sound run the gamut from simple experiments with a digital spectrum analyzer to large scale research projects with a dedicated computer. The research itself, however, can be divided into three distinct areas: physical acoustics, psychoacoustics, and digital signal processing. In principle, these areas have much in common; but in practice, they are surprisingly isolated. We will discuss each of them briefly.

Researchers in physical acoustics are primarily concerned with the physics of musical instruments and concert halls. They recognize sound as a three-dimensional wave phenomenon and analyze it accordingly. This is a marked contrast to the other two areas in which the one-dimensional amplitude-versus-time signal of a microphone or speaker is considered to be adequate. In this report we consider only one-dimensional signals, but we refer to physical acoustical experiments where appropriate.

Psychoacoustics deals with the perceptual effects of carefully defined test sounds. Early workers in this area made important discoveries about the nature of hearing by observing the response to sine wave and white noise input signals. More recent investigations, however, have been considerably less fruitful. To a large extent, this simply reflects the limitations of the psychoacoustical approach. Unfortunately, these limitations have been so widely ignored that they are well worth stating explicitly.

Psychoacoustics represents an engineering approach to the problem of sound analysis. It is very good at determining classes of acoustic signals which are perceptually equivalent, but it is very poor at discovering the internal processing scheme which explains this equivalence. That information must ultimately come from neurophysiological studies. Meanwhile, there is little value to psychoacoustical experiments which attempt to deduce this processing scheme on the basis of responses to highly artificial test sounds.

There is also a question about the usefulness of such artificial test signals in the investigation of timbre. Timbre is a highly multidimensional phenomenon in which the important factors depend strongly on the particular sounds being compared. A far more promising approach to the analysis of timbre would be to start with real musical sounds and systematically eliminate those features which are not perceptually significant. Indeed, this latter approach is the one taken in this investigation.

The primary focus of this study, though, is on digital signal processing. This is by far the most recent, and probably the most active, branch of musical sound analysis. Much of this interest stems from the virtually unlimited potential of the computer itself as a musical instrument. To exploit this capability, however, we first must determine what kinds of waveforms produce a given perceptual effect. Hence, the computer provides not only a powerful means of timbre investigation, but also a powerful motivation for it.

Unfortunately, digital processing of musical sounds to date has been nearly as much an art as a science. The ubiquity of digital equipment has attracted a great many investigators, but has encouraged very little rigor. In fact, much work in this area has been reported only by word of mouth. In this report, we take considerable pains to establish a clear and unified framework in which this research can be usefully conducted. Indeed, we view this as one of the primary contributions of this study.

Most of the useful techniques for digital processing of musical sound have come from the field of digital speech processing. This is not surprising since both these fields are ultimately concerned with the way sound is perceived. However, these fields share two more immediate goals as well: efficient encoding of signals, and easy modification of signals. Each of these is sufficiently important to merit further discussion.

Digitizing a typical speech waveform results in 80,000 bits of data per second; for music, the result is 800,000 bits per second. In both cases, the perceptually significant information can be transmitted at one percent of this data rate. Since bandwidth and memory are expensive commodities, methods for accomplishing this data compression are of great interest.

It is also very useful to be able to change one perceptual feature of a music or speech signal without affecting the others. In particular, much effort is devoted to separating pitch from the other temporal aspects of a given sound. The ear performs this operation automatically, but accomplishing it analytically remains difficult.

By far the most successful technique for attaining these goals has been that of analysis-synthesis. The fundamental assumption of this approach is that the signal can be well represented by a model whose parameters are varying with time. The analysis is devoted to determining the values of these parameters, while the synthesis is simply the output of the model itself. However, the success of this method depends very much on the appropriateness of the model. In computer music applications, three distinct classes of models have been found to be useful: additive models, subtractive models, and nonlinear models. We will discuss each of these briefly.

Additive analysis-synthesis attempts to represent the signal as a sum of sine waves; the instantaneous amplitudes and frequencies of these sine waves are the parameters to be estimated. This is a very computation intensive approach, but it is capable of truly impressive fidelity. In our view, this simply reflects the excellent match between the additive model and the human hearing system. This fidelity makes additive analysis-synthesis the method of choice for investigating timbre.

Subtractive analysis-synthesis takes a nearly opposite approach. The signal is modeled as a pulse train which passes through a time varying filter. This amounts to modeling the process by which the sound is produced as opposed to the process by which it is perceived. In the case of speech signals, the pulse train amplitude and frequency and the filter coefficients are usually estimated by the technique of linear prediction. This method offers tremendous flexibility in modifying and resynthesizing the sound, but at the cost of reduced fidelity. There have been several outstanding attempts to apply it to musical signals [Petersen, 1976; Moorer, 1979; Lansky and Steiglitz, 1981], but its potential lies more in the creation of novel musical effects than in timbral analysis.

Nonlinear models, too, are more important for sound generation than for sound analysis. The most notable of these is the FM representation of Chowning [1973]. The basis of this technique is a summation formula which expresses a sine wave with sinusoidal frequency modulation as a sum of harmonic sine waves; the relative amplitudes of each harmonic depend on the modulation index. A number of related methods have also been proposed [Moorer, 1976; Saunders, 1977], all featuring extremely efficient synthesis with no consideration of analysis. Indeed, it is only recently that an analysis-synthesis interpretation has even been developed for these models [LeBrun, 1979; Justice, 1979].

In this report we deal exclusively with additive analysis-synthesis. The first successful application of this technique to the study of timbre was in the work of Fletcher [1962; 1963; 1965; 1967] which led to a number of important discoveries. This is all the more impressive today in that Fletcher depended entirely on analog equipment. Early digital implementations of this approach were developed by Luce [1963], Freedman [1965], Risset [1966], Beauchamp [1969], and Keeler [1972]. However, the current state of the art was defined by Moorer and Grey in a landmark series of investigations [Moorer, 1975; Grey, 1975; Grey and Moorer, 1977; Grey, 1978; Grey and Gordon, 1978; Moorer, 1978].

It is their work which constitutes the starting point for the reseach to be reported here.

1.2 Overview

We begin in Chapter Two with a careful introduction to additive analysis-synthesis. We also review the work of Moorer and Grey and discuss its implications. This provides the necessary background for the chapters to follow.

In Chapter Three, we investigate the use of the phase vocoder for additive analysis-synthesis. In particular, we seek to understand those errors which arise when the phase vocoder magnitude and phase-derivative signals are undersampled or modified prior to resynthesis. To this end, we consider the relation between these signals and the parameters of the additive model; this leads to the identification of two independent sources of error. We show that the first of these involves crosstalk between adjacent frequency channels and can be eliminated through the development of a tracking version of the phase vocoder. Alternatively, restrictions can be placed on the phase-derivative signal to preserve the absolute phase. The second source of error appears to be inherent in the phase vocoder formulation.

In Chapter Four, we use the tracking phase vocoder to investigate differences between solo and ensemble sounds. In particular, we seek to identify those cues which are sufficient to produce an ensemble sensation. We show that the primary requirement is that there be at least four to eight harmonics, each of which has a characteristic amplitude modulation proportional to its frequency. In addition to this, we examine a number of issues related to the quality of the ensemble sensation. Lastly, we consider the implications of these results for the efficient synthesis of ensemble sounds from solos.

II. ADDITIVE ANALYSIS-SYNTHESIS

Additive analysis-synthesis has proven to be a powerful tool for the investigation of timbre. In this chapter we present a careful introduction to the additive analysis-synthesis technique. We begin by examining the connection between the additive model and both physical and psychological acoustics. We then review the work of Moorer and Grey upon which our own research has been based. Finally, we consider some of the more recent developments in this field.

2.1 Physical acoustics and psychoacoustics

A violinist produces sound by drawing a horsehair bow across a string. This induces oscillations of the string which are transmitted to the violin body and to the surrounding air. These oscillations propagate throughout the room as a mechanical wave disturbance in which pressure and particle velocity vary periodically. Thus, a three-dimensional sound field is created with the property that the displacement-versus-time at any point is related to that of the string. However, the displacement waveform in the sound field also includes the effects of filtering by the violin body and of reflections from surfaces within the room.

A listener perceives sound by sampling this three-dimensional sound field at two distinct points. In each ear, pressure variations are channelled through the auditory canal to the tympanic membrane or *eardrum* (Figure 1). Oscillations of this membrane are transmitted to the oval window of the cochlea through a series of three small bones which provide impedance matching. It is in the cochlea that the actual transduction of sound to neural impulses is accomplished.

The cochlea is a snail-shaped structure consisting of a coiled tube filled with saline solution (Figure 2). Running the length of this tube through the

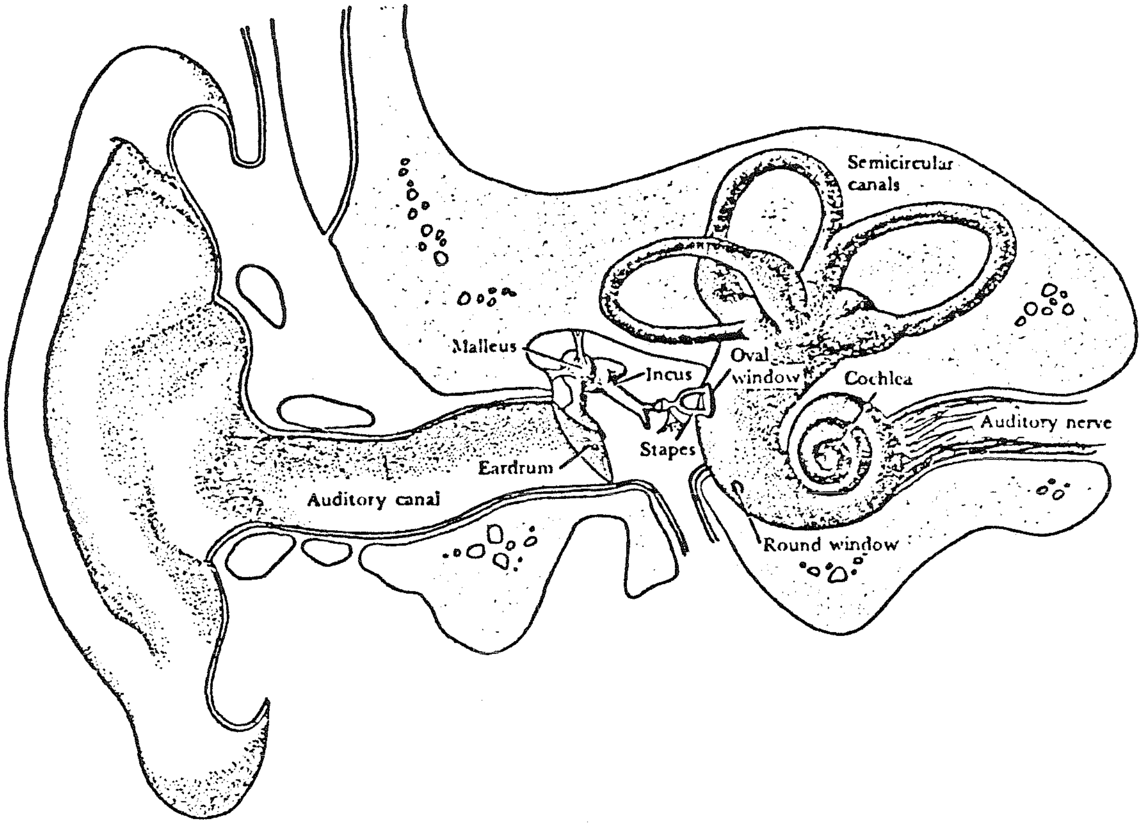


Figure 1. Anatomy of the peripheral auditory system.
(from Lindsay and Norman, 1972)

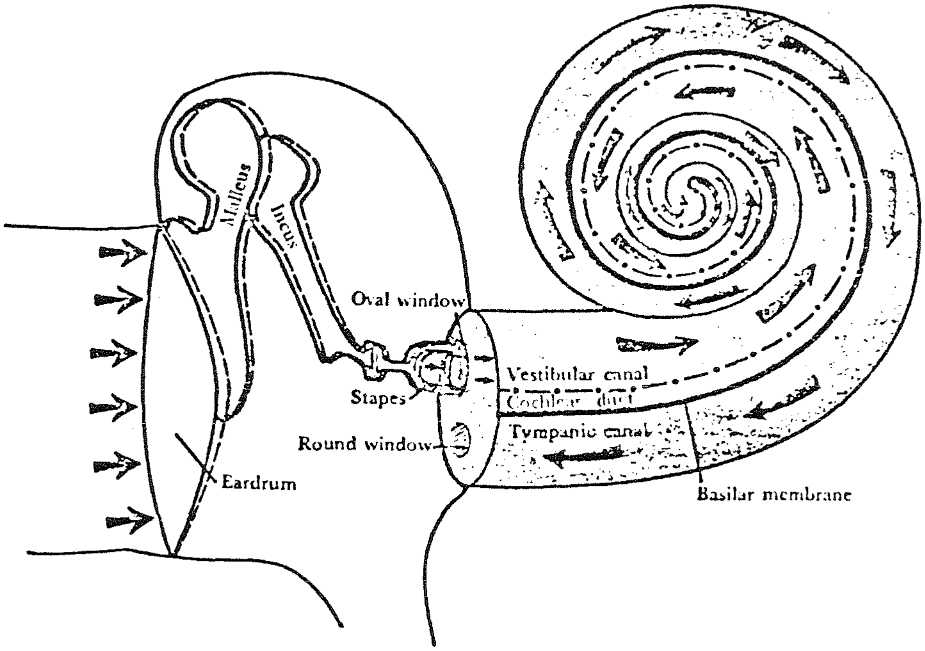


Figure 2. A schematic view of the cochlea.
(from Lindsay and Norman, 1972)

center is the basilar membrane, a thin sheet with a dense array of motion-detecting hair cells. Audible sounds produce pressure variations at the oval window which propagate through the fluid of the cochlea to excite traveling waves on the basilar membrane. The resulting displacements are translated by the hair cells into the neural impulses which form the input to the brain.

A crucial feature of this system is that the width and stiffness of the basilar membrane vary considerably over the length of the cochlea. Consequently, a given traveling wave typically exhibits a maximum amplitude over only one small region of the membrane. In particular, if the input sound is a pure sinusoid, then the location of maximal displacement on the membrane is determined solely by the frequency of the input. A high frequency tone produces a peak excitation near the oval window while lower frequency tones result in maximal displacements proportionally farther along the membrane. Hence, the cochlea and basilar membrane effectively perform a Fourier transform of the incoming sound.

Another important and closely related feature of this system has emerged from psychoacoustic investigations. The hair cells are distributed along the basilar membrane such that there is a roughly constant number covering any 1/3 octave region of frequency space along the membrane. Two pure tones whose frequencies lie within this *critical bandwidth* are heard as a single tone of complex timbre; more widely separated tones are heard distinctly. This is a statement about the resolution of the Fourier transform.

Further psychoacoustic investigations have disclosed yet another interesting aspect of this system. It seems that the brain devotes considerable attention to the magnitude portion of this Fourier transform while largely ignoring the phase. This is shown by numerous experiments in which the relative phasing of harmonics is varied with only a subtle effect on the perceived sound. Additional information of this type can be found in any standard acoustics text (eg. Backus, 1969).

2.2 Sound recording and reproduction

Audible sound can be described entirely by the time variation of pressure at the tympanic membrane of each ear. It is possible to record and recreate these variations exactly [Schroeder, 1970], but such prodigious efforts are necessary only for studying the spatial features of the sound. In general, it is sufficient merely to record and reproduce the free field pressure variation at a point well removed from the listener. The resulting sound is audibly inferior to the live sound, but retains nearly all the perceptually important features.

In this investigation we work exclusively with the free field pressure variation, but with one additional complication: analog-to-digital and digital-to-analog conversion of the audio signal. In principle, this conversion can be accomplished with negligible degradation of the signal; in practice, errors can be introduced both by quantization and by sampling and lowpass filtering [Blesser, 1978; 1981]. The work reported here uses 16 bit quantization with a 50 KHz sampling frequency and an 8-pole Butterworth lowpass filter. This results in imperfections which are audible only under worst case test conditions.

The complete sound recording and reproduction sequence is shown in Figure 3. The crucial feature of this sequence is that it does not significantly alter the perceived sound. It follows that the signal in the cochlea and the signal in the computer both carry the same perceptual information. The challenge is to determine which features of the signal carry which pieces of information. It is here that the focus shifts from physical and psychological acoustics to digital signal processing.

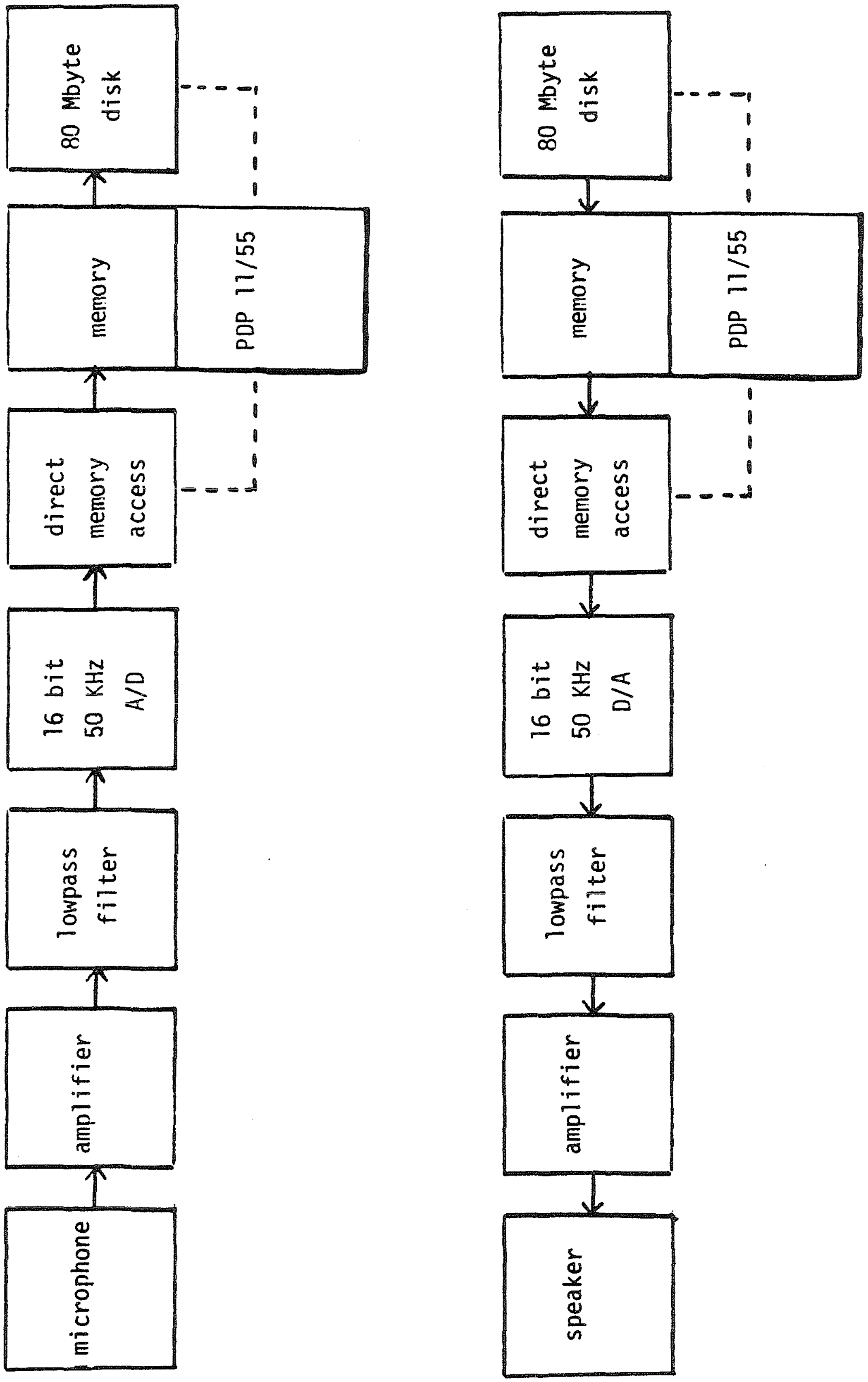


Figure 3. The sound recording and reproduction sequence.

2.3 The additive model

The digitized signal from a typical violin tone is shown in Figure 4. The waveform is basically periodic, but the amplitude, frequency, and waveshape all change significantly over the course of ten to twenty cycles. This quasi-periodic behavior is typical of a great many Western musical instruments. In fact, it will be taken as the defining characteristic of the signals with which this investigation is concerned.

The goal now is to find a representation of these signals in which the perceptually important information is easily identifiable. The underlying periodicity immediately suggests an expansion in terms of Fourier series. Of course, this cannot be done exactly because the periodicity is only approximate. But the fact that the cochlea itself performs a kind of Fourier analysis is an additional and powerful motivation to proceed in this direction. Indeed, an examination of the cochlea suggests that the human auditory system is hardwired for a bandpass filter bank representation of the audio signal. The additive model is an attempt to mimic this representation in a very simple mathematical form.

In the additive model, the signal is treated as a sum of nearly harmonic sinusoids or *partials* . Formally, this is stated as

$$x(n) = \sum_{k=1}^{M} A_k(n) \sin\{ 2\pi nT \, [kf + F_k(n)] \} \tag{2.1}$$

where $x(n)$ is the signal at time nT, T is the time between consecutive samples, f is the fundamental frequency of the tone, $A_k(n)$ is the amplitude of the k^{th} partial at time nT, $F_k(n)$ is the frequency deviation of partial k at time nT, and M is the total number of partials to be included in the sum - typically ten to twenty.

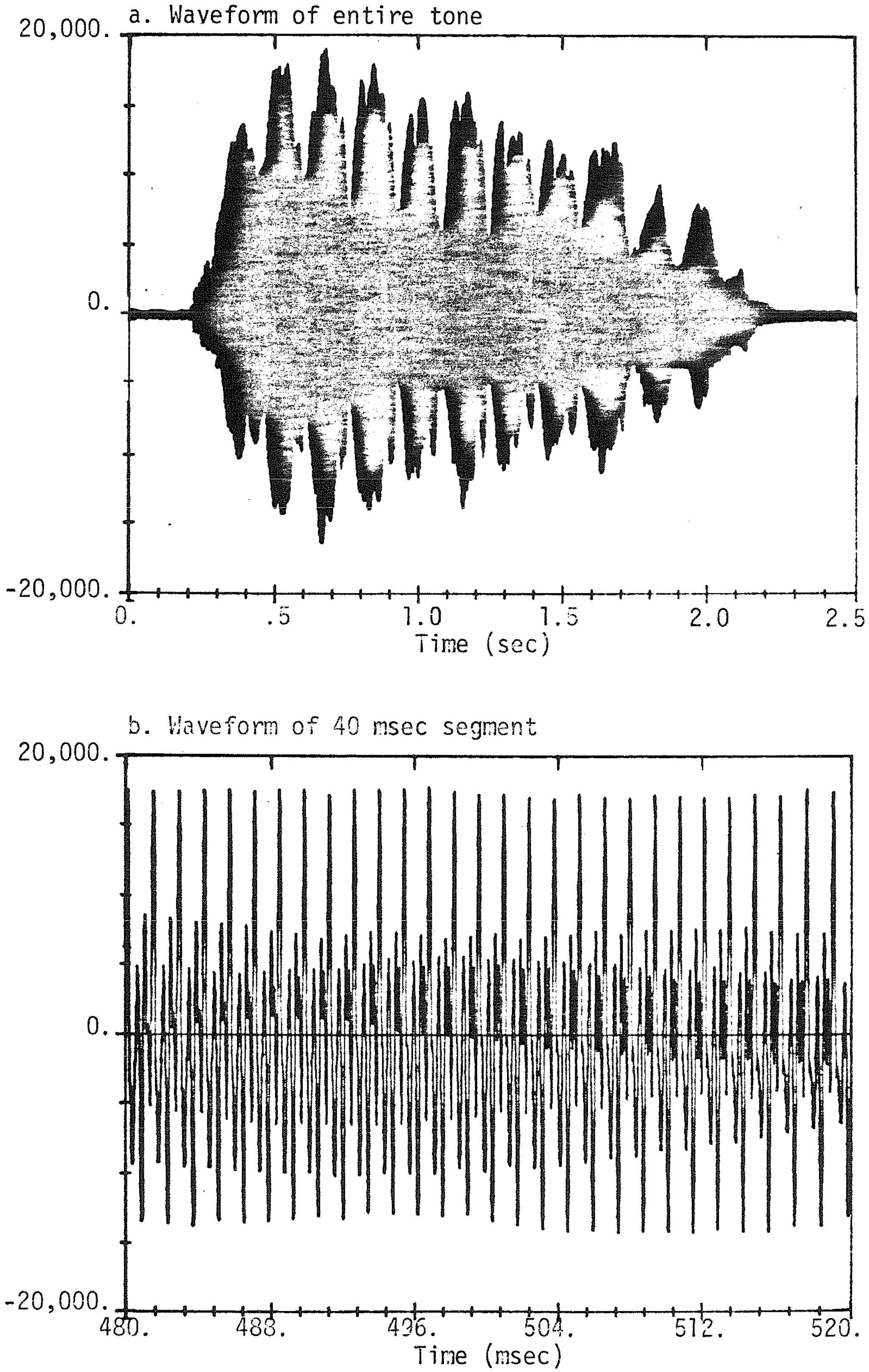


Figure 4. Waveform of a typical digitized violin tone. The tone is F5 (698 Hz) played with moderate vibrato.

It is generally assumed that the fundamental frequency is known or can be easily determined and that the amplitudes and frequencies of each partial are slowly varying. The goal of the analysis is then to estimate those amplitudes $A_k(n)$ and frequency deviations $F_k(n)$ for each partial. There are many techniques for accomplishing this, but by far the most important have been the heterodyne filter [Moorer, 1975] and the phase vocoder [Moorer, 1978]. (Actually, the phase vocoder can be used as an additive analysis-synthesis system without direct reference to the additive model; however, we find it instructive to make the connection explicit.)

2.4 The heterodyne filter

The heterodyne filter can be viewed as a bank of analysis devices in which each device extracts one partial from the signal and estimates its amplitude and frequency as a function of time. The algorithm is given by

$$a_k(n) = \sum_{m=n}^{n+L-1} x(m)\sin(2\pi k\hat{f}mT) \tag{2.2}$$

$$b_k(n) = \sum_{m=n}^{n+L-1} x(m)\cos(2\pi k\hat{f}mT) \tag{2.3}$$

$$\hat{A}_k(n) = [a_k^2(n) + b_k^2(n)]^{\frac{1}{2}} \tag{2.4}$$

$$\hat{\theta}_k(n) = atan \left[\frac{a_k(n)}{b_k(n)} \right] \tag{2.5}$$

$$\hat{F}_k(n) = \frac{1}{2\pi} \frac{d\hat{\theta}_k(n)}{dt} \tag{2.6}$$

where $\hat{f}$ is the known or previously determined fundamental frequency, k is the partial number, $x(n)$ is the input waveform at time nT, and L is chosen so that the summation extends over exactly one period of the input waveform. The variables $\widehat{A}_k(n)$, $\widehat{\theta}_k(n)$, and $\widehat{F}_k(n)$ are the desired estimates of the actual amplitude, phase, and frequency respectively.

The key to this approach is the choice of L such that $\widehat{f} = \frac{1}{LT}$. Equations (2.2) and (2.3) can be viewed as applying a filter with z transform $H(z) = \frac{1-z^{-L}}{1-z^{-1}}$ to the signals $x(m)sin(2\pi\widehat{f}mT)$ and $x(m)cos(2\pi\widehat{f}mT)$ respectively. With the proper choice of L, this places zeros of transmission at all harmonic frequencies $r\widehat{f}$ except for $r=k$. With the proper choice of $\widehat{f}$, these zeros annihilate all partials of the waveform $x(n)$ except for the k^{th} partial. The sequences $a_k(n)$ and $b_k(n)$ then constitute an accurate phase quadrature representation of that partial, and equations (2.4) thru (2.6) simply convert to polar form. However, any error in $\widehat{f}$ will lead to incomplete cancellation of the other partials and a corresponding ripple on the amplitude and frequency estimates. To eliminate this ripple, these estimates are themselves subjected to considerable additional smoothing before finally being accepted.

Although the heterodyne filter is fairly straightforward, its implementation in the sampled-data domain involves several subtleties which reappear in the phase vocoder and are therefore deserving of mention. For example, the operation described in equation (2.5) produces only the principal value of the phase angle while the derivative in equation (2.6) assumes a continuous phase function. Methods for unwrapping the principal value into a continuous function are given by Schafer [1969] and Moorer [1975]. Another example is presented by the derivative itself. This operation is performed either by first fitting a polynomial to the phase function [Moorer, 1975] or by filtering the phase function with a bandlimited differentiator [Rabiner and Gold, 1975].

The most important issue in the implementation of the heterodyne filter is the question of how often the formulas (2.2) thru (2.6) must be evaluated to avoid significant degradation of the estimates. The Nyquist theorem suggests that the estimates be computed at no less than twice the highest frequency present in the signals $\hat{A}_k(n)$ and $\hat{F}_k(n)$. However, these signals are fundamentally non-bandlimited because of the nonlinear operations by which they were obtained. Moorer therefore suggested that the calculations be carried out at the original input sample rate. This issue will be examined more closely in Chapter 3.

2.5 The use of the heterodyne filter in the investigation of timbre

It is easy to show that the heterodyne filter can accurately extract the amplitudes and frequencies of each harmonic in a synthetic tone of constant fundamental frequency $\hat{f}$. The more important issue is how it performs on actual instrument waveforms. This question was extensively investigated by Grey as the first step in his far-reaching study of timbre [Grey and Moorer, 1977].

Grey tested a number of musically sophisticated listeners in a high quality sound recording and reproduction environment with stimuli consisting of isolated tones from a variety of instruments. The tones were presented both in their original tape-recorded form and in modified versions based on the additive model. Grey found that those tones which were resynthesized directly from the amplitude and frequency estimates of the heterodyne filter were nearly indistinguishable from the originals. Furthermore, those differences which were perceptible seemed more related to articulation and background noise than to actual timbre. This result established additive analysis-synthesis, and the heterodyne filter in particular, as a valid technique for investigating timbre; it indicated that the entire procedure did indeed preserve the perceptually significant information in the musical signal.

Grey further discovered that each amplitude and frequency function $\hat{A}_k(n)$ and $\hat{F}_k(n)$ could be replaced by a several line-segment approximation with only a small additional loss of fidelity in the resynthesized version. This tremendously reduced the amount of data required by the additive model but with very little sacrifice of important information. Furthermore, it confirmed the intuitively appealing notion that the overall shape of the amplitude and frequency functions was far more important than their detailed structure. This was an additional suggestion of the extent to which the additive model matched the representation within the human auditory system itself.

Having established the shape of the amplitude and frequency functions as the primary determinant of perceived timbre, Grey next attempted to link specific aspects of the shape with specific features of the perceived sound. In this endeavor, he was only moderately successful; in the process, though, he provided an outstanding paradigm for the investigation of timbre:

1) Acquire several samples of the sounds to be investigated, and fit them to the additive model using a technique such as the heterodyne filter or phase vocoder.

2) Equalize the various samples in terms of perceived loudness, pitch, and duration by appropriate scaling of the amplitude and frequency functions in the additive model representation.

3) Form a hypothesis as to which feature of the additive model representation is related to a particular perceptual feature.

4) Test the hypothesis by modifying only that one feature in the additive model representation and then resynthesizing the sound.

The crucial feature of this approach is the continued reliance on additive analysis-synthesis, not only to provide an instructive representation of the signal, but also to equalize all perceptual aspects of the signal except the one specifically under examination. We will adhere closely to this methodology throughout Chapter 4.

The difficult step in the above procedure is the formation of a likely hypothesis for testing. Grey, in his search for promising hypotheses, made extensive use of the technique of multidimensional scaling. This is a procedure in which each possible pair in a set of stimuli is rated on the basis of perceived similarity. These ratings form the input to a computer program which represents the stimuli as a set of points in a multidimensional space where distance is proportional to perceived dissimilarity. By collapsing this space into a relatively few dimensions while preserving the distance relations, the program can suggest the subconscious groupings which led to the observed similarity judgements. These groupings can in turn suggest those particular aspects of the signal which are perceived to be most important.

In this investigation, we reject the use of multidimensional scaling for two simple but rarely mentioned reasons. First, the groupings obtained with this technique are extremely dependent on the particular stimuli in the original sample set. This means that they are not nearly so fundamental as is often assumed. Second, and most importantly, the groupings are only suggestive of the particular features which the brain appears to target. The correct identification of these features still depends very much on the ingenuity of the person examining the groupings. The contribution of the multidimensional scaling procedure is therefore rather small.

2.6 More recent developments

The primary limitation of the heterodyne filter is the requirement that the fundamental analysis frequency $\hat{f}$ always be very near to the actual fundamental frequency of the signal. Hence, this technique can be applied only to isolated tones with no vibrato. To overcome these restrictions, Moorer suggested the use of the phase vocoder [1978]. However, discussion of this technique will be deferred until Chapter 3.

With the work of Moorer and Grey, the additive model became widely recognized as the standard for high fidelity sound synthesis; however, the model itself has attracted little additional research. Two investigators attempted to further reduce the amount of data required by the additive model, each with moderate success [Strawn, 1979; Charbonneau, 1981]. Beyond this, timbre research has continued to consist either of analyses of natural instrument tones with no attempt at resynthesis, or of perceptual tests of highly artificial electronic tones.

Research in the digital analysis of musical sound has also shifted its focus from additive analysis-synthesis. A number of studies have been conducted with the goal of developing improved nonlinear analysis-synthesis techniques [LeBrun, 1979; Arfib, 1979; Beauchamp, 1982]. These techniques have the merit of being computationally efficient, but with a corresponding reduction in fidelity. Attention has also been focused on the automatic transcription of music with an additive model, but with less emphasis on accurate reproduction of timbre [Piszczalski and Galler, 1977; 1981].

One recent and novel idea which could be useful in future investigations of timbre is that of critical band analysis-synthesis [Petersen and Boll, 1981]. This technique models the critical band phenomena of the basilar membrane with a kind of *constant-Q* Fourier transform in which the width of each frequency bin is proportional to its center frequency. This maps a given input signal into a representation which is even more similar to that of the cochlea than the additive model representation. However, it remains to be seen how successful this technique will be in practice.

III. THE PHASE VOCODER

The phase vocoder has proven to be an extremely useful tool for performing additive analysis-synthesis. In this chapter we seek an understanding of the errors which arise when the phase vocoder magnitude and phase-derivative signals are modified prior to resynthesis. We begin with a review of the phase vocoder and its relation to short-time Fourier analysis. We then consider the relation between the phase vocoder magnitude and phase-derivative signals and the parameters of the additive model; this leads to the identification of two independent sources of error. We show that the first of these involves crosstalk between adjacent frequency channels and can be eliminated by the development of a tracking version of the phase vocoder. Alternatively, restrictions can be placed on the phase-derivative signal to preserve the absolute phase. The second source of error appears to be inherent in the phase vocoder formulation.

3.1 History

The phase vocoder is an analysis-synthesis technique based upon the time-dependent Fourier transform. It was originally developed by Flanagan and Golden [1966] as a device for reducing the bandwidth of speech signals; however, its usefulness in modifying pitch and timing was also immediately apparent. Unfortunately, the early implementations of this technique were so computation intensive that its attraction was very limited.

The following decade brought a number of advances in our understanding of short-time Fourier analysis culminating in the work of Portnoff [1976]. Portnoff showed that the phase vocoder could be formulated as an analysis-synthesis identity system; the synthesized output could be made identical to the input both in theory and in practice. Furthermore, Portnoff described efficient techniques for performing both the analysis and synthesis. It was this work which led Moorer [1978] to suggest that the phase vocoder replace his earlier heterodyne filter technique for additive analysis-synthesis.

Moorer's investigation of the phase vocoder was largely empirical; he simply demonstrated its effectiveness in the analysis and synthesis of musical sounds. He observed that using the phase vocoder to modify sounds could result in strange errors, but he did not attempt to analyze this phenomenon. A similarly empirical viewpoint was adopted by Flanagan and Christensen [1980] in their evaluation of phase-vocoder-based schemes for reducing the bandwidth of speech signals. Only in a companion paper by Flanagan [1980] was there any attempt to analyze the phase vocoder itself. The novel feature of our work is to explicitly relate the phase vocoder magnitude and phase-derivative signals to the parameters of the additive model. This provides a basis for understanding the errors which arise from the undersampling and modification of these signals.

3.2 Short-time Fourier analysis

We begin our investigation by reviewing the fundamentals of short-time Fourier analysis. Our treatment of this area is basically that of Rabiner and Schafer [1978] with some appropriate modifications.

A proper understanding of the phase vocoder begins with an understanding of the time-dependent Fourier transform. A useful definition of the time-dependent Fourier transform is

$$X(n,\omega) = \sum_{m=-\infty}^{\infty} h(n-m)\, x(m)\, exp(-j\omega m) \qquad (3.1)$$

where $x(m)$ is the signal and $h(n-m)$ is an appropriate window function. There are two distinct ways in which this equation can be viewed: the Fourier transform interpretation, and the linear filtering interpretation.

In the first interpretation, we assume that n is fixed; then $X(n,\omega)$ is just the normal Fourier transform of the sequence $h(n-m)x(m)$. A sufficient condition for this transform to exist is that it be absolutely summable. Since the window $h(n-m)$ is usually of finite duration, this condition is easily satisified.

In this interpretation, the role of the window is to select a portion of the signal to be Fourier transformed. The specific portion selected is determined by the value of n. It is also clear from this interpretation that the Fourier transform $X(n,\omega)$ is the convolution of the desired signal transform with the Fourier transform of the window. Consequently, the window should be chosen to have a transform with the magnitude characteristic of a narrow band, lowpass filter.

In the second interpretation, we assume that ω is fixed; then $X(n,\omega)$ can be viewed as the convolution of the impulse response $h(n)$ with the sequence $x(n)exp(-j\omega n)$. In this form, the similarity with Moorer's heterodyne filter is unmistakable. Indeed, if we choose

$$h(n) = \begin{cases} 1 & -L < n \leq 0 \\ 0 & otherwise \end{cases} \tag{3.2}$$

where L is the number of samples in a single period of $x(n)$, then equation (3.1) can be viewed simply as a complex version of equations (2.2) and (2.3). However, it is generally advantageous to choose an $h(n)$ which more closely approximates an ideal lowpass filter.

In this interpretation, the multiplication of $x(n)$ by $exp(-j\omega n)$ shifts the spectrum of $x(n)$ in the region of ω to baseband (and also to 2ω). The baseband spectrum is extracted by the narrow band, lowpass filter with impulse response $h(n)$. The virtue of this interpretation is that it makes explicit the fact that the time-dependent Fourier transform is a discrete-time, linear, shift-invariant system.

Based on these two interpretations, two distinct resynthesis procedures have been developed. In each of these, we assume that $X(n,\omega)$ has been evaluated at N frequencies $\omega_k = \frac{2\pi}{N}k$ where $k = 0, 1, ..., N-1$. The first synthesis technique is the overlap addition method, derived from the Fourier transform interpretation:

$$y(n) = \sum_{r=-\infty}^{\infty} \frac{1}{N} \sum_{k=0}^{N-1} X(rR,\omega_k)\, exp(j\omega_k n) \tag{3.3}$$

This procedure inverts the Fourier transform $X(n,\omega)$ for every R^{th} value of n to produce windowed segments of $x(n)$ spaced by R samples in time. These segments are overlapped and added to produce the resynthesized version of $x(n)$.

The second resynthesis technique is the filter bank summation method, derived from the linear filtering interpretation:

$$y(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(n,\omega_k)\, exp(j\omega_k n) \tag{3.4}$$

In this procedure, the resynthesized signal is simply the sum of signals from each band of the filter bank, translated back to the original center frequencies of the band.

With either of these synthesis techniques, the synthesized $y(n)$ will be a close approximation to the original $x(n)$ provided that the sampling rate of $X(n,\omega)$ is sufficiently great. From the linear filtering interpretation, it is clear that the sampling rate in time must be at least twice the bandwidth of the filter. It can then easily be shown that the Fourier transform interpretation imposes a related restriction on the sampling in frequency such that the required overall sampling rate is on the order of 2 to 4 times that of $x(n)$. In return for this higher sampling rate, we obtain a very general representation of $x(n)$ with tremendous flexibility and usefulness. However, not even this tradeoff is always necessary.

By substituting equation (3.1) into equation (3.4), we find that the filter bank summation technique can be used to resynthesize $x(n)$ exactly and at the original sampling rate, provided only that some very modest restrictions are applied to $h(n)$. To do this, the sampling rate in time is maintained at twice the filter bandwidth, but the sampling rate in frequency (i.e., the value of N) is reduced. The key requirement is simply that $h(n) = 0$ for all (nonzero) values of n which are integer multiples of N. With this one additional constraint, equations (3.1) and (3.4) can be rewritten as an analysis-synthesis identity system:

$$X(n,k) = \sum_{m=-\infty}^{\infty} h(n-m)\, x(m)\, exp(-j\, \frac{2\pi}{N} km) \tag{3.5}$$

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(n,k)\, exp(j\, \frac{2\pi}{N} kn) \tag{3.6}$$

where $k = 0, 1, \ldots, N-1$. If $X(n,k)$ is expressed in magnitude and phase form, then equations (3.5) and (3.6) can also be taken as the definition of the phase vocoder.

The Fourier transform interpretation can also be employed to develop more efficient formulations of equations (3.5) and (3.6). In both cases, the key to improved efficiency is to manipulate the equation into the form of a discrete Fourier transform. For equation (3.5), we seek $f(m-n)$ such that

$$X(n,k) = \sum_{m=0}^{N-1} f(m-n)\, exp(-j\, \frac{2\pi}{N} km) \tag{3.7}$$

If N is a power of 2, then a substantial reduction in computation can be obtained by employing a Fast Fourier Transform algorithm to perform the calculation.

It turns out that this manipulation is always possible provided that the window $h(n)$ is of finite duration, say L samples. If $L=N$, then the manipulation is trivial. In the more general case, it can be shown [Schafer and Rabiner, 1973] that the required function $f(m-n)$ is

$$f(m-n) = u_n((m-n))_N \tag{3.8}$$

where the notation $((\ \ ))_N$ means that the integer inside the double set of parentheses is to be interpreted modulo N, and

$$u_n(q) = \sum_{r=-\infty}^{\infty} x(Nr+q+n)\, h(-Nr-q) \tag{3.9}$$

for $q = 0, 1, \ldots, N-1$. This looks complicated but simply indicates that the sequence $x(m+n)h(-m)$ is broken into segments of length N samples, and that the segments are added together to produce $u_n(q)$ which is then circular shifted by n modulo N.

The advantage of this approach can easily be demonstrated by example. The direct implementation of equation (3.5) requires $2LN$ real multiplications per output sample; equation (3.7) requires $L + 2N\log_2 N$. In practice, the values of L and N are chosen so that the frequency bins of the transform are at least as closely spaced as the harmonics. If the sampling rate is 50 KHz and the fundamental frequency is 200 Hz, then appropriate values for L and N are L=1200 and N=256. This translates to a factor of 140 reduction in the required computation load.

A similar savings can be effected by manipulating equation (3.6) to give

$$x(n) = \sum_{m=n-RQ+1}^{n+RQ-1} h(n-m)\, \frac{1}{N} \sum_{k=0}^{N-1} V(m,k)\, exp(j\, \frac{2\pi}{N} kn) \tag{3.10}$$

where

$$V(m,k) = \begin{cases} X(m,k) & m = 0, \pm R, \pm 2R, \ldots \\ 0 & otherwise \end{cases} \tag{3.11}$$

and $L = 2RQ-1$. It is assumed that $X(n,k)$ has been computed at the minimum allowable rate; hence there is one sample of $X(n,k)$ for every R samples of $x(n)$. Rather than interpolating the values of $X(n,k)$ and then evaluating equation (3.6), this technique transforms back to the time domain first and then interpolates.

On the other hand, in situations where only a few frequency channels of the phase vocoder are actually required, it may be more efficient to calculate each channel separately. In this case, the sequence $x(n) exp(-j \frac{2\pi}{N} kn)$ can be obtained with relatively little computation by calculating it prior to the convolution with $h(n)$. The calculation of $X(n,k)$ can then be viewed as a simple lowpass filtering problem.

Lastly, we note that the filter $h(n)$ can easily be made to obey the constraint that $h(rN)=0$ for $r = \pm 1, \pm 2, \ldots$ by defining it as a suitably windowed version of the ideal lowpass impulse response

$$h_{ideal}(n) = \frac{sin(\frac{\pi n}{N})}{\frac{\pi n}{N}} \tag{3.12}$$

This approach also guarantees that the filter $h(n)$ will have a linear phase response. A number of appropriate classical windows are discussed by Harris [1978]. Alternatively, a suitable window can be designed using the McClellan-Parks-Rabiner optimal FIR filter design program [1975].

3.3 The phase vocoder

We now consider the phase vocoder specifically. In principle, the phase vocoder is just another technique for performing short-time Fourier analysis-synthesis; in practice, however, it presents several additional complications, all of which can be attributed to its reliance on magnitude and phase.

A useful definition of the phase vocoder is provided by the following pair of equations:

$$X(n,k) = \sum_{m=-\infty}^{\infty} h(n-m)\, x(m)\, exp(-j\,\frac{2\pi}{N}km) \tag{3.13}$$

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} |X(n,k)|\; exp(j\,\varphi(n,k))\; exp(j\,\frac{2\pi}{N}kn) \tag{3.14}$$

If $\varphi(n,k)$ is simply $arg[X(n,k)]$, then equations (3.13) and (3.14) are identical to equations (3.5) and (3.6), and everything is fine. However, the prime attraction of the phase vocoder is that it can provide a representation of $x(n)$ in which the perceptually significant parameters - amplitude and frequency - are explicit. This is what makes it so useful in data rate reduction and in independent modification of pitch and timing. Hence, taking some liberties with notation, we define

$$\varphi(n,k) = \int_0^{nT} \dot{\varphi}(t,k)\, dt \; + arg[X(0,k)] \tag{3.15}$$

$$\dot{\varphi}(n,k) = \frac{d}{dt}\, arg\,[X(n,k)] \tag{3.16}$$

We will refer to $|X(n,k)|$, $\varphi(n,k)$, and $\dot{\varphi}(n,k)$ as the *magnitude*, *phase*, and *phase-derivative* signals, respectively. In contrast, we will reserve the terms *amplitude* and *frequency* for describing the signals which constitute the input to the phase vocoder. For convenience, we also define

$$a(n,k) = \mathrm{Re}\, X(n,k) \tag{3.17}$$

$$b(n,k) = \mathrm{Im}\, X(n,k) \tag{3.18}$$

We now examine the problems which arise as a consequence of the transformation to magnitude and phase derivative.

The first problem is that the bandwidths of $|X(n,k)|$ and $\dot{\varphi}(n,k)$ can be significantly greater than those of $a(n,k)$ and $b(n,k)$; this is a consequence of the highly nonlinear transformation required for the conversion to magnitude and phase derivative. This problem does not necessarily prevent the phase vocoder from being useful for data rate reduction; rather, it reflects the fact that the magnitude and phase-derivative signals are generally more slowly varying than $a(n,k)$ and $b(n,k)$, but that they can occasionally be more rapidly varying.

The bandwidths of $|X(n,k)|$ and $\dot{\varphi}(n,k)$ are important because they determine the minimum allowable sampling rate for these signals. These signals can always be sampled at the same rate as $a(n,k)$ and $b(n,k)$ if no subsequent modifications are intended; but there is no guarantee that such signals will be usefully related to the actual amplitudes and frequencies. In the case of speech signals, Flanagan [1980] suggested that $|X(n,k)|$ and $\dot{\varphi}(n,k)$ could probably be bandlimited to one quarter of the filter bandwidth. However, for musical signals, where fidelity is a far more important criterion, Moorer [1978] recommended that the conversion be carried out at the original input sampling frequency. Unfortunately, this results in a tremendous increase in the computation requirements.

When $X(n,k)$ is required at the input sampling rate, it is still most efficient to first obtain it at the minimum output sampling rate as described in Section 3.2; the decimated version can then be interpolated back to the original sampling rate. However, the transform technique of equation (3.10) can now no longer be used, and the interpolation alone requires $2LN$ multiplications per output sample. This is a tremendous increase in computation even without considering the calculations required to perform the conversion to magnitude and phase derivative.

The other problem in the implementation of the phase vocoder has to do with the way in which this conversion is actually performed. The magnitude and phase can easily be obtained as

$$|X(n,k)| = [a^2(n,k) + b^2(n,k)]^{\frac{1}{2}} \tag{3.19}$$

$$\varphi(n,k) = atan\left[\frac{b(n,k)}{a(n,k)}\right] \tag{3.20}$$

But the phase in equation (3.20) is only the principal value. To differentiate the phase, a continuous phase function must be available. One alternative is to calculate

$$\dot{\varphi}(n,k) = \frac{a(n,k)\dot{b}(n,k) - b(n,k)\dot{a}(n,k)}{a^2(n,k) + b^2(n,k)} \tag{3.21}$$

where $\dot{a}(n,k)$ and $\dot{b}(n,k)$ are obtained by filtering $a(n,k)$ and $b(n,k)$ with a bandlimited differentiator. This method was used by Flanagan and Christensen [1980]. However, Moorer [1978] showed that a preferable solution is to calculate only angle differences. It can easily be shown that

$$\Delta\varphi(n,k) = atan\left[\frac{b(n,k)a(n-1,k) - a(n,k)b(n-1,k)}{a(n,k)a(n-1,k) - b(n,k)b(n-1,k)}\right] \tag{3.22}$$

with initial conditions $a(0,k) = 1$, $b(0,k) = 0$. By using the angle differences in place of $\dot{\varphi}(n,k)$, the original phase can be reconstructed exactly. Moorer saw this simply as a way to recreate discontinuities in $x(n)$, but we will show that it is in fact crucial to the effective use of the phase vocoder.

In the sections to follow, we present a number of results obtained using only several channels of a phase vocoder; the values of N in these examples are similar but not identical to those which would be used in a multi-channel Fourier transform implementation. The filters for this vocoder were all designed by applying a Blackman window to the ideal impulse response of equation (3.12). This gives $h(n) = [\ .42 + .50 \cos(\frac{2\pi}{L}n) + .08 \cos(\frac{2\pi}{L}2n)]\ h_{ideal}(n)$ where $n = 0, \pm 1, \pm 2, \ldots, \pm\frac{L}{2}$. Whenever phase derivatives (as opposed to angle differences) were required, they were obtained by filtering the unwrapped phase with a bandlimited differentiator designed via the McClellan-Parks-Rabiner optimal FIR filter design program. The filtering was performed by fast convolution using the overlap-add method. Unless otherwise specified, the input *and* output sampling rates in all examples are 50 KHz.

3.4 Error sources: estimated amplitude and frequency

Because the phase vocoder can be efficiently implemented as an analysis-synthesis identity system, it has been widely assumed that the relation between the magnitude and phase-derivative signals and the parameters of the additive model (i.e., amplitude and frequency deviation) is unimportant. This is true when the synthesis is performed without undersampling or modification, but such instances are the exception. We now consider the more general case.

The effects of linear (ie. additive and multiplicative) modifications of $X(n,k)$ were investigated by Allen and Rabiner [1977]. They showed that the consequences of such modifications depended on the particular technique used to perform the synthesis: overlap addition versus filter bank summation. They further showed that, for the filter bank summation technique, the effect of an additive spectral modification was to produce an additive component in the reconstructed signal, while the effect of a multiplicative spectral modification was to convolve the original signal with a time-limited, window-weighted version of the time response due to the modification. However, they left unanswered the entire question of nonlinear modification.

In our investigation, we are particularly concerned with nonlinear modifications of the type

$$[\ \dot{\varphi}(n,k) + \frac{2\pi}{N}k\ ] \rightarrow \alpha[\ \dot{\varphi}(n,k) + \frac{2\pi}{N}k\ ] \tag{3.23}$$

where α is a scaling constant of order unity. This type of modification is frequently employed to introduce changes in pitch. We are also concerned with the bandwidths of $|X(n,k)|$ and $\dot{\varphi}(n,k)$ because of the implications for sampling rate modifications. In both cases, our approach is to relate the magnitude and phase-derivative signals to the actual amplitudes and frequencies of the input signals. In the process, we generalize some of the results of Flanagan [1980].

As a prelude, we recall that the Hilbert transform of a signal $y(n)$ is defined as the convolution of that signal with the impulse response

$$h(n) = \frac{\sin^2(\frac{\pi n}{2})}{\frac{\pi n}{2}} \qquad n \neq 0 \tag{3.24}$$

where $h(0)=0$. Thus, the Hilbert transform of $y(n)$ is

$$\hat{y}(n) = \sum_{m \neq 0} y(n-m) \frac{\sin^2(\frac{\pi m}{2})}{\frac{\pi m}{2}} \tag{3.25}$$

(In this chapter, $\hat{y}(n)$ always indicates the Hilbert transform of $y(n)$; this should not be confused with the notation of Chapter 2 in which it referred to the estimate of $y(n)$.) A more useful definition of the Hilbert transform can be obtained via the inverse Fourier transform:

$$\hat{y}(n) = \frac{1}{2\pi}\left[\int_{0}^{\pi} -jY(\theta)\, exp\,(j\,\theta n)\, d\,\theta + \int_{\pi}^{2\pi} jY(\theta)\, exp\,(j\,\theta n)\, d\,\theta\right] \tag{3.26}$$

where $Y(\theta)$ is the Fourier transform of $y(n)$. In this view, the Hilbert transform is obtained by retarding the phases of all sinusoidal components by $\frac{\pi}{2}$. Furthermore, the Fourier transform of $y(n) + j\hat{y}(n)$ is nonzero only for positive frequencies. This makes the Hilbert transform very useful in representing narrow band signals; it is frequently used this way in communication theory [eg. Schwartz, 1966], and its use in the analysis to follow is patterned after the communication theory approach.

We begin by assuming a signal $x(n)$ which is bandlimited to frequencies between ω_1 and ω_2. Such a signal can always be expressed in the form

$$x(n) = s_r(n)cos\,(\omega_o nT) + s_q(n)sin\,(\omega_o nT) \tag{3.27}$$

or equivalently,

$$x(n) = A(n)cos\,(\omega_o nT+\theta(n)) \tag{3.28}$$

where

$$A(n) = [s_r^2(n) + s_q^2(n)]^{\frac{1}{2}} \tag{3.29}$$

$$\theta(n) = atan\left[-\frac{s_q(n)}{s_r(n)}\right] \tag{3.30}$$

Here, ω_o is any frequency not less than $\frac{\omega_2}{2}$, and T is the sampling period.

There are infinitely many pairs of $s_r(n)$ and $s_q(n)$ (or $A(n)$ and $\theta(n)$) which satisfy equation (3.27) for a given $x(n)$. However, there is a "best" pair, at least in a certain sense. To obtain it, we impose the additional assumption that $s_r(n)$ and $s_q(n)$ are bandlimited to frequencies less than ω_o. (Note, however, that we make no such assumptions about $A(n)$ and $\theta(n)$.) We can now take the Hilbert transform of equation (3.27) to obtain

$$\hat{x}(n) = s_r(n)\sin(\omega_0 nT) - s_q(n)\cos(\omega_0 nT) \tag{3.31}$$

Equations (3.27) and (3.31) are linear in $s_r(n)$ and $s_q(n)$; hence, they can be solved to yield

$$s_r(n) = x(n)\cos(\omega_0 nT) + \hat{x}(n)\sin(\omega_0 nT) \tag{3.32}$$

$$s_q(n) = x(n)\sin(\omega_0 nT) - \hat{x}(n)\cos(\omega_0 nT) \tag{3.33}$$

which is, in fact, the "best" pair in that these particular $s_r(n)$ and $s_q(n)$ are the lowpass functions of minimum bandwidth.

In the analysis to follow, we always assume that $s_r(n)$ and $s_q(n)$ are defined as in equations (3.32) and (3.33). It is then easy to show that $A(n)$ and $\theta(n)$ in equations (3.29) and (3.30) are given by

$$A(n) = [x^2(n) + \hat{x}^2(n)]^{\frac{1}{2}} \tag{3.34}$$

$$\theta(n) = atan\left[\frac{\hat{x}(n)}{x(n)}\right] - \omega_0 nT \tag{3.35}$$

This may be recognized as the motivation for the frequently defined *analytic signal*

$$z(n) = x(n) + j\hat{x}(n) \tag{3.36}$$
$$= A(n)\, exp\,(j\,\theta(n))\, exp\,(j\,\omega_0 nT) \tag{3.37}$$

from which equation (3.28) can be obtained by taking the real part.

We now consider the relation between $A(n)$ and $\theta(n)$ and the phase vocoder signals $|X(n,k)|$ and $\varphi(n,k)$. With no loss of generality, we assume that ω_0 is such that $\frac{2\pi k}{NT} = \omega_0$. We then have (equation 3.38)

$$X(n,\omega_0) = \sum_{m=-\infty}^{\infty} h(n-m)x(m)\cos(\omega_0 mT) - j\sum_{m=-\infty}^{\infty} h(n-m)x(m)\sin(\omega_0 mT)$$

Substituting equation (3.27) for $x(n)$ gives (equations 3.39 and 3.40)

$$a(n,\omega_o)=\sum_{m=-\infty}^{\infty} h(n-m)s_r(m)[1+\cos(2\omega_o mT)]+\sum_{m=-\infty}^{\infty} h(n-m)s_q(m)\sin(2\omega_o mT)$$

$$b(n,\omega_o)=\sum_{m=-\infty}^{\infty} h(n-m)s_r(m)\sin(2\omega_o mT)-\sum_{m=-\infty}^{\infty} h(n-m)s_q(m)[1-\cos(2\omega_o mT)]$$

where a constant factor of $\frac{1}{2}$ has been absorbed into the filter impulse response $h(n)$.

We now note that $s_r(n)$ and $s_q(n)$ are bandlimited to ω_o and that $h(n)$ is chosen to be the impulse response of a lowpass filter. We therefore make the (excellent) assumption that $h(n)$ is a sufficiently good filter to make the terms at $2\omega_o$ negligible. We then have

$$a(n,\omega_o) = \sum_{m=-\infty}^{\infty} h(n-m)\, s_r(m) \qquad (3.41)$$

$$b(n,\omega_o) = -\sum_{m=-\infty}^{\infty} h(n-m)\, s_q(m) \qquad (3.42)$$

Furthermore, if $h(n)$ is the impulse response of an ideal lowpass filter with cutoff frequency ω_o, then

$$X(n,\omega_o) = s_r(n) - j\, s_q(n) \qquad (3.43)$$

and $|X(n,k)|$ and $\varphi(n,k)$ are identical to $A(n)$ and $\theta(n)$. This result was also obtained by Flanagan [1980], but in a less rigorous fashion.

In the more general case, the impulse response $h(n)$ differs significantly from the ideal; in fact, it is a windowed version of the ideal. We can then convert equations (3.41) and (3.42) back to magnitude and phase form to obtain

$$|X(n,\omega_o)| = \left|\sum_{m=-\infty}^{\infty} h(n-m)\, A(m)\, exp(j\,\theta(m))\right| \qquad (3.44)$$

$$\varphi(n,\omega_o) = arg\left[\sum_{m=-\infty}^{\infty} h(n-m)\, A(m)\, exp(j\,\theta(m))\right] \qquad (3.45)$$

which is the result we are seeking.

Equations (3.44) and (3.45) show the connection between the phase vocoder magnitude and phase-derivative signals and the parameters of the additive model; it is instructive to examine them in detail. In particular, we consider the phase vocoder signals as estimates of the additive model parameters and show that there is an undesirable coupling of the amplitude and phase through the filter convolution. This coupling is automatically unscrambled in the resynthesis, but it can be a source of error when the phase is modified prior to resynthesis. We seek ways to minimize this error.

We first examine equation (3.44) and assume that the frequency deviation of the signal $x(n)$ from ω_0 is zero. Then $\theta(n) = 0$, and the phase vocoder magnitude is simply the magnitude of the true amplitude convolved with the impulse response $h(n)$. Furthermore, we note that $A(n)$ is always positive; hence, the absolute value is inconsequential except for very rapid attacks and decays where a filter with a sharp cutoff may introduce some ringing.

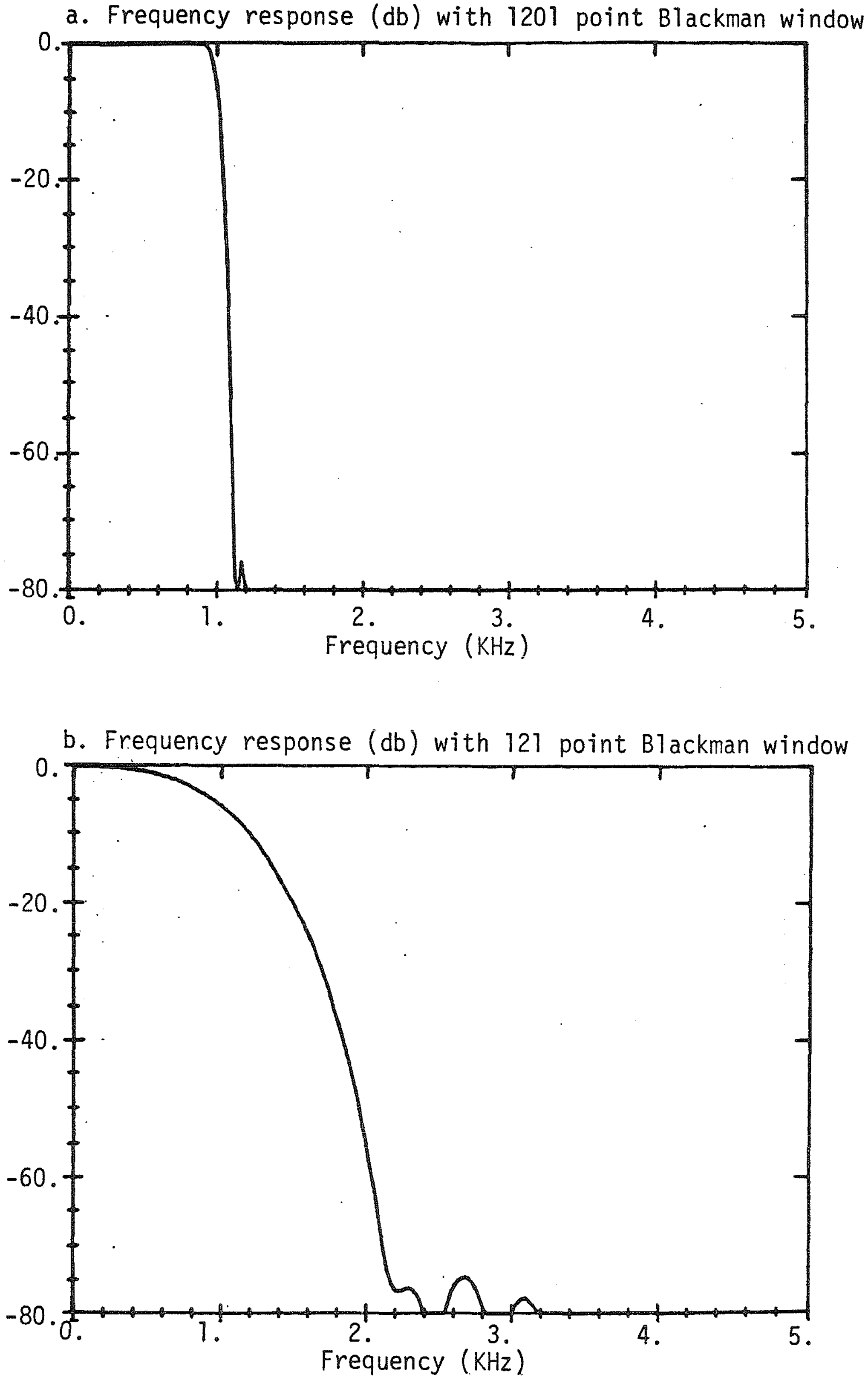
From this, two immediate conclusions can be drawn. First, the filter bandwidth should be as wide as possible to minimize smearing of the amplitude estimate. (Of course, the bandwidth cannot be greater than the fundamental frequency of the tone being analyzed.) Secondly, the bandwidth of $|X(n,k)|$ is simply that of the lowpass filter (except possibly during rapid attacks and decays) provided that the analysis frequency and the instantaneous frequency are closely matched.

If we now assume that the analysis frequency and the instantaneous frequency are not well matched but that the latter is still constant, we can write $\theta(n) = \frac{2\pi}{N}\beta n$. Then equation (3.44) can be viewed as taking a kind of Fourier transform of the impulse response $h(n-m)$ weighted by the amplitude $A(n)$, and evaluating the transform at frequency $\frac{2\pi}{N}\beta$; if the actual amplitude is constant then the estimated amplitude is scaled by $H(\frac{2\pi}{N}\beta)$.

Again, two conclusions are possible. First, the filter magnitude response should be as flat as possible. Secondly, the bandwidth of $|X(n,k)|$ can be made arbitrarily large, for example, by choosing $\theta(n)$ to be a linearly increasing frequency (a *chirp*). Then the time response of $|X(n,k)|$ is simply the frequency response of $h(n)$ with a time scale depending on the rate of increase of $\theta(n)$. (Of course, a chirp is not strictly a narrowband input, but the conclusion is still valid.)

For equation (3.45) it is more difficult to make exact statements, but it is clear that any variation in the amplitude $A(n)$ distorts the averaging which is used to estimate $\theta(n)$. In particular, a sharp dip in amplitude tends to flatten the filter impulse response and introduce significant phase errors. In contrast, a sharp peak in amplitude tends to reinforce the filter impulse response and not create any significant distortion in the phase signal. Furthermore, distortions in phase are considerably magnified in the phase-derivative signal. This pattern is unfortunate in that sharp amplitude minima are quite common in some sounds, but fortunate in that frequency errors which occur during amplitude minima are not easily heard. As in equation (3.44), this distortion can be eliminated by matching the analysis frequency to the instantaneous frequency.

Confirmation of the above observations is provided by Figures 5 thru 9. Figure 5 shows the frequency response $|H(\omega)|$ for two different filters, each of which meets the requirements for a phase vocoder with N=25. In Figures 6 thru 9, these filters are used to illustrate the behavior of the magnitude and phase-derivative signals for k=1 (ie. the channel centered at 2000 Hz). Figures 6 and 8 show how the magnitude signal is affected by changes in the frequency of $x(n)$,


a. Frequency response (db) with 1201 point Blackman window

b. Frequency response (db) with 121 point Blackman window

Figure 5. Frequency response of two possible filters for N=25.

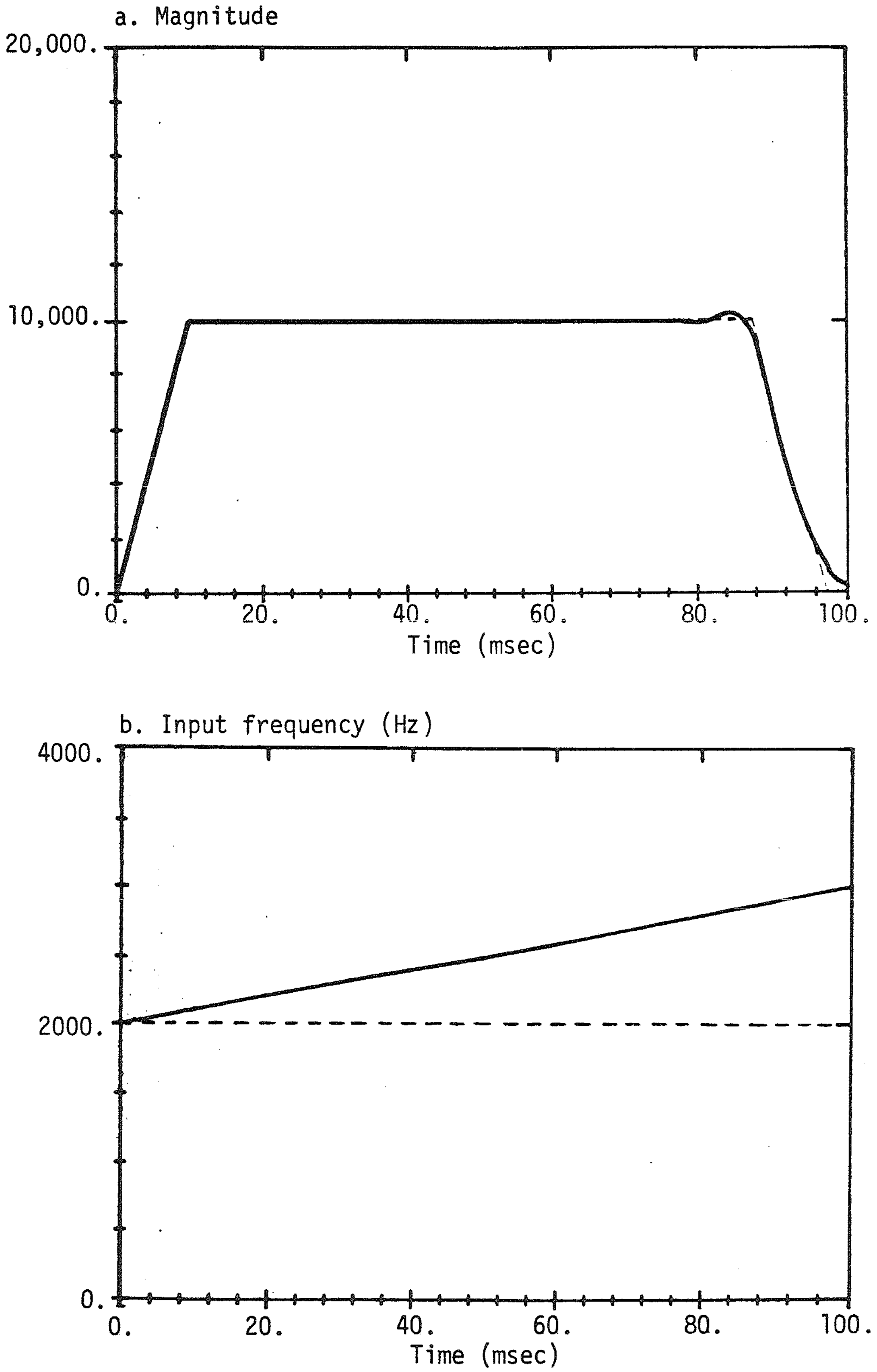


Figure 6. Magnitude signal for k=1, N=25, and filter of Figure 5a with input frequency varying as shown in (b.).

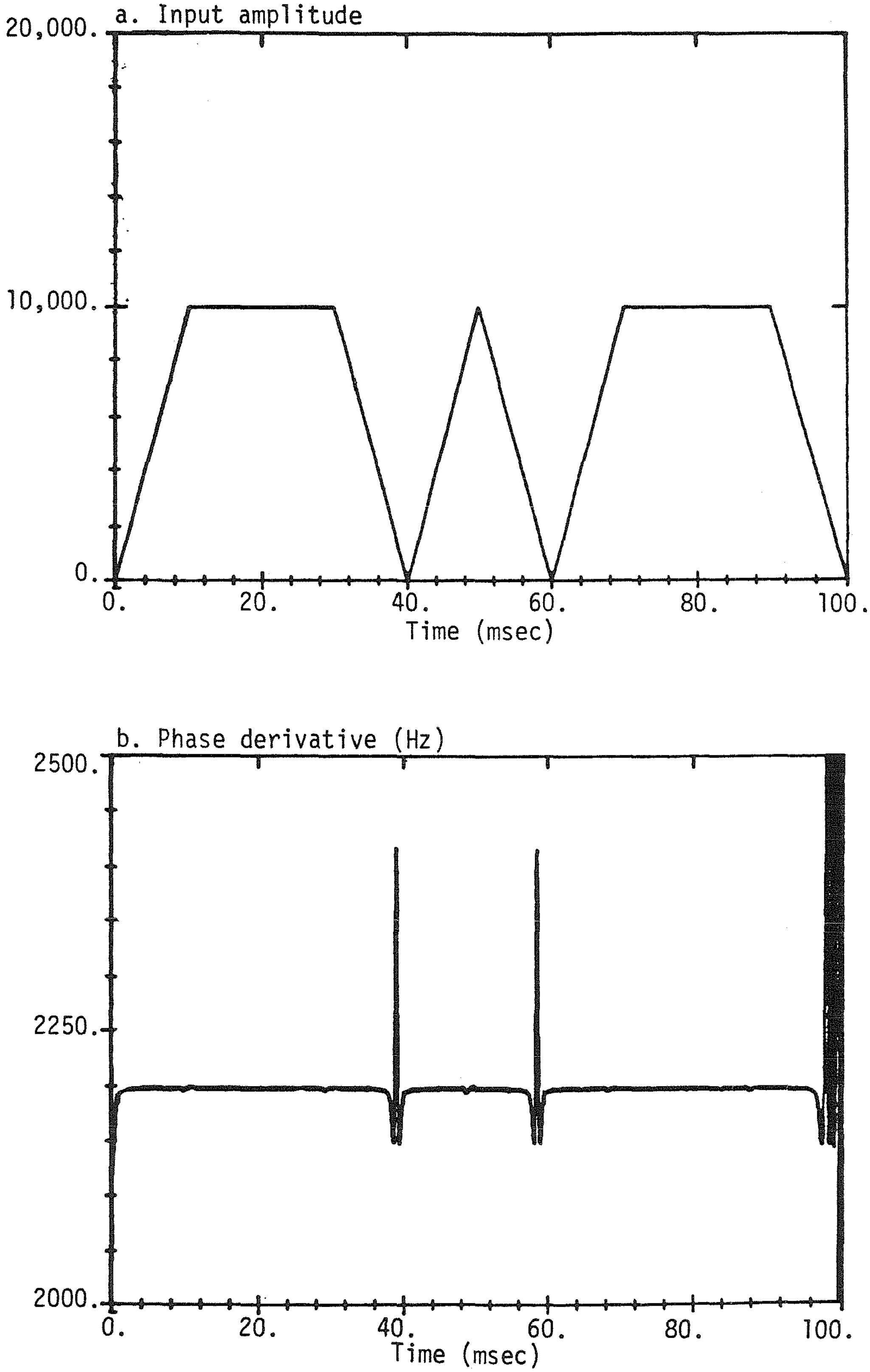


Figure 7. Phase-derivative signal for k=1, N=25, and filter of Figure 5a with input amplitude varying as shown in (a.) and input frequency of 2200 Hz.

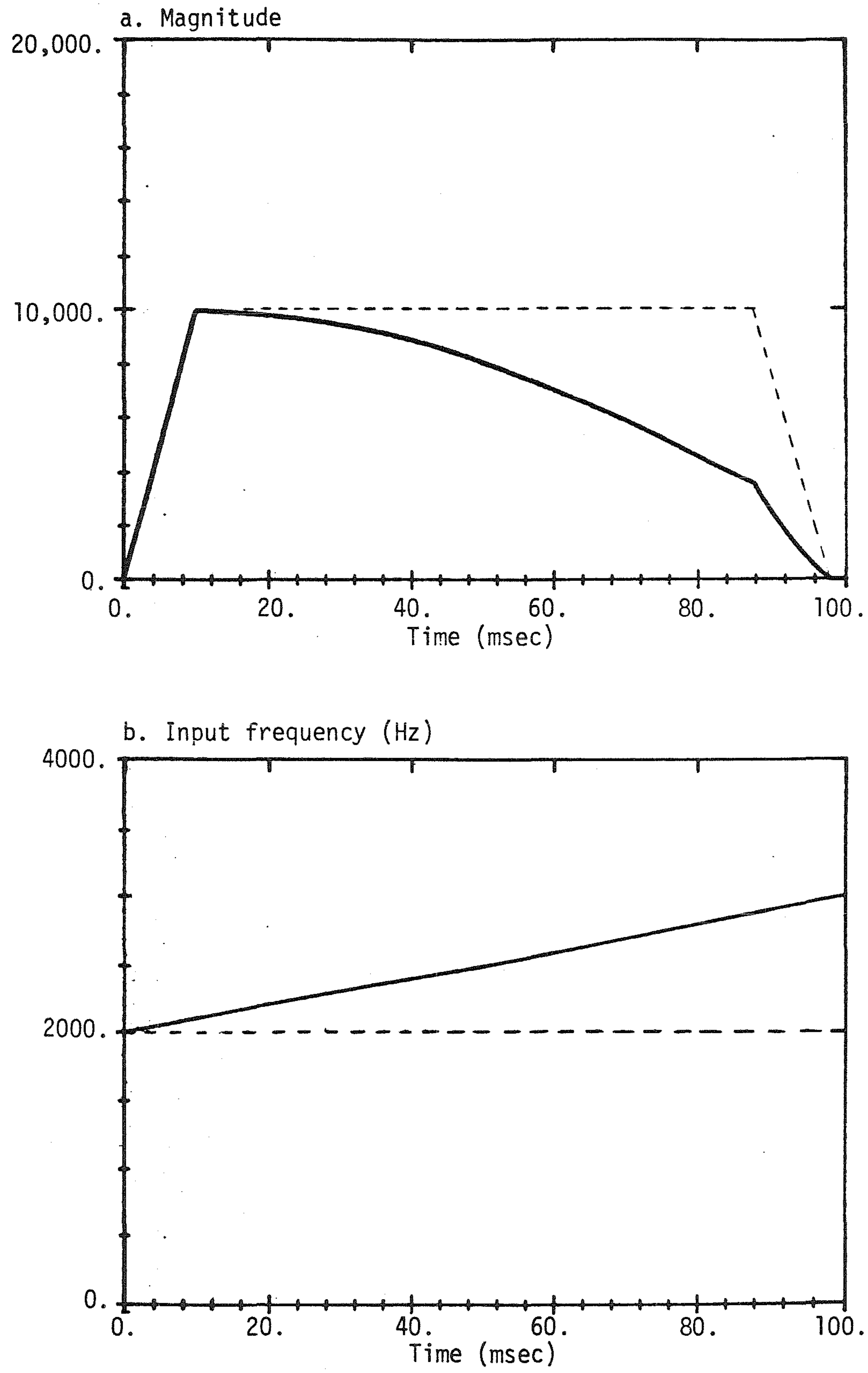


Figure 8. Magnitude signal for k=1, N=25, and filter of Figure 5b with input frequency varying as shown in (b.).

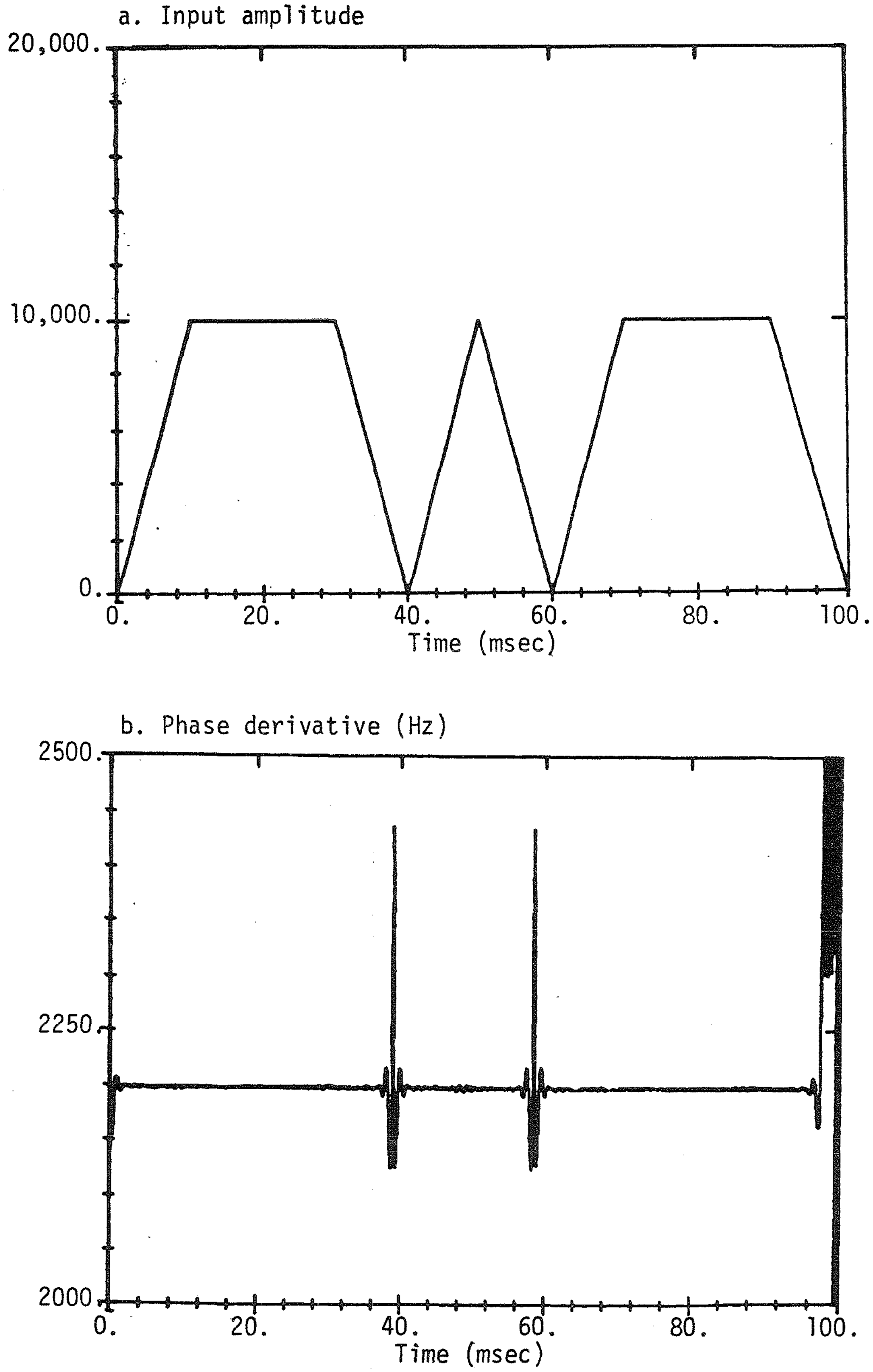


Figure 9. Phase-derivative signal for k=1, N=25, and filter of Figure 5b with input amplitude varying as shown in (a.) and input frequency of 2200 Hz.

while Figures 7 and 9 show the effect on the phase derivative of changes in the amplitude of $x(n)$. It is interesting to note that, in Figures 6 and 8, the input frequency changes by half the channel separation while only significantly distorting the magnitude for the filter with the gradual roll-off. However, in Figures 7 and 9, a frequency displacement of only $\frac{1}{10}$ the channel separation introduces severe distortions in the phase-derivative signal for either filter; but the distortions occur only at amplitude minima. It should also be noted that these examples are more severe than would normally be encountered in music or speech.

The underlying message of equations (3.44) and (3.45) is that the magnitude and phase-derivative signals are accurate estimates of the true amplitude and frequency only when the input signal is well within the filter bandpass. This is scarcely surprising, but it is nevertheless quite important. First, this observation provides a useful perspective for examining the role of the phase vocoder in modifying pitch independently of amplitude; we return to this point momentarily. Secondly, this observation is important in determining the minimum allowable sampling rate for $X(n,k)$ and a $\dot{\varphi}(n,k)$, because the distortions in these signals tend to increase their bandwidth; this issue is addressed more fully in Section 3.6. Lastly, this observation is important if the phase vocoder is to be used with the additive model to study timbre; this is a primary motivation for the tracking phase vocoder of Section 3.7.

A typical music or speech signal is actually a sum of signals of the type defined by equation (3.28) with harmonically (or almost harmonically) related frequencies. For the composite signal to be reconstructed exactly, the relative phasing of the individual harmonics (or partials) must be faithfully reproduced. This is a nontrivial task when the phase must be reconstructed from its derivative. Fortunately, the ear is generally insensitive to this relative phasing. This is also true of the phase vocoder provided that the individual partials are well within their respective filter bandwidths. Unfortunately, typical music and speech signals vary sufficiently that this condition is never satisfied for very long.

Even when a given partial is not confined to the center of any one bandpass filter, the phase vocoder can still reconstruct this partial exactly by adding in the information from adjacent frequency channels. However, this can be done only if the relative phasing of each channel is properly maintained. Otherwise, destructive interference between the different channels will severely distort the amplitude of the reconstructed partial. (An example of this is given in Figure 10 for a signal with linearly increasing frequency.) This fact has been noted before [Rabiner, 1978], but has apparently not been fully appreciated.

Since the discrete versions of differentiation and integration are not exact, the relative phasing of different frequency channels can very easily be lost. In the sampled-data domain, an ideal differentiator has the frequency response

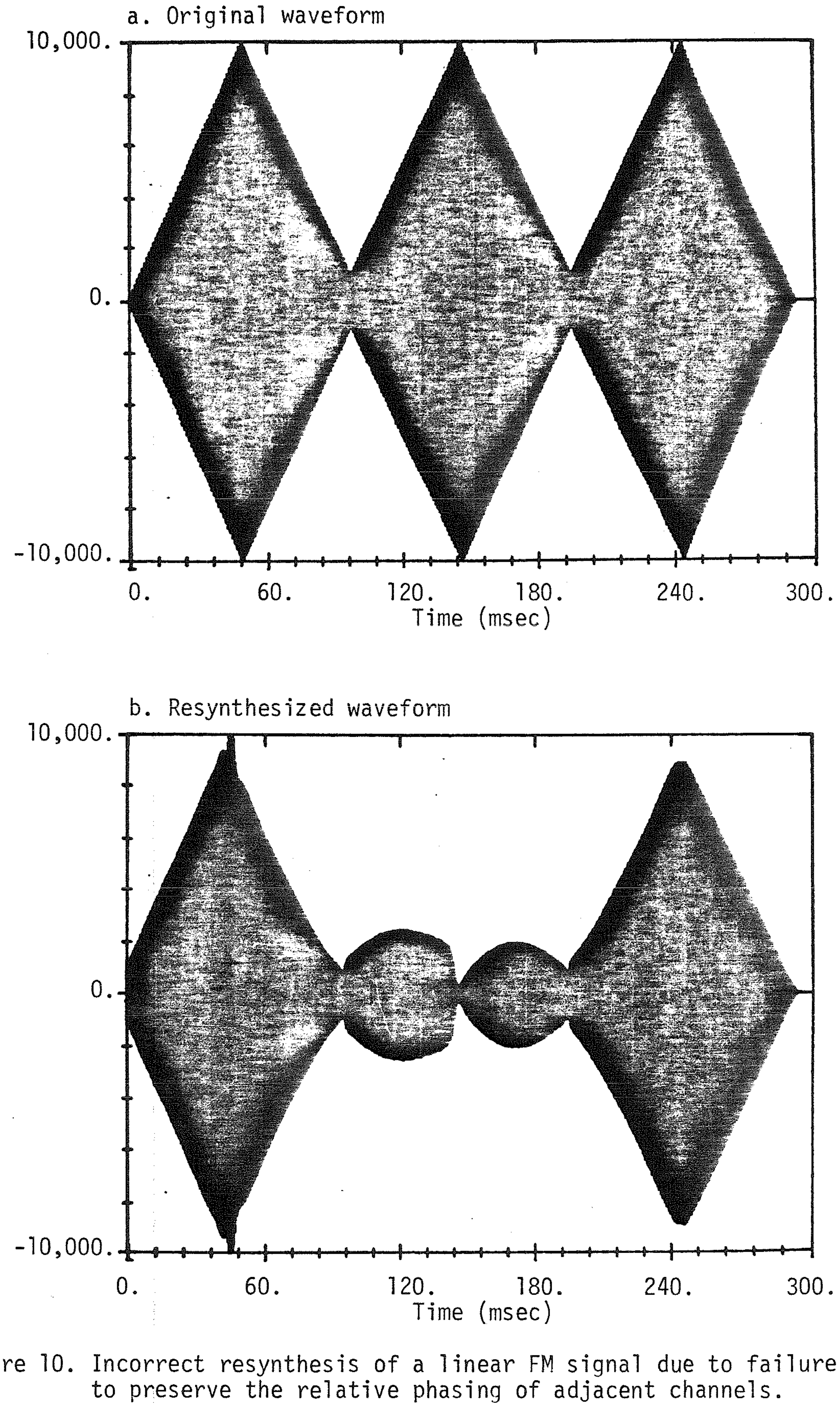
$$H(\omega) = \begin{cases} j\,\omega & 0 \le \omega \le \pi \\ j\,(2\pi - \omega) & \pi \le \omega < 2\pi \end{cases} \tag{3.46}$$

In contrast, it can easily be shown that the frequency response of a system which simply takes angle differences is

$$H(\omega) = 2j\; exp\,(-j\,\frac{\omega}{2})\; sin\,(\frac{\omega}{2}) \tag{3.47}$$

It follows that the angle difference is a good approximation to the phase derivative for small values of ω. An ideal integrator would have the frequency response

$$H(\omega) = \begin{cases} \dfrac{1}{j\,\omega} & 0 \le \omega \le \pi \\ \dfrac{1}{j\,(2\pi - \omega)} & \pi \le \omega < 2\pi \end{cases} \tag{3.48}$$



a. Original waveform

b. Resynthesized waveform

Figure 10. Incorrect resynthesis of a linear FM signal due to failure to preserve the relative phasing of adjacent channels.

But the infinite gain at DC cannot be obtained with a finite impulse response. Therefore, it is customary to perform the integration by simply accumulating the derivative values. This can be shown to result in a frequency response of

$$H(\omega) = \frac{exp\,(j\,\frac{\omega}{2})}{2jsin\,(\frac{\omega}{2})} \tag{3.49}$$

This is clearly the right thing to do for the angle differences, but not for the phase derivative except for small values of ω. Since typical music and speech signals include quite large values of ω, the use of the phase derivative is inevitably a source of error. This error is quite significant because it leads to audible interchannel interference.

In our research, we found only two reliable ways of preserving this phasing. The first of these is to use the angle differences in place of the phase-derivative signal, and to reconstruct the phase by simple addition. Even then, however, great care must be taken to avoid altering the angle differences. For example, decimation cannot be performed via the usual lowpass filtering technique; rather the successive angle differences are simply added together. Interpolation of the angle differences must also be avoided unless it is performed simply by equally subdividing each difference. The second technique avoids all of these problems by allowing a single frequency channel to track each partial individually; this is the tracking phase vocoder of Section 3.7.

We now ask whether the angle differences can be modified as in equation (3.23) without affecting the reconstructed amplitude. It is clear that the proposed modification is simply a linear transformation of the phase:

$$\theta(n,k) \rightarrow \alpha\theta(n,k) + (\alpha-1)\frac{2\pi}{N}kn \tag{3.50}$$

To examine this, we consider the $x(n)$ defined in equation (3.28) with

$\theta(n) = \beta n^2$; this can be viewed as a kind of *locally narrowband* signal. If $A(n)$ is constant, then it can easily be shown analytically that the modification in question will not alter the reconstructed amplitude. However, it is not clear whether this still holds true when $A(n)$ is varying.

To examine this further, we constructed a number of examples on the computer. For instance, Figures 11 and 12 show the magnitude and phase derivative for two adjacent channels when the input frequency increases linearly from the center of one channel to the center of the next while the input amplitude is changing severely. The individual signals are badly distorted, but as long as the relative phasing of adjacent channels was accurately reconstructed, we could detect no significant amplitude distortions in any example (Figure 13). Upon refelection, this appears reasonable because the transformation of equation (3.50) is precisely the one which keeps the phase-derivative signals of Figures 11b and 12b properly synchronized. We therefore conclude that pitch modification can in fact be accomplished while faithfully reproducing the amplitude-versus-time for each partial.

However, it does not follow from this that pitch modifications will never alter the timbre of a sound. Indeed, a mechanism by which timbre alteration can still occur will now be described.

3.5 Error sources: actual amplitude and frequency

The phase vocoder makes the tacit assumption that the signal in each filter channel is (at least over a suitable time period) a pure sine wave. Of course, the phase vocoder is still an analysis-synthesis identity system even when this condition is violated; however, its usefulness for efficient encoding and pitch modification depends heavily on this assumption. In the preceding section, we showed that the magnitude and phase-derivative signals in a given channel could

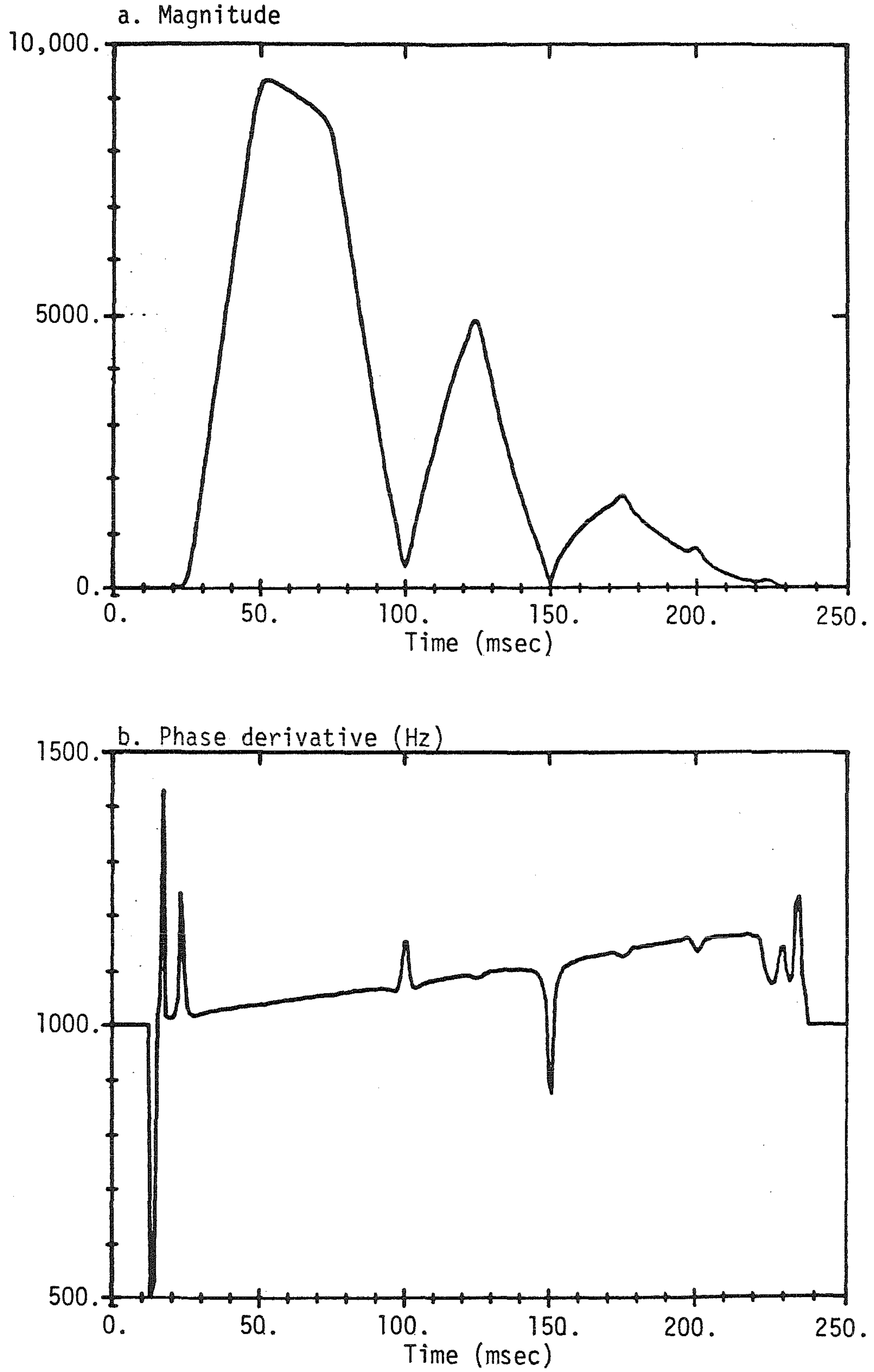


Figure 11. Magnitude and phase derivative for k=5, N=250, and filter of Figure 16 with input signal of Figure 13a. The frequency increases linearly from 1000 Hz to 1200 Hz.

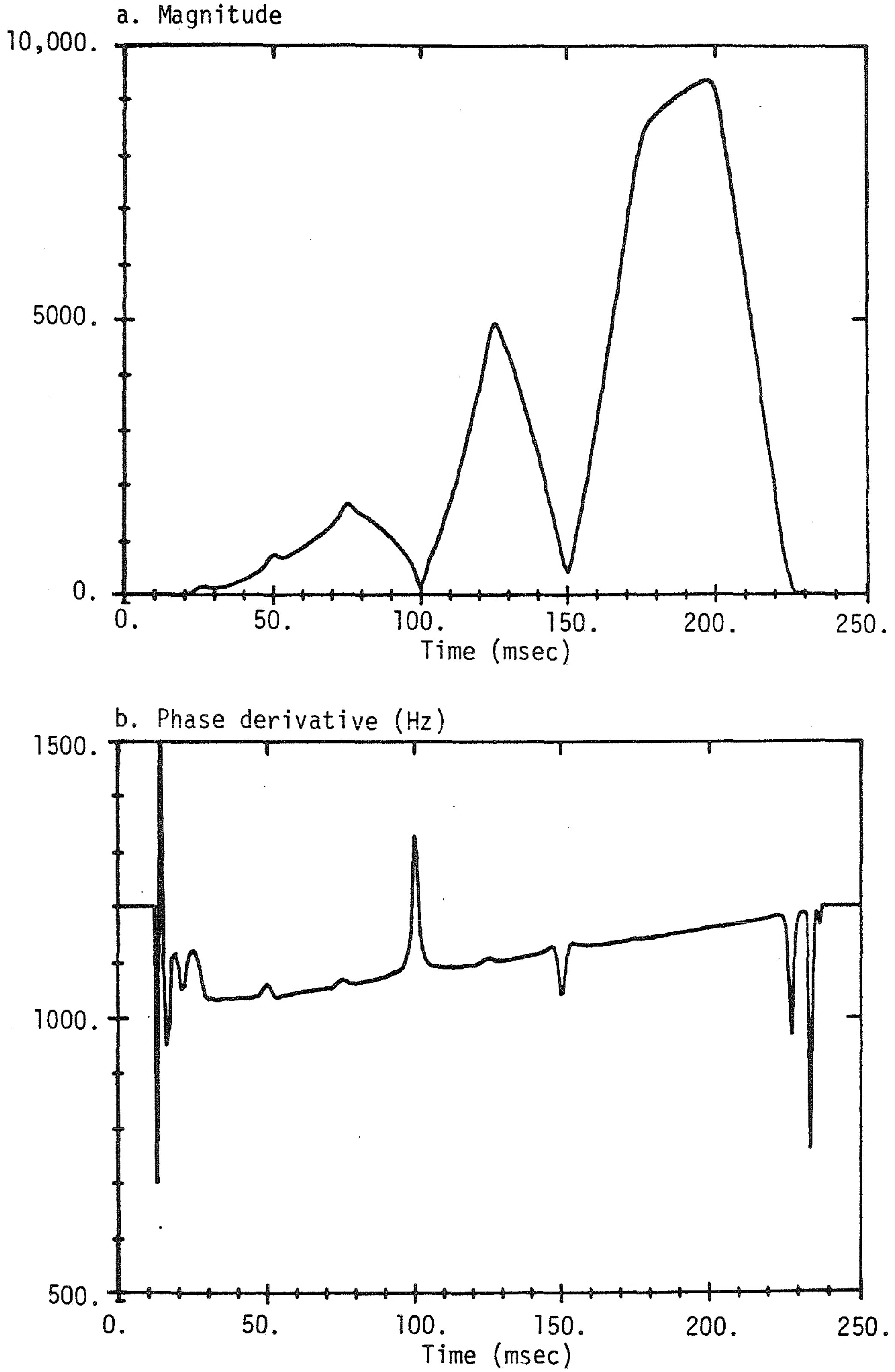
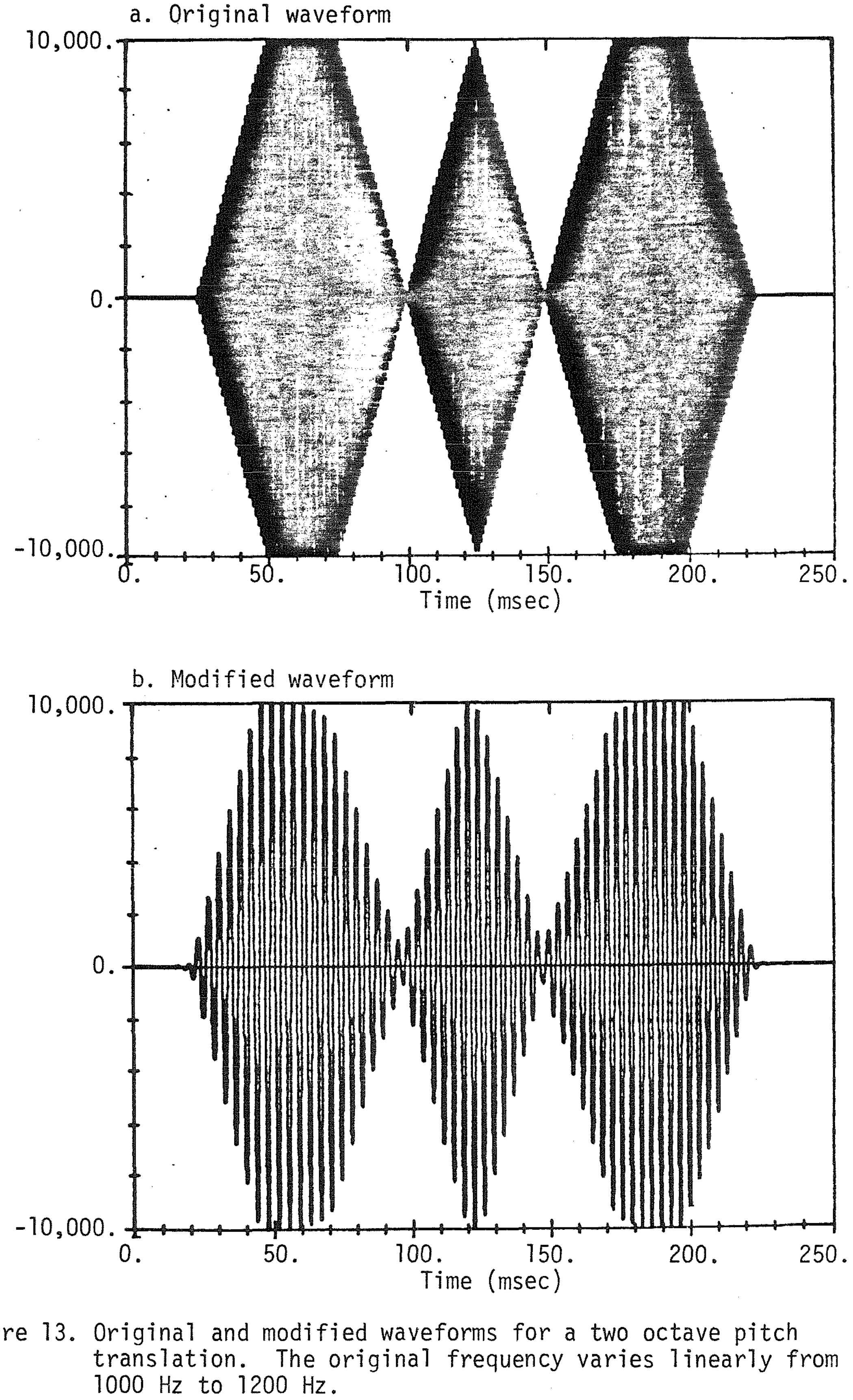


Figure 12. Magnitude and phase derivative for k=6, N=250, and filter of Figure 16 with input signal of Figure 13a. The frequency increases linearly from 1000 Hz to 1200 Hz.



a. Original waveform

b. Modified waveform

Figure 13. Original and modified waveforms for a two octave pitch translation. The original frequency varies linearly from 1000 Hz to 1200 Hz.

be distorted, even when the input to that channel was a pure sinusoid of slowly varying amplitude and phase. In this section, we examine the more general case in which the input signal is a sum of sine waves, or a sine wave plus noise, or simply noise. Again, we rely on some well known results from communication theory [Schwartz, 1966].

We begin by considering an input signal $x(n)$ consisting entirely of bandlimited Gaussian noise centered about ω_o. Such a signal can still be expressed in the form of equation (3.27) by taking $s_r(n)$ and $s_q(n)$ to be independent Gaussian random variables with a probability distribution

$$p(s) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{s^2}{2\sigma^2}) \tag{3.51}$$

where σ^2 is the expected value of $[s(n)]^2$, ie. $\sigma^2 = <s(n)s(n)>$. Equivalently, equation (3.28) can be used, in which case $A(n)$ and $\theta(n)$ are independent random variables with probability distributions

$$p(A) = \frac{1}{\sigma^2} A \; exp(-\frac{A^2}{2\sigma^2}) \tag{3.52}$$

$$p(\theta) = \frac{1}{2\pi} \qquad 0 \le \theta < 2\pi \tag{3.53}$$

These are the Rayleigh distribution and the uniform distribution, respectively.

Furthermore, the autocorrelation functions $R_r(m) = <s_r(n)s_r(n+m)>$ and $R_q(m) = <s_q(n)s_q(n+m)>$ can be expressed as

$$R_r(m) = R_q(m) = R_x(m)cos(\omega_o mT) + \hat{R}_x(m)sin(\omega_o mT) \tag{3.54}$$

where $R_x(m) = <x(n)x(n+m)>$. Taking the Fourier transform of this result and assuming that the two-sided spectral density $G_x(\omega)$ is symmetric about ω_o, it can then easily be shown that the spectral densities of $s_r(n)$ and $s_q(n)$ are

$$G_r(\omega) = G_q(\omega) = 2G_x(\omega+\omega_o) \tag{3.55}$$

Hence, we can still use the analysis of the preceding section; however, we will not attempt to calculate the autocorrelation functions for $A(n)$ and $\theta(n)$.

The above analysis is very easily extended to the case of signal plus noise. With no loss of generality, we can write

$$\begin{aligned} x(n) &= a(n)\cos(\omega_0 nT) + n(n) && (3.56)\\ &= [a(n)+n_r(n)]\cos(\omega_0 nT) + n_q(n)\sin(\omega_0 nT) && (3.57)\\ &= A(n)\cos(\omega_0 nT + \theta(n)) && (3.58)\end{aligned}$$

(The phase of the signal term is arbitrary and does not affect the result.) The new $s_r(n)$ is again a Guassian-distributed random variable, but now with mean value $a(n)$. It can then be shown that the probability distribution of $A(n)$ is

$$p(A) = \frac{1}{\sigma^2} A \, exp\left(-\frac{A^2+a^2}{2\sigma^2}\right) I_0\left(\frac{aA}{\sigma^2}\right) \tag{3.59}$$

where $I_0(z)$ is the zeroth order modified Bessel function of the first kind:

$$I_0(z) = \frac{1}{2\pi}\int_0^{2\pi} exp(z \cos(\theta))d\theta \tag{3.60}$$

The probability distribution of the phase is approximately

$$p(\theta) = \sqrt{\frac{a}{\pi\sigma^2}} \cos(\theta)\, exp\left(-\frac{a\sin^2(\theta)}{\sigma^2}\right) \qquad |\theta| < \frac{\pi}{2} \tag{3.61}$$

provided that the signal-to-noise ratio is sufficiently large [Cahn, 1960]. Alternatively, we can simply view the signal-plus-noise case as a subset of the case to follow.

We now consider the case where more than one sine wave occupies a given frequency channel. This occurs when the filter bandwidth is greater than the fundamental frequency, or when more than one voice is sounding, or when reverberation or noise is present. If the individual sine waves are

$$x_1(n) = a_1(n)\cos(\omega_0 nT + \theta_1(n)) \tag{3.62}$$

$$x_2(n) = a_2(n)cos(\omega_o nT + \theta_2(n)) \tag{3.63}$$

then it can easily be shown that their sum $x(n) = x_1(n) + x_2(n)$ has an amplitude and phase given by

$$A(n) = [a_1^2(n) + 2a_1(n)a_2(n)cos(\theta_2(n)-\theta_1(n)) + a_2^2(n)]^{\frac{1}{2}} \tag{3.64}$$

$$\theta(n) = atan\left[\frac{-a_2(n)sin(\theta_2(n)-\theta_1(n))}{a_1(n) + a_2(n)cos(\theta_2(n)-\theta_1(n))}\right] \tag{3.65}$$

Typically, $\theta(n)$ varies linearly with time at a rate determined by the instantaneous frequency. Now, if $a_1(n) = a_2(n)$, then

$$A(n) = |a_1(n)\cos\left[\frac{\theta_2(n) - \theta_1(n)}{2}\right]| \tag{3.66}$$

$$\theta(n) = \frac{\theta_2(n)+\theta_1(n)}{2} + \Theta(n) \tag{3.67}$$

where $\Theta(n)$ equals 0 or π depending on the sign of $cos(\frac{\theta_2(n)-\theta_1(n)}{2})$. This shows that the instantaneous frequency of the sum is the average of the instantaneous frequencies except for occasional spikes due to the phase flipping. On the other hand, if $a_1(n) \gg a_2(n)$, then

$$A(n) \approx a_1(n)[1 + \frac{2a_2(n)}{a_1(n)}cos(\theta_2(n)-\theta_1(n))]^{\frac{1}{2}} \tag{3.68}$$

$$\approx a_1(n) + a_2(n)cos(\theta_2(n)-\theta_1(n)) \tag{3.69}$$

$$\theta(n) \approx atan\left[\frac{a_2(n)}{a_1(n)}sin(\theta_2(n)-\theta_1(n))\right] \tag{3.70}$$

$$\approx -\frac{a_2(n)}{a_1(n)}sin(\theta_2(n)-\theta_1(n)) \tag{3.71}$$

which shows that a small sine wave introduces a simultaneous amplitude and frequency modulation to a larger sine wave which occupies the same channel. Furthermore, the frequency of the modulation is simply the difference between the two individual frequencies. While the above results are quite easily obtained, they have important implications for the phase vocoder which have not previously been fully appreciated.

Equations (3.66) and (3.67) are important because they show that destructive interference of two pure sine waves can lead to phase flips of 180 degrees and corresponding spikes in the phase-derivative signal. A further example of this is given in Figures 14 and 15. In both cases, it is clear that the phase-derivative spike is merely an artifact of the analysis technique; a narrow band filter centered at the peak frequency of the spike would not detect any energy at any time.

Furthermore, we conducted tests in which a pair of sine waves were analyzed and resynthesized both with and without phase flipping (ie., both without and with suppression of spikes). We found that the ear was utterly incapable of distinguishing between the two; this is not surprising in that the phase flips occur only at amplitude nulls. Hence, we conclude that the phase flips within a given channel of the phase vocoder are not perceptually significant. Again we note, however, that the phase signal in each channel must be accurately reproduced to avoid crosstalk between adjacent channels upon resynthesis.

Equations (3.68) thru (3.71) also describe a situation which is quite common in the analysis of musical sound. In essence, all of these equations are saying that the filter bandwidths are too wide (i.e., there are too few frequency channels) to separate the signal into sinusoidal components. Consequently, the composite amplitudes and frequencies include modulation terms which tend to increase the bandwidths of $|X(n,k)|$ and $\varphi(n,k)$. In addition, the composite amplitudes and frequencies have pitch information inextricably linked with

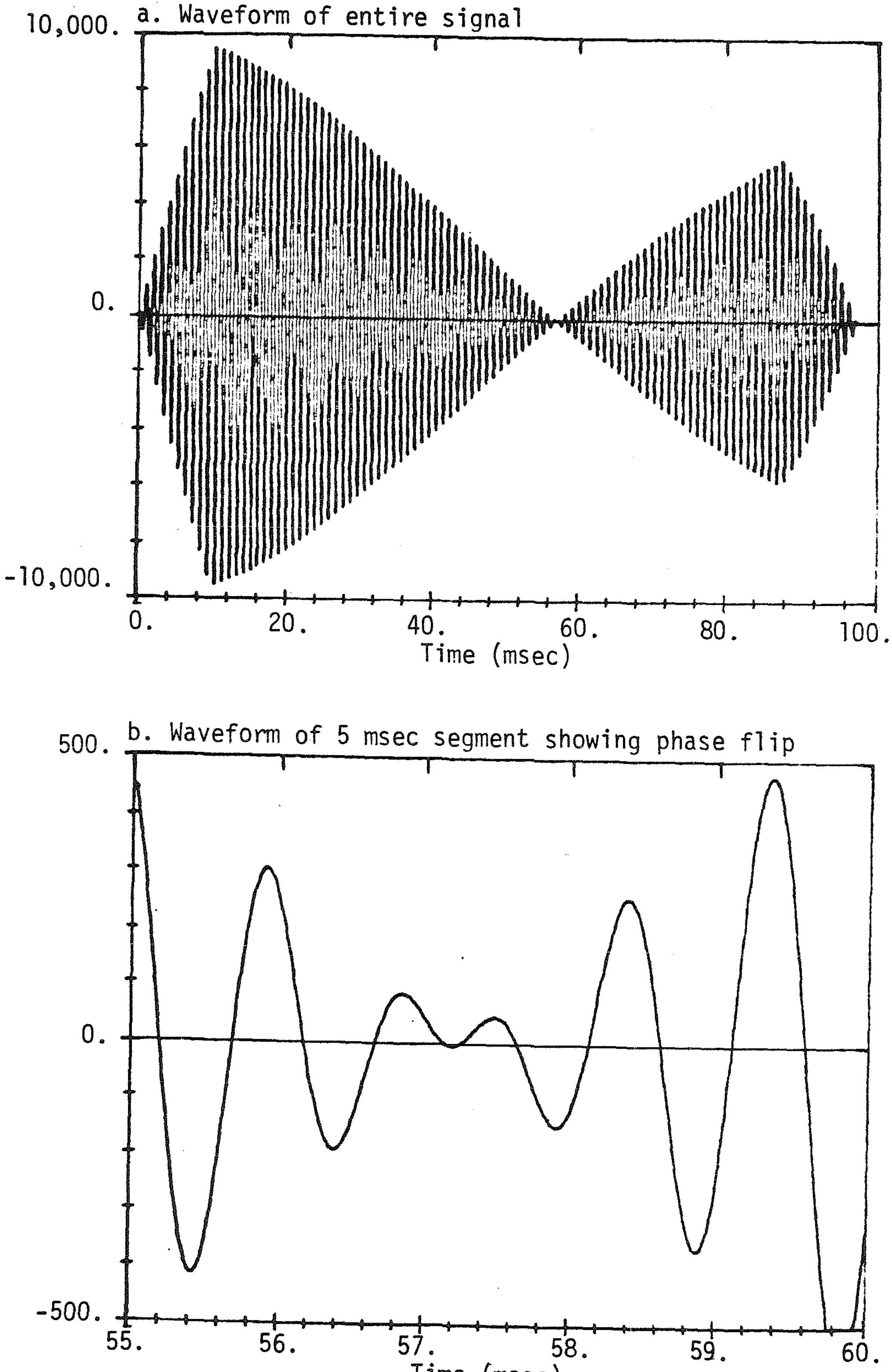
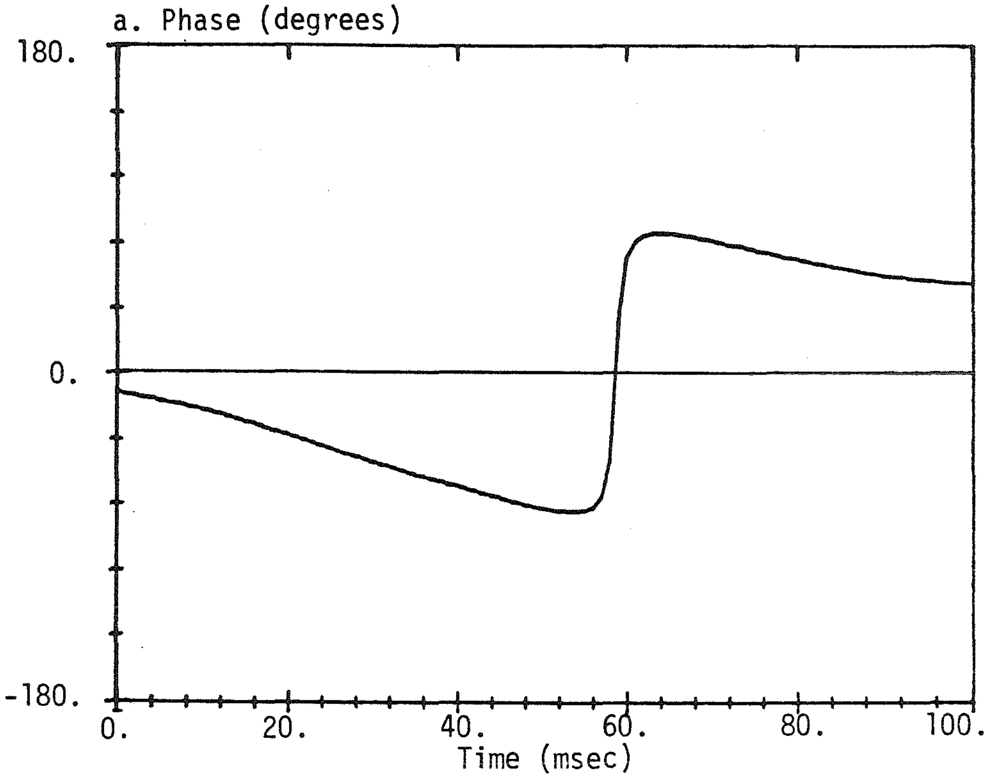


Figure 14. Phase flipping in the sum of a 1000 Hz sine wave and a 990 Hz to 1000 Hz linear FM sine wave.



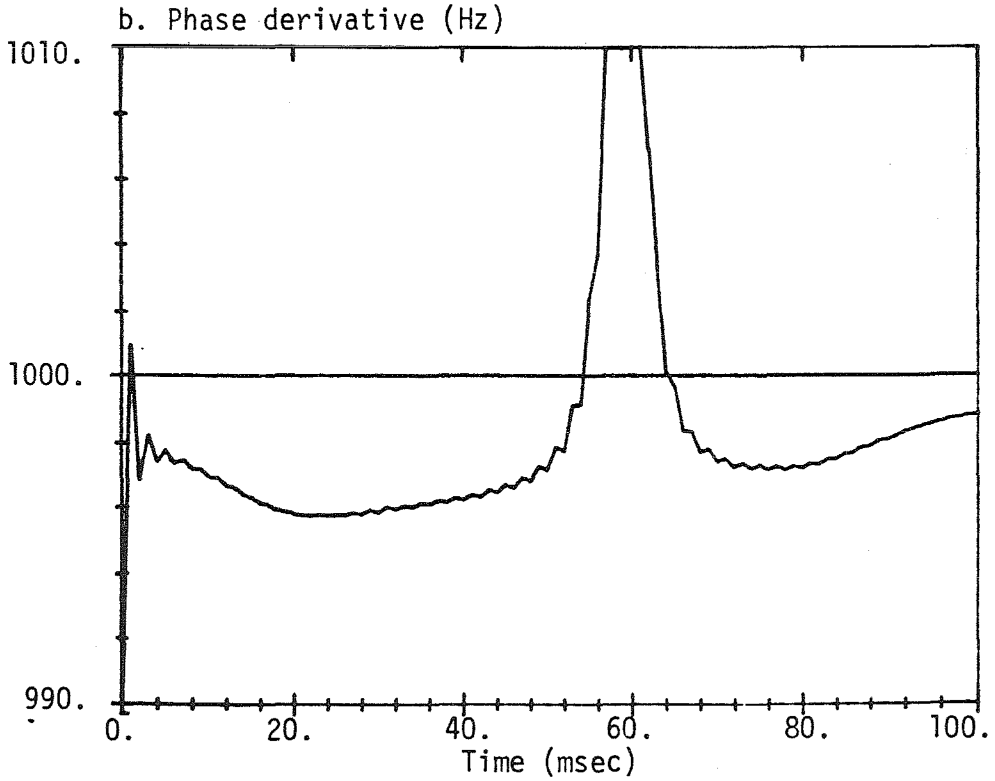


Figure 15. Phase and phase derivative for k=5, N=250, and filter of Figure 16 with input signal of Figure 14.

temporal variations. As a result, pitch modifications of the type defined by equation (3.23) can have undesirable effects on timbre.

For example, suppose that the actual signal has components at 500 Hz and 600 Hz (an interval of a *minor third* in musical terminology). Translating this signal downward by one octave results not in components at 250 Hz and 300 Hz as desired, but rather in components at 225 Hz and 325 Hz (an interval of nearly a *fifth*). Even in music involving only a single voice, this effect still occurs (albeit on a smaller scale) because of sympathetic vibration of undamped strings, reverberation, noise, etc. This suggests that the filter bandwidths should be made as narrow as possible. However, some amount of undesirable modulation appears to be inevitable in any practical phase vocoder.

It is also worth noting that even in a phase vocoder which does have only a single sine wave in any given filter bandpass, there is still a potential for timbre modification when pitch translation is performed. This occurs because modifiying the pitch also modifies the center frequencies of any broad resonances or *formants* which may exist. In the case of speech, this modification can be quite objectionable.

Lastly, we consider an extension of the above analysis which is relevant to Chapter 4 to follow. This is the case of M independent sine waves within a given filter bandwidth (i.e., an ensemble). With a good deal of algebra, it can be shown that

$$A(n) = \left|\left[\sum_{l=1}^{M} a_l^2(n) + \sum_{l=1}^{M} \sum_{m \neq l} a_l(n) a_m(n) \cos(\theta_l(n) - \theta_m(n))\right]^{\frac{1}{2}}\right| \qquad (3.72)$$

$$\theta(n) = atan\left(\frac{\sum_{l=1}^{M} a_l(n) \sin(\theta_l(n))}{\sum_{l=1}^{M} a_l(n) \cos(\theta_l(n))}\right) + \Theta(n) \qquad (3.73)$$

where $\Theta(n)$ again equals 0 or π depending on the sign of $A(n)$ prior to taking the absolute value in equation (3.68). If $\theta_l(n) = \beta_l nT + \theta_l(0)$, then the amplitude has a sum of modulating terms with the most rapid variation corresponding to the most widely separated sine waves within the filter bandwidth. In this case, it is also instructive to examine the phase-derivative. Recalling that $\frac{d}{dt}\, atan(u) = \frac{1}{1+u^2}\frac{du}{dt}$, it can be shown that (equation 3.74)

$$\dot{\theta}(n) = \frac{\sum\limits_{l=1}^{M} a_l^2(n)\beta_l + \sum\limits_{l=1}^{M}\sum\limits_{m\neq l} a_l(n)a_m(n)\beta_m \cos((\beta_l-\beta_m)nT+\theta_l(0)-\theta_m(0))}{\sum\limits_{l=1}^{M} a_l^2(n) + \sum\limits_{l=1}^{M}\sum\limits_{m\neq l} a_l(n)a_m(n)\cos((\beta_l-\beta_m)nT+\theta_l(0)-\theta_m(0))}$$

This indicates a frequency modulation with at least a qualitative similarity to the amplitude modulation in equation (3.72).

3.6 Bandwidths of the magnitude and phase-derivative signals

The fact that the magnitude and phase signals are not strictly bandlimited has troubled investigators from the time that the phase vocoder was first conceived. The phase-derivative signal $\dot{\varphi}(n,k)$ presents an even worse case. While the differentiation of the phase signal does not actually increase its bandwidth, it does multiply the spectrum by ω and thus enhances the higher frequency components. Flanagan [1980] showed that this problem could be partially circumvented by using the signals $|X(n,k)|^2$ and $|X(n,k)|^2\dot{\varphi}(n,k)$. It can be shown that these signals are bandlimited by the filter $h(n)$ and therefore can be sampled at the same rate as $a(n,k)$ and $b(n,k)$. However, it is still necessary to obtain $\dot{\varphi}(n,k)$ initially, and this must be done at a much higher sampling rate.

Moorer [1978] reported that the non-bandlimited nature of the conversion to magnitude and phase made it necessary to calculate these signals at the original input sampling rate. We have already noted (Section 3.3) that this results in a tremendous increase in computation load. Furthermore, there is no *a priori* reason to be satisfied with the input sampling frequency; a particularly wideband signal might require interpolation to an even higher sampling rate before it could be computed without aliasing. In this section, however, we show that useful magnitude and phase signals can be calculated at the same sampling rate as $a(n,k)$ and $b(n,k)$, provided that proper care is taken.

We begin by considering the magnitude signal. It was shown in Section 3.4 that the magnitude signal is bandlimited to the filter bandwidth except for two perturbing influences. Furthermore, in the case of sine wave inputs, these perturbations can generally be ignored. If the instantaneous frequency is centered within the filter bandpass, then the only perturbation is the absolute value in equation (3.44); this is significant only when the actual amplitude is switched on or off very suddenly. If the instantaneous frequency is not centered within the filter bandpass, then a rapid change in frequency can significantly increase the bandwidth of the magnitude signal; but this rarely occurs. An important example of this bandlimiting is given in Figure 17, in which the response of the magnitude signal to a step change in amplitude is seen to be very nearly the step response of the filter.

For more complex narrow band signals such as those in the preceding section, the above argument collapses due to the rapid changes in phase. It is still possible to show bandlimiting in the case of two sine waves with slightly different frequencies, but the more general case of signal-plus-noise is analytically intractable. Consequently, we are forced to determine the bandwidth of the magnitude signal numerically for some representative cases. In the experiments to follow, we use the filter of Figure 16 when the channel

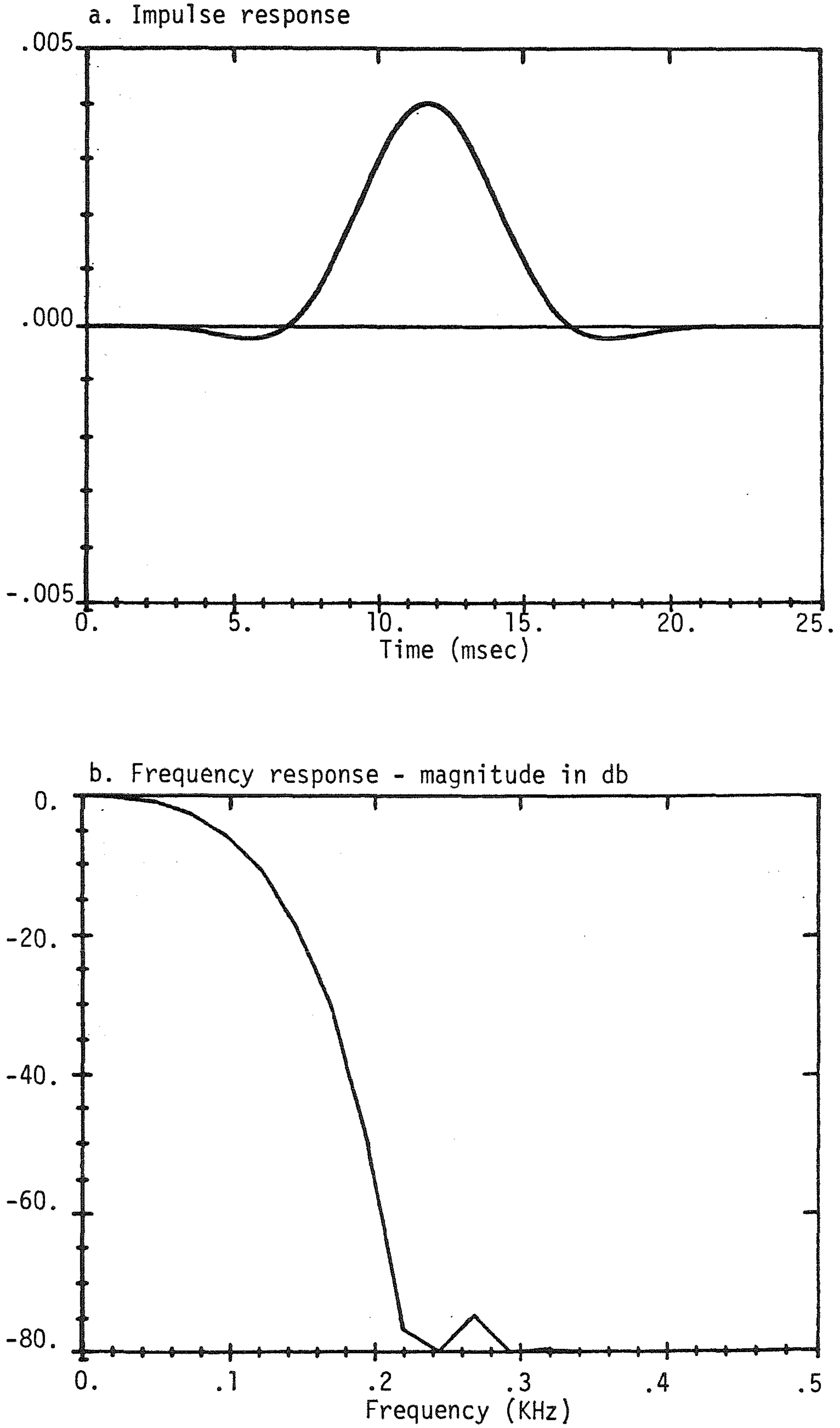


Figure 16. Impulse and frequency response for filter with N=250 and 1201 point Blackman window.

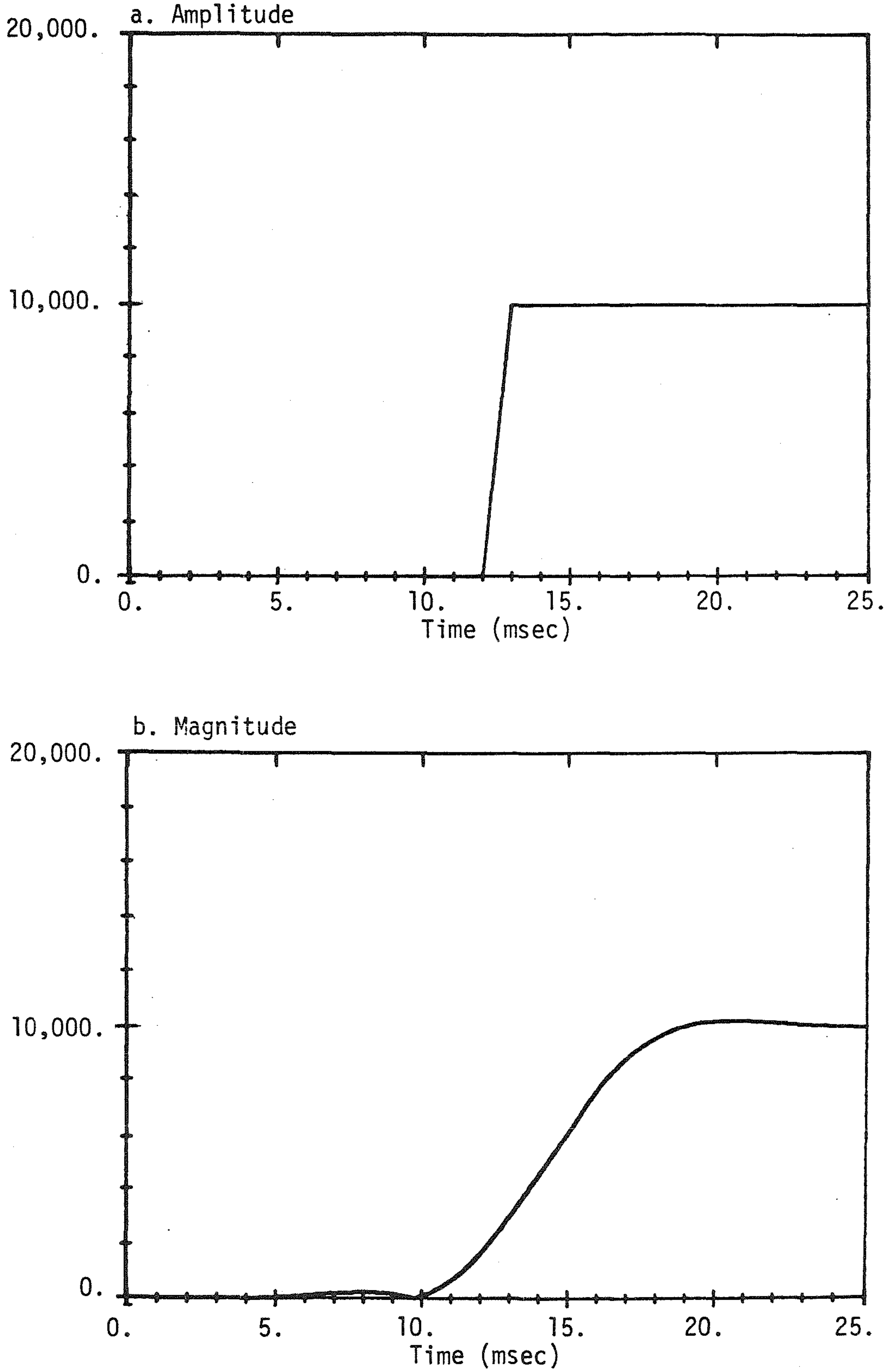


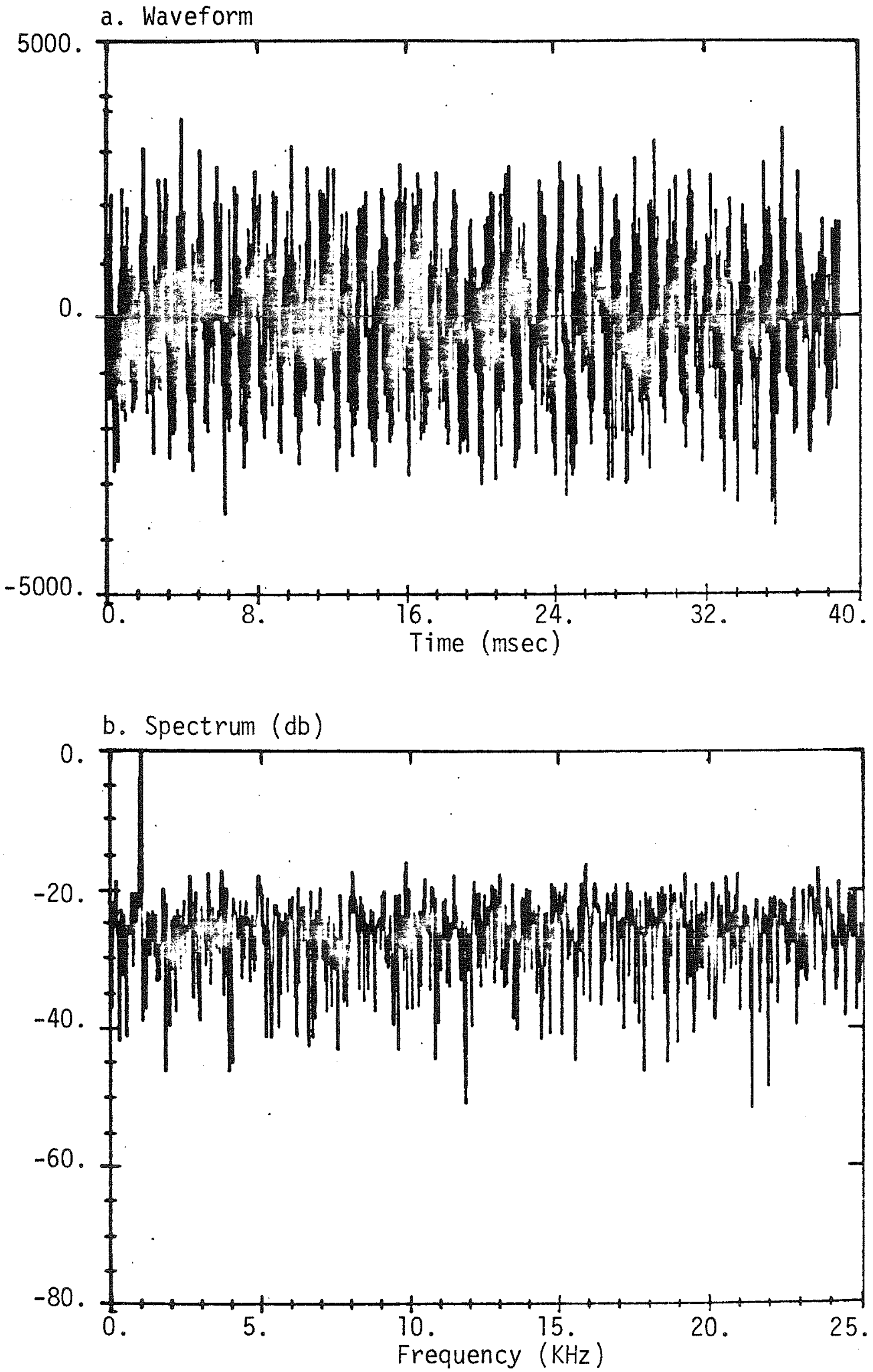
Figure 17. Step response of the magnitude signal with the filter of Figure 16. The input amplitude is shown in (a.).

separation is 200 Hz and the filters of Figure 5 when the separation is 2000 Hz. This provides a means of realistically examining narrowband signals while also comparing the effects of different filters which cannot always be implemented with narrow bandwidths. To determine the approximate bandwidth of the magnitude signal, we apply a 3-term Blackman-Harris window [Harris, 1978] and take a 2048 point Fast Fourier Transform, using zero fill where appropriate.

Figure 18 shows the waveform and spectrum of a signal which is actually the sum of a 1 KHz sine wave and white Gaussian noise. The noise was generated by summing independent samples from a uniform distribution as described in Rabiner and Gold [1975]. Figure 19 shows the magnitude signal obtained from the phase vocoder with the filter of Figure 16 centered at 1 KHz. The spectrum in Figure 19b makes it clear that the magnitude signal in this case is still effectively bandlimited by the filter.

A more demanding example is provided by the white Gaussian noise signal of Figure 20. Indeed, we consistently found this case to be the one in which the magnitude signal attained its greatest bandwidth. The exact value of this bandwidth depends upon which particular definition is adopted, but we did observe two consistent trends. First, the spectrum of the magnitude signal depended surprisingly little on the sharpness of the filter cutoff; this is illustrated in Figures 21 and 22 for a center frequency of 2000 Hz and the two filters of Figure 5. Secondly, the bandwidth of the magnitude signal (as measured by the -40 db point) was usually less than the channel separation. Occasionally, however, it was as much as twice the separation.

We did not conduct any tests of the perceptual effects of undersampling, but we conjecture that such effects would be least evident precisely for those instances in which the magnitude signal attains its greatest bandwidth - i.e., the



a. Waveform

b. Spectrum (db)

Figure 18. Waveform and spectrum of 1 KHz sine wave plus white Gaussian noise.

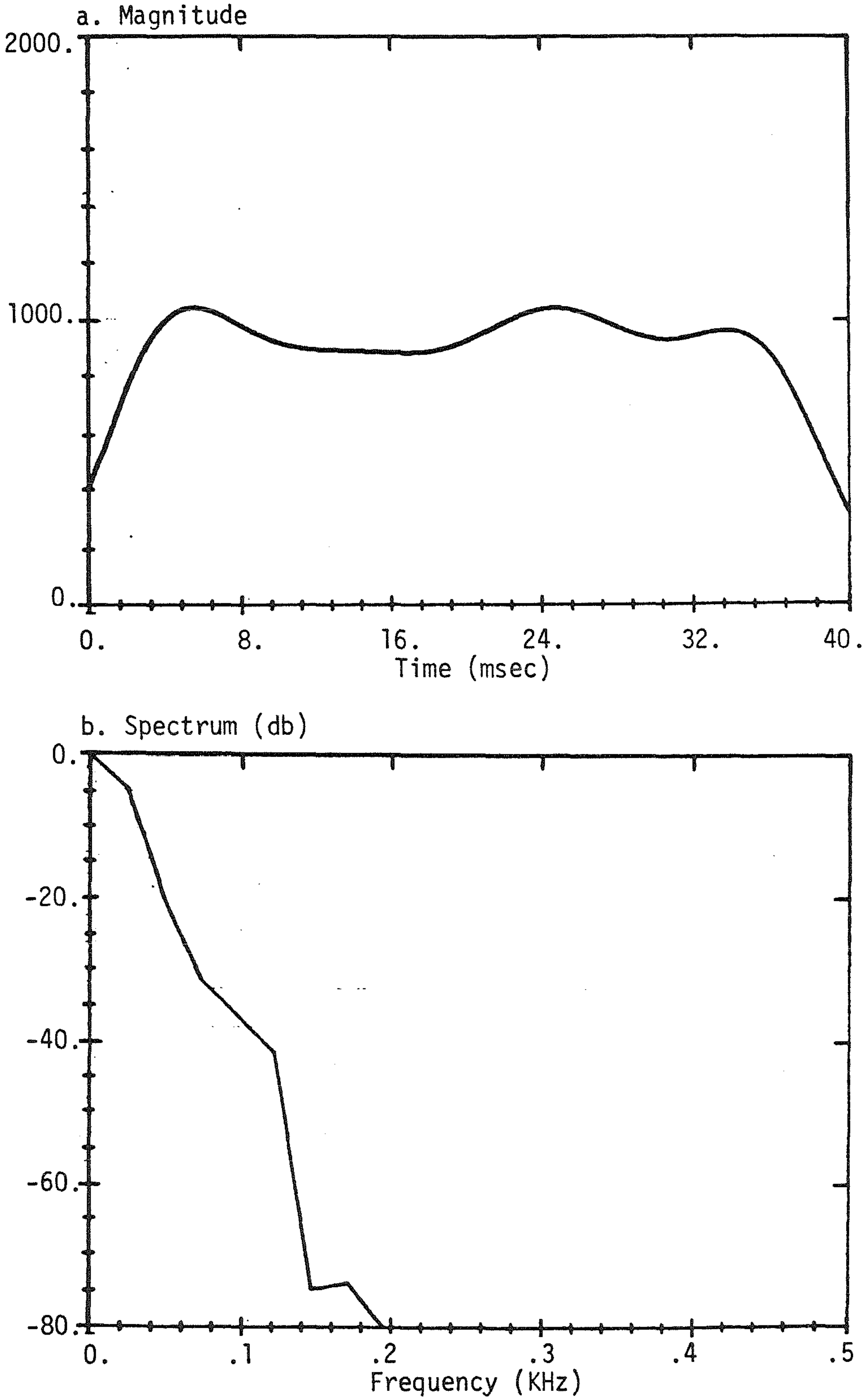


Figure 19. Magnitude for k=5, N=250, and filter of Figure 16 with input signal of Figure 18. The spectrum of the magnitude signal is shown in (b.).

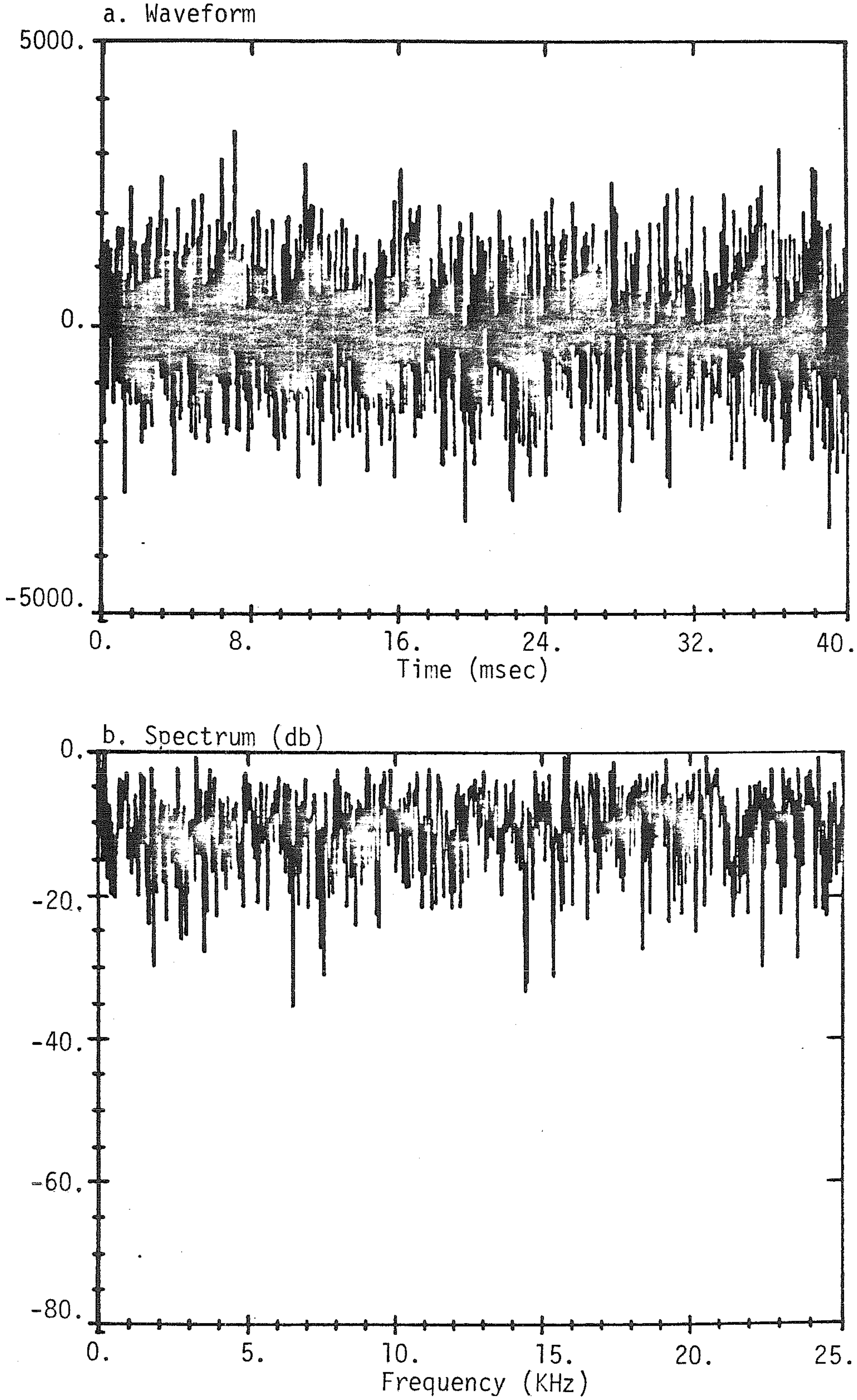


Figure 20. Waveform and spectrum of white Gaussian noise.

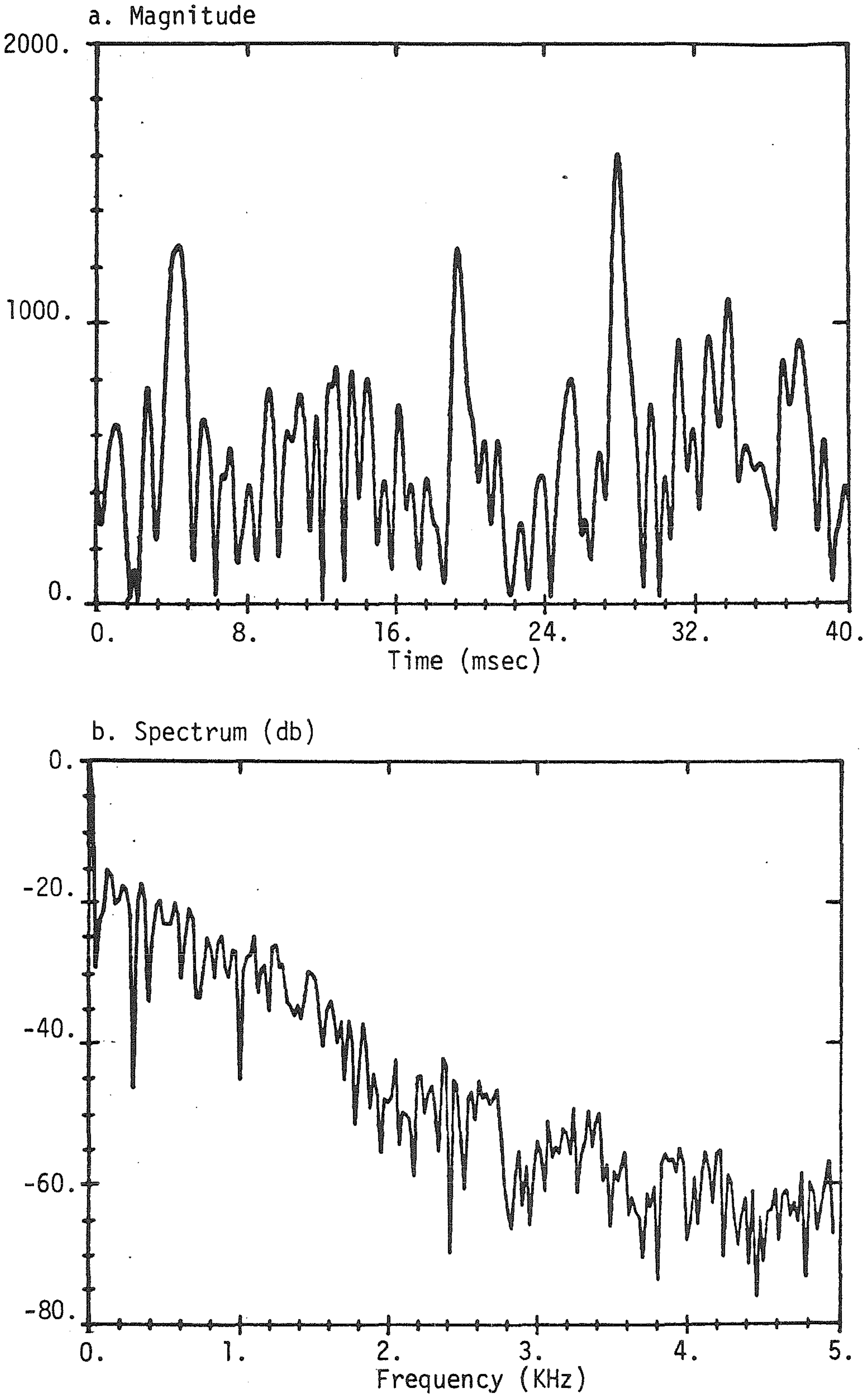


Figure 21. Magnitude for k=1, N=25, and filter of Figure 5a with input signal of Figure 20. The spectrum of the magnitude signal is shown in (b.).

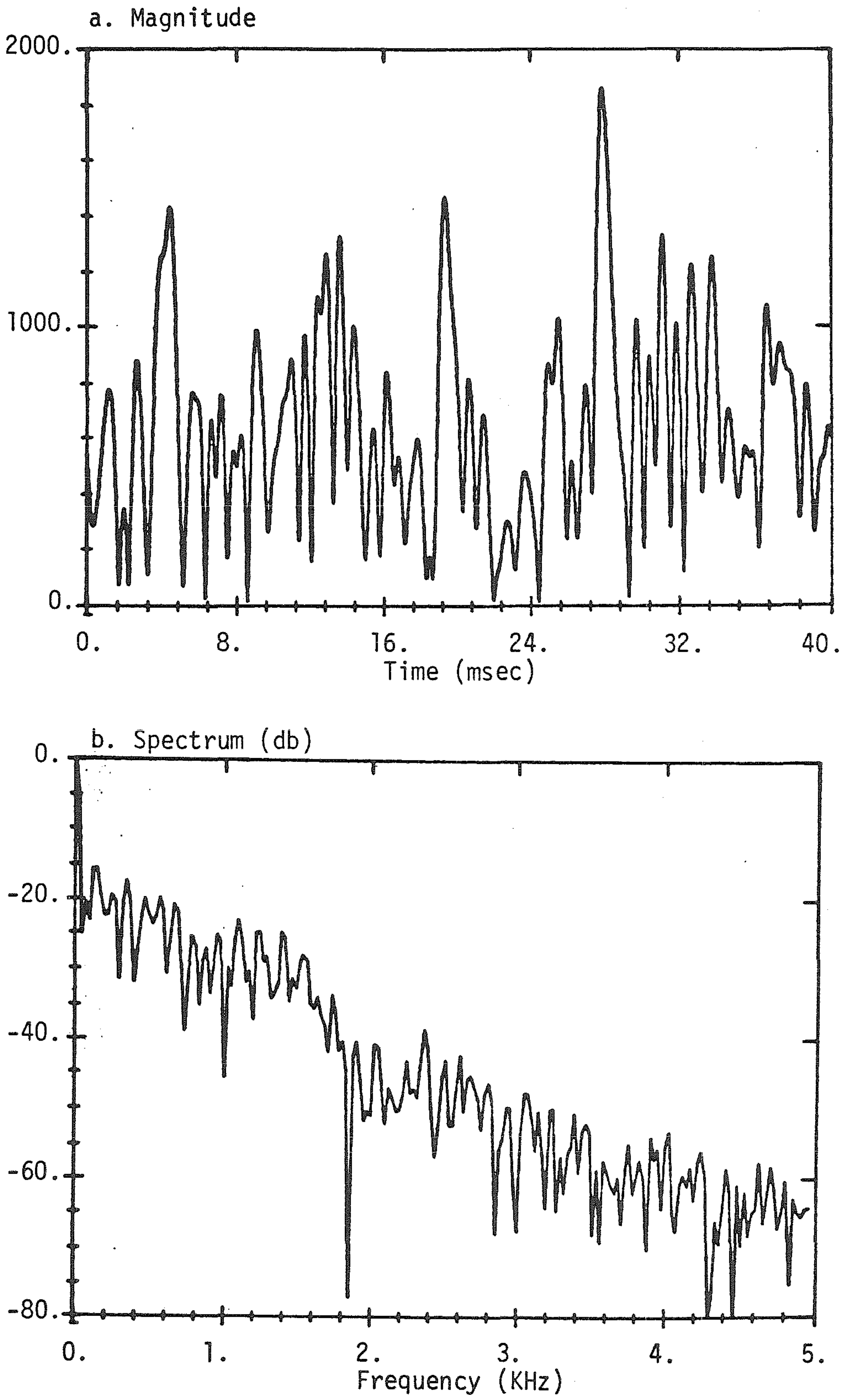


Figure 22. Magnitude for k=1, N=25, and filter of Figure 5b with input signal of Figure 20. The spectrum of the magnitude signal is shown in (b.).

case of noise alone. Consequently, we found it sufficient to sample the magnitude signal at twice the channel separation; this is the same sampling rate as is typically required for $a(n,k)$ and $b(n,k)$ with a nonideal filter.

We now consider the phase-derivative signal. For a pure sinusoid, the most rapid possible change in phase corresponds to the highest frequency within the filter bandwidth. Hence, it suffices to sample the phase at twice the filter bandwidth. This bandlimiting can also be easily demonstrated in the case of simple frequency modulation. The frequency modulation introduces sidebands in the input signal $x(n)$ with a spacing equal to the modulating frequency. When this frequency becomes too large, the sidebands fall outside the filter bandpass, and the signal within the filter becomes a sine wave of constant frequency. But in the case of more complex signals, we again must rely on numerical examples to form our conclusions.

We first examine the response of the phase-derivative signal to a step change in frequency (assuming that both initial and final frequencies are well within the filter bandpass). This is an important and nontrivial case. It turns out that - as with the magnitude signal - the step response of the phase derivative is very nearly that of the filter; this is clearly illustrated in Figure 23 in which the filter of Figure 16 is centered at 1 KHz. This indicates that the phase-derivative signal is also bandlimited by the filter with regard to step changes in frequency.

The bandlimiting due to the filter is also evident in the case of signal-plus-noise (Figure 24). We therefore turn directly to the case of noise alone. Figure 25 shows the phase-derivative signals corresponding to the magnitude signals of Figures 21a and 22a. The most obvious features are the ubiquitous spikes occurring (as predicted in Section 3.5) at the minima of the magnitude signals.

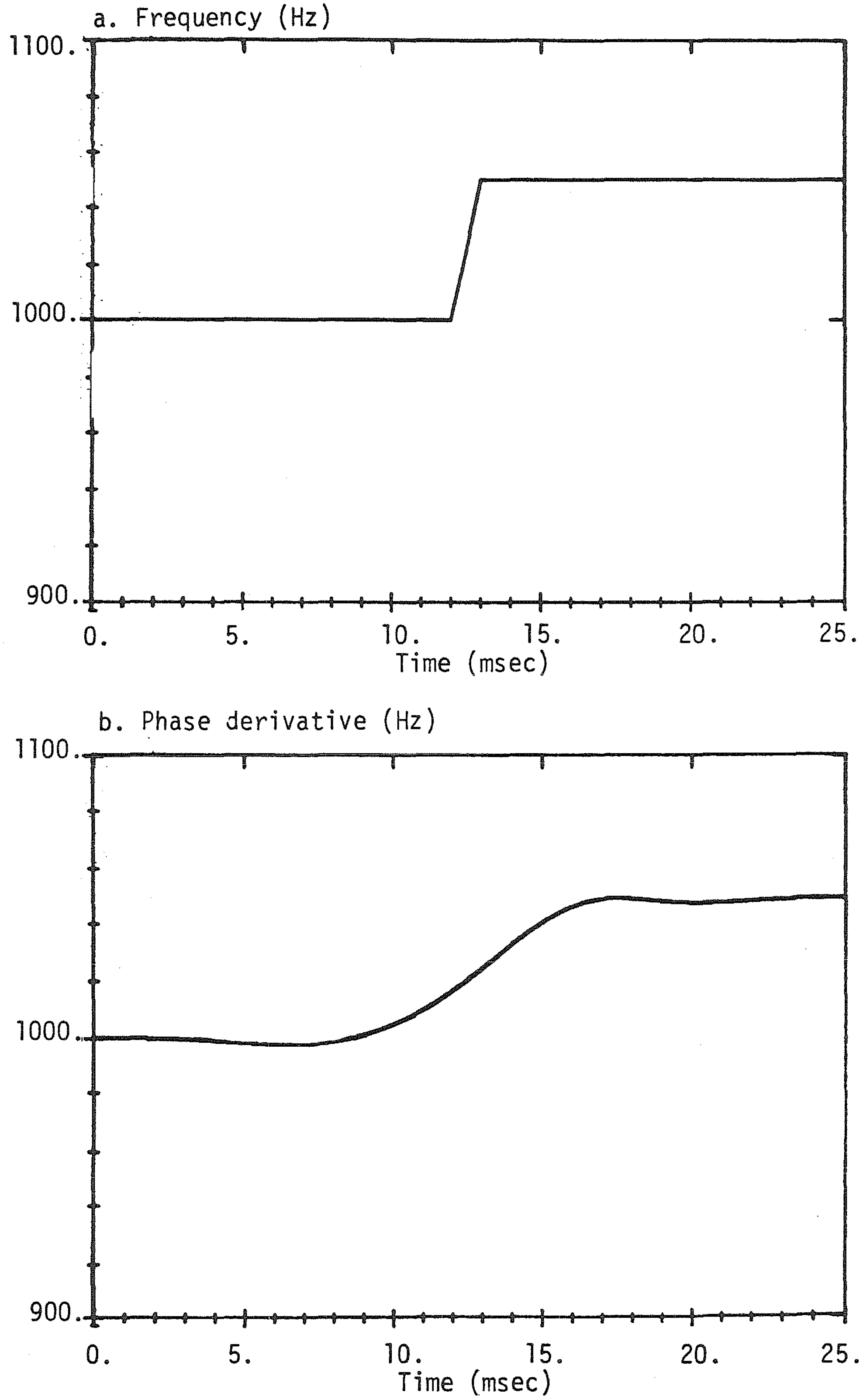


Figure 23. Step response of the phase-derivative signal with the filter of Figure 16. The input frequency is shown in (a.).

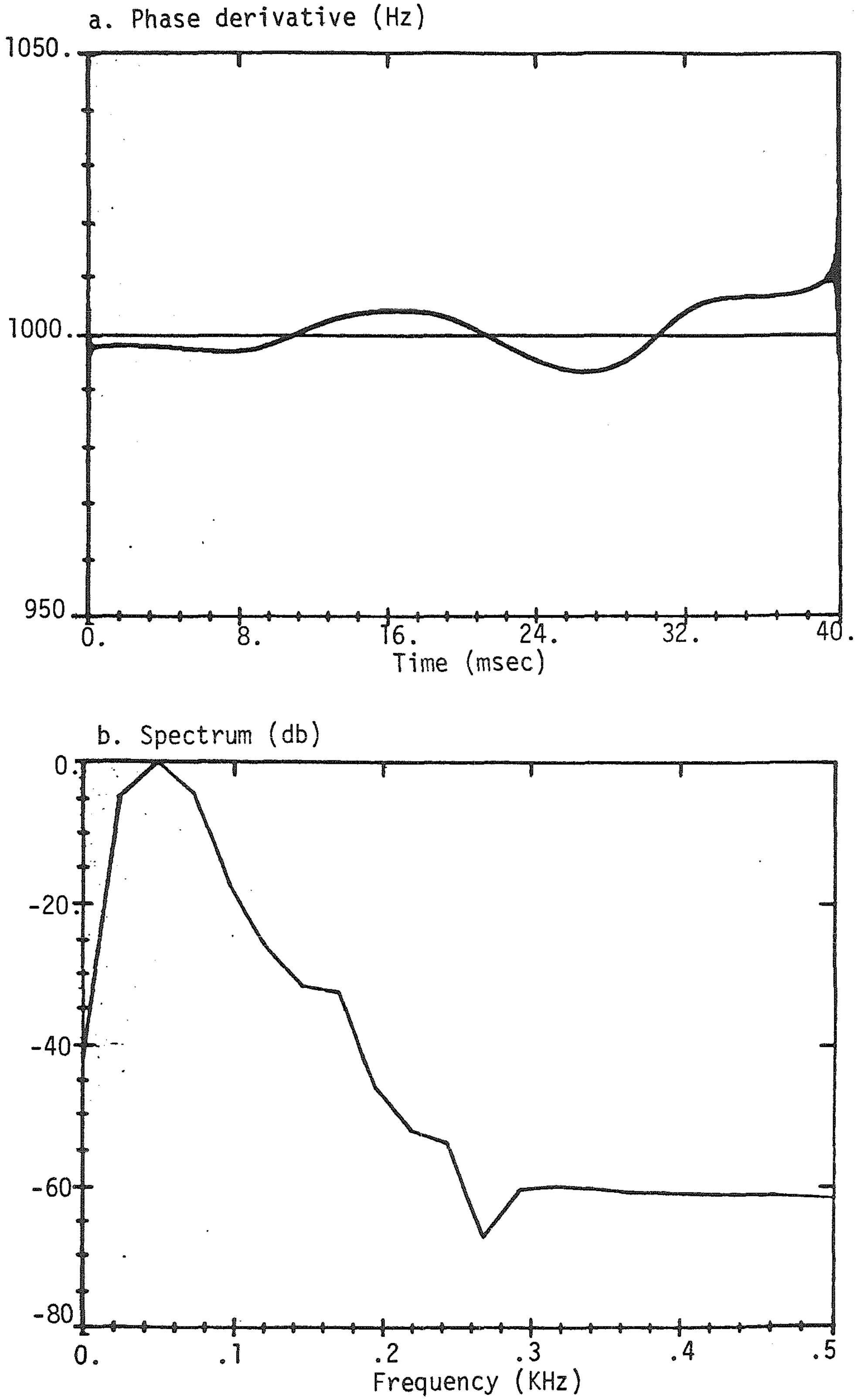
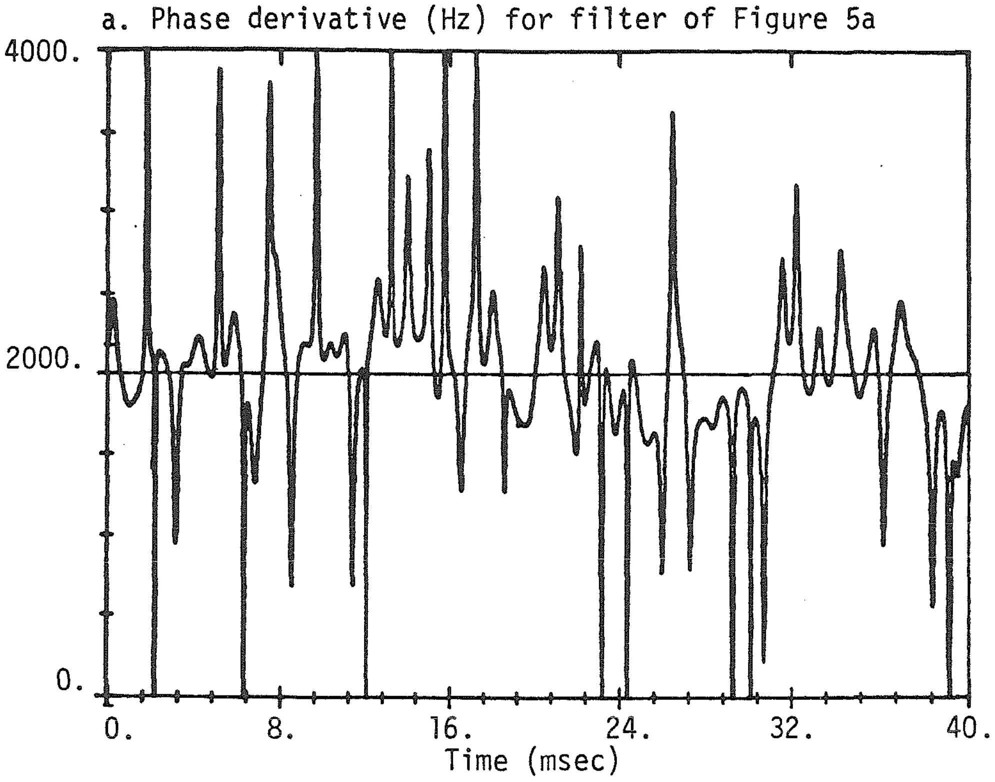


Figure 24. Phase derivative for k=5, N=250, and filter of Figure 16 with input signal of Figure 18. The spectrum of the phase derivative is shown in (b.).



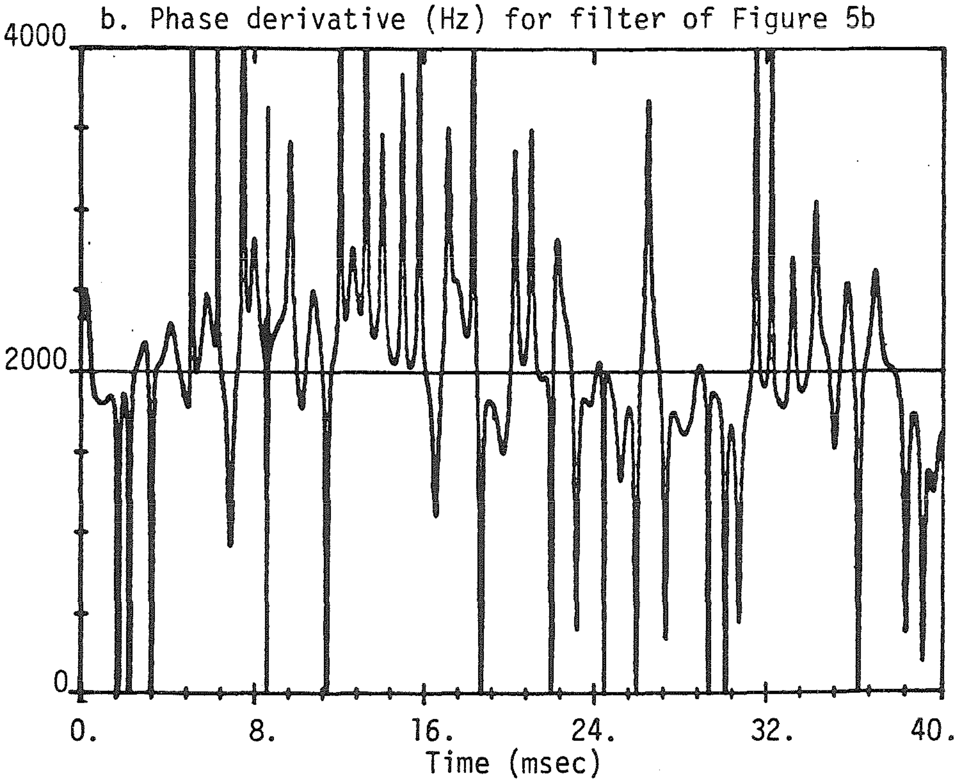


Figure 25. Phase derivative for k=1, N=25, and filters of Figure 5 with input signal of Figure 20.

These spikes reflect changes in phase so rapid that not even the original 50 KHz sampling rate is always sufficient.

A similar problem arises when an actual input signal is preceded by a block of zeros. The first nonzero phase value occurs when the rightmost point of the filter impulse response overlaps the first nonzero signal value. This produces a step change in phase which is not bandlimited by the filter. In fact, this step is as rapid as the sampling rate at which it is computed. It follows that the phase-derivative signal will contain a spike whose magnitude depends on the sampling rate. (An alternative is to prohibit the initial block of zeros and to simply compute an initial phase which is not differentiated. But then it is unclear how to treat this initial phase when modifying pitch.)

We have seen that while the magnitude signal is not bandlimited in theory, it is nevertheless extremely bandlimited in practice. However, the above examples show that the phase and phase-derivative signals are not even particularly bandlimited in practice. Rather than abandoning all hope of computing these signals at lower sampling rates, we now examine the errors which arise in such instances.

We first note that, in the absence of spikes, the bandwidth of the phase-derivative signal is substantially smaller. This is illustrated in Figure 26 for the filter of Figure 16 centered at 1 KHz with a white Gaussian noise input. We made measurements on a number of such examples and found that the bandwidth (again measuring from the -40 db point) of the phase-derivative signal without spikes was typically less than twice the channel separation. It was occasionally as much as four times this separation, but only when amplitude minima were so pronounced as to produce distortions in the phase-derivative which bordered on being spikes themselves. As with the magnitude signal, these phase-derivative

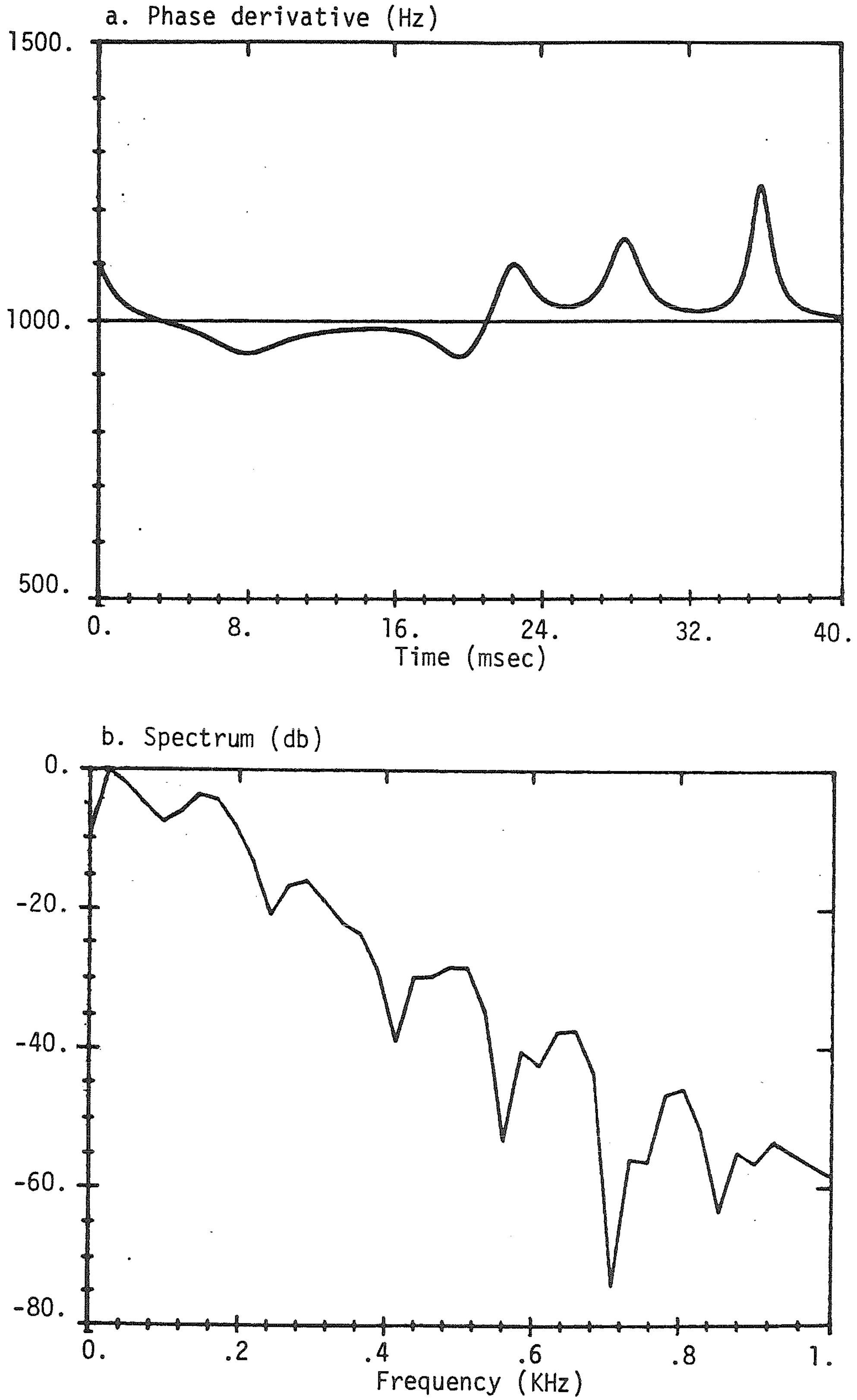


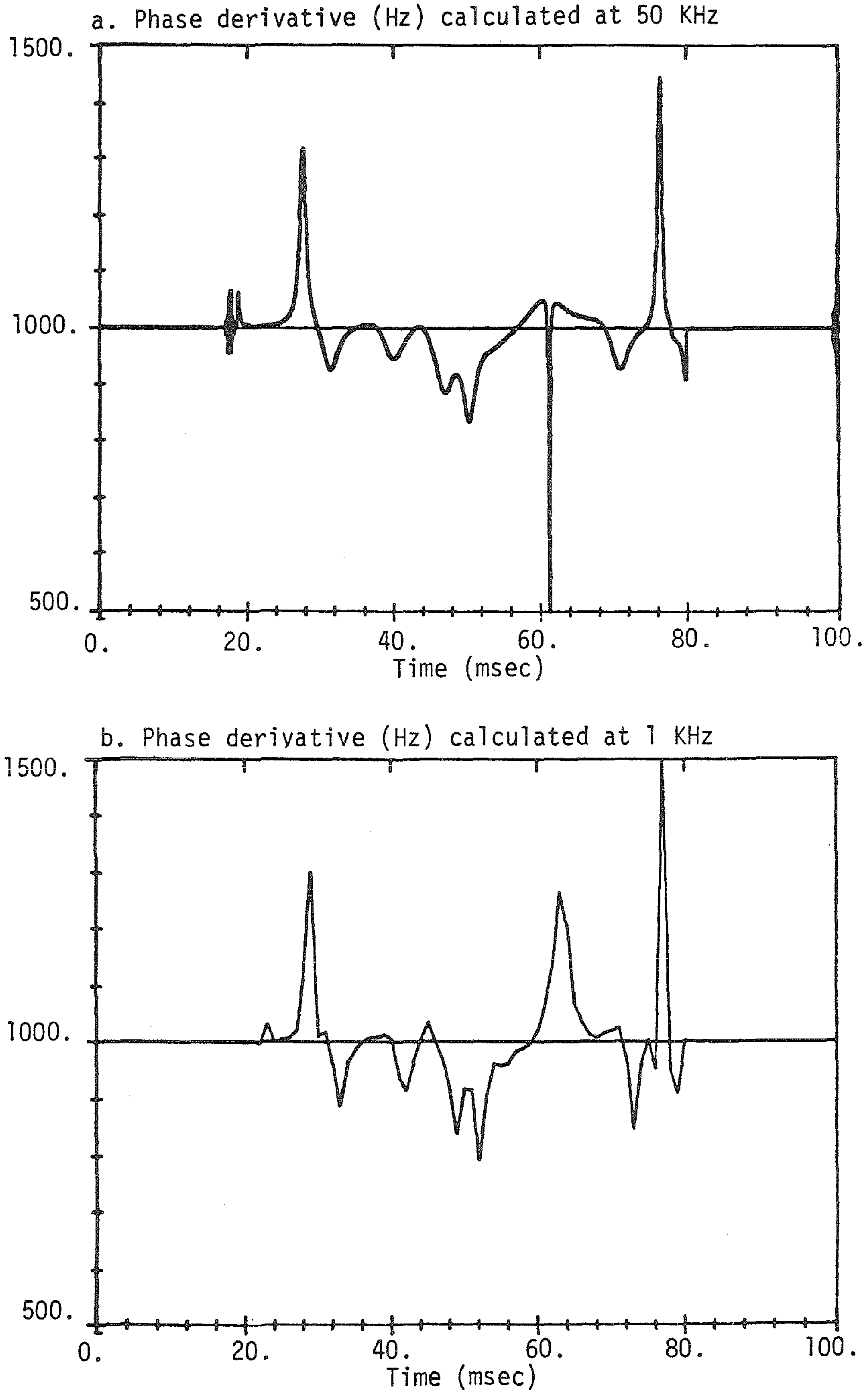
Figure 26. Phase derivative for k=5, N=250, and filter of Figure 16 with input signal of Figure 20. The spectrum of the phase derivative is shown in (b.).

signals could still be fairly accurately computed at twice the channel separation (i.e., the same rates as $a(n,k)$ and $b(n,k)$).

We now consider the phase-derivative signal with spikes. The consequences of computing this signal at too low a sampling rate can be seen in Figure 27 in which the filter of Figure 16 is centered at 4 KHz with a white Gaussian noise input. The 1 KHz signal captures all of the features of the 50 KHz signal except for the central spike. This error is purely local and inaudible because it occurs at a minimum of the amplitude signal. However, it can have global and audible consequences if its integration produces a constant error in the reconstructed phase.

The angle difference avoids this problem because the phase can always be trivially and exactly reconstructed. The angle difference signal still contains spikes which vary with the sampling rate at which the differences are computed, but the instantaneous absolute phase is always preserved (except for possible differences of 2π). Consequently, any errors in representing the spikes are purely local, and there is no need to use the original 50 KHz sampling rate.

The tracking phase vocoder (Section 3.7), offers even greater possibilities for bandwidth reduction. For example, with the tracking phase vocoder, the absolute phase in each channel is no longer critical to a perceptually satisfactory reproduction of the signal. Hence, the spikes can be suppressed entirely without significantly altering the reconstruction. Furthermore, the tracking phase vocoder adjusts the channel center frequency to match the instantaneous frequency of the input; as a result, many of the phase distortions which increase the bandwidth of the stationary vocoder signals do not even arise. Lastly, the tracking phase vocoder can be used with filter bandwidths significantly less than the spacing between harmonics. This reduces the



a. Phase derivative (Hz) calculated at 50 KHz

b. Phase derivative (Hz) calculated at 1 KHz

Figure 27. Phase-derivative signals calculated at two different sample rates for k=20, N=250, and filter of Figure 16 with input of Figure 20.

required sampling rates even further.

3.7 The tracking phase vocoder

Conceptually, a pitch-tracking phase vocoder is a simple extension of the standard phase vocoder described in the preceding sections. Indeed, a kind of pitch-tracking phase vocoder was described by Malah [1979] as a time domain algorithm for harmonic bandwidth reduction. However, the advantages of such a vocoder have not previously been fully appreciated. In this section, we enumerate those advantages and show how a tracking version of the phase vocoder can easily be implemented.

From the standpoint of timbre investigation, the most important feature of a tracking phase vocoder is that it minimizes distortions in the amplitude and frequency estimates. As shown in Section 3.4, these distortions arise when the channel center frequency and the instantaneous frequency are not well matched. These distortions are of no consequence in the resynthesis because they cancel, although they do increase the bandwidths of the magnitude and phase signals. More importantly, however, these distortions make it extremely difficult to determine the true behavior of the amplitude and frequency signals. Consequently, it is difficult to develop simple models from which to proceed. This is particularly true for the kinds of complex musical signals which are analyzed in Chapter 4 (eg., violin with vibrato and violin ensemble).

Another important advantage of a pitch tracking phase vocoder is that the partial under analysis remains centered within the filter bandpass; consequently, contributions from adjacent channels are negligible, and the potential crosstalk can be ignored. As a result, it is no longer essential to preserve the absolute phase in each channel. For example, frequency spikes due to interference can be automatically suppressed with no audible effect.

In addition, a tracking phase vocoder lends itself to a variety of bandwidth compression schemes. For example, unusually narrowband filters can be used to retain only a narrow frequency region surrounding each harmonic. This is also an effective noise reduction technique. The limitation to this approach is (as noted in Section 3.4) that the magnitude signal is itself subject to the bandlimiting of the filter; for a narrow enough filter, the magnitude will be significantly smeared.

On the other hand, a tracking version of the phase vocoder also presents several disadvantages. It is more complex to implement, yet its accuracy (in terms of resynthesis) can never exceed that of the standard phase vocoder. Furthermore, a tracking phase vocoder inevitably embodies assumptions which are unnecessary for the standard phase vocoder; consequently, it is significantly less robust. For example, attempting to track white Gaussian noise can lead the vocoder on a random walk through the frequency domain. Hence, we usually specify a magnitude threshold below which tracking is suppressed. This also prevents the vocoder from losing lock when it encounters a frequency spike. Even with this precaution, however, the ultimate performance depends on the intelligence of the tracking algorithm.

For isolated tones, we are able to track individual partials without difficulty using a fairly simple algorithm. However, for more complex inputs it is necessary to provide a more global tracking strategy as well; the tracking of individual partials must be supplemented with some form of pitch tracking. In such situations we have found it preferable to perform the analysis in two passes: We first apply a standard pitch detection algorithm [Moorer, 1973; Tucker and Bates, 1978] and correct it by hand if necessary. We then run the tracking phase vocoder in a mode in which the tracking is predetermined.

We now consider the actual implementation of a tracking phase vocoder. The introduction of tracking means that the Fourier transform interpretation of Section 3.2 is no longer appropriate. Consequently, the efficient Fast Fourier Transform computational techniques of Section 3.2 are no longer applicable. Instead, the analysis is performed independently for each partial of interest, and the synthesis is performed from the additive model.

The algorithm itself is fairly straightforward. To begin, we choose an appropriate value of ω and simply evaluate equation (3.1) for every R^{th} value of n (where R represents an appropriate decimation factor so that the sampling rate of $X(n,\omega)$ is twice the filter bandwidth). The restrictions on the filter impulse response are no longer significant, so $h(n)$ can be any desired lowpass filter. For each calculated value of $X(n,\omega)$, we immediately compute $|X(n,\omega)|$ according to equation (3.19) and $\Delta\varphi(n,\omega)$ according to equation (3.22). An estimate of the instantaneous frequency is now given by $\omega+\frac{2\pi}{RT}\Delta\varphi(n,\omega)$. Whenever the short-time average value of $\Delta\varphi(n,\omega)$ becomes too large, it is time to adjust ω.

The details of this adjustment determine the kind of feedback control loop which is established; for typical input signals, however, we find that almost any reasonable tracking scheme is sufficient. The critical issue is that the phase of $X(n+R,\omega_{new})$ bears no relation to the phase of $X(n,\omega_{old})$. Consequently, it is essential that the calculation at time n be repeated for the new ω. This establishes a new reference phase from which the differences can now be calculated.

An example of the tracking phase vocoder output is given in Figure 28 for the input signal of Figure 13a. It is clear that both magnitude and angle difference signals are considerably less distorted than for the stationary phase

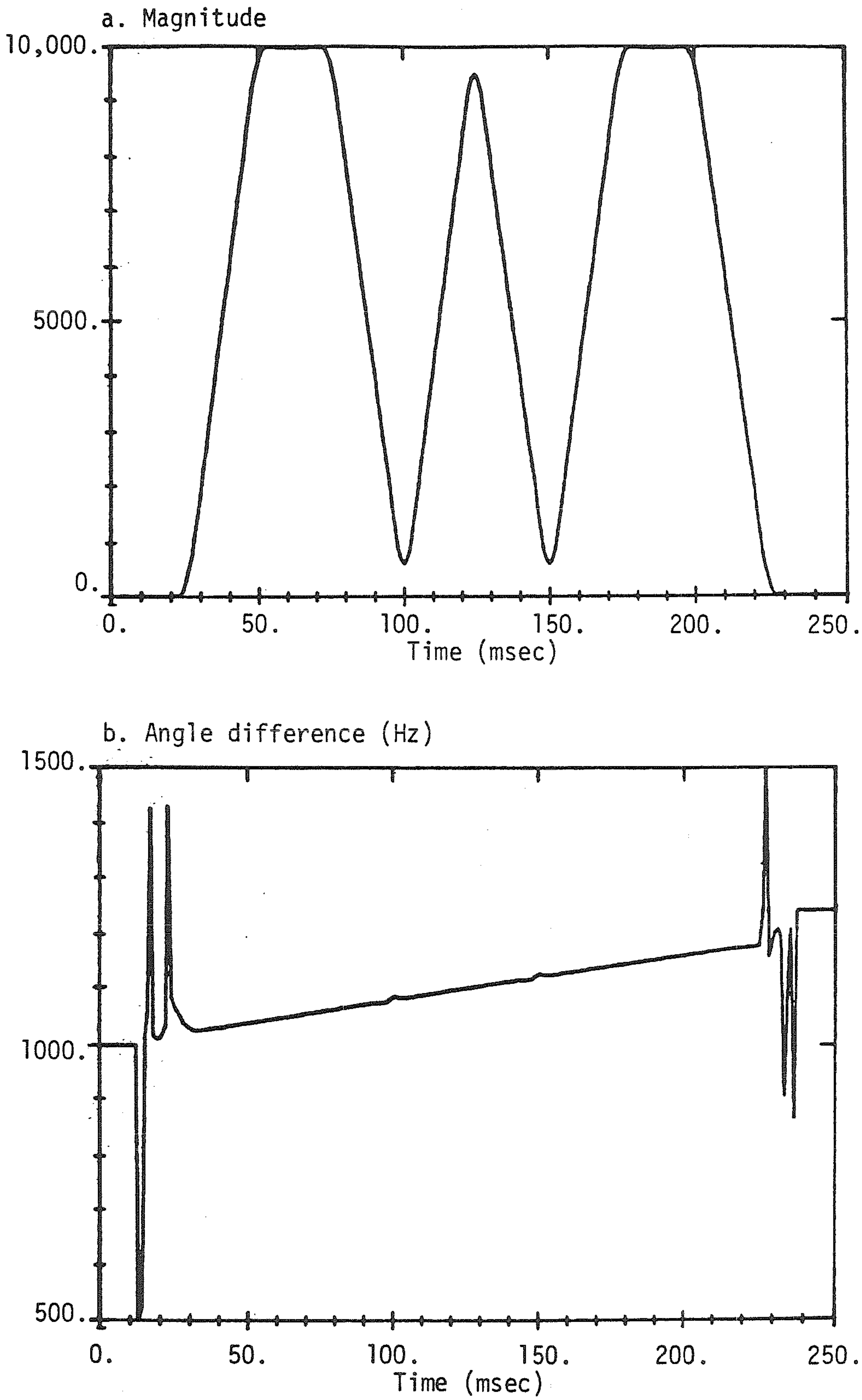


Figure 28. Magnitude and angle difference for the input signal of Figure 13a using the tracking phase vocoder with the filter of Figure 16.

vocoder signals of Figures 11 and 12. However, the resynthesized signal is actually inferior to that of the stationary phase vocoder (which is perfect) due to the bandlimiting of the magnitude signal in the tracking version. In general, the choice of an appropriate filter for the tracking phase vocoder involves a tradeoff between smearing the magnitude signal and retaining undesirable noise. In many applications, it is possible that a simple rectangular impulse response will suffice; in this event, the computations are greatly simplified, and the tracking phase vocoder is simply a tracking heterodyne filter.

3.8 Conclusions

As a consequence of the preceding analysis, we can draw four significant conclusions:

1) The angle difference signal is not, as is frequently assumed, merely a simple approximation to the phase-derivative signal. Rather, it is superior to the phase-derivative signal, because it permits accurate reconstruction of the phase.

2) The magnitude and angle difference signals can, in fact, be accurately calculated at the minimum output sampling rate in most cases of interest. At worst, this produces errors of 2π during frequency spikes.

3) Pitch modification can be performed with negligible modification of the amplitude signal; but this does not guarantee negligible modification of the timbre.

4) A tracking phase vocoder provides a useful alternative to the standard version, particularly in the investigation of timbre. It eliminates distortions in the magnitude and phase signals, and entirely avoids the problem of interchannel interference.

These conclusions amount to a significant improvement in the phase vocoder analysis technique. This improved technique is the basis for the results in Chapter 4.

IV. ADDITIVE ANALYSIS-SYNTHESIS OF ENSEMBLE SOUNDS

A violin section playing in unison sounds very different from a solo violin. This is true even when the comparison is carried out via a low fidelity monaural sound reproduction system. It is strange, then, that while numerous attempts have been made to synthesize ensemble sound, virtually none have been made to analyze it. In this chapter we undertake such an analysis with two basic goals: 1) to identify the minimal cues required for the ensemble sensation, and 2) to identify the cues which enhance that sensation.

4.1 History

The ensemble or *choir* effect has a long history amongst organ manufacturers and a somewhat more recent one amongst computer musicians. While both of these groups obviously have some idea of what characterizes ensemble sound, they have concentrated their attention on trial-and-error schemes for synthesizing it. Furthermore, they have tended to report their schemes only by word of mouth or via patent disclosures. This sort of approach has led to some reasonable ensemble simulations, but has not contributed much understanding of the underlying perceptual cues.

A rare discussion of these schemes in a scientific context appeared in an article by Le Caine in 1956. Le Caine identified ensemble sound with beating between partials of slightly different frequencies. Consequently, the simulation techniques which he described consisted of adding together multiple voices in such a way that the individual frequencies were always different. However, none of these techniques was entirely satisfactory.

A more recent and more efficient approach to ensemble simulation consists of adding together variably time-delayed versions of the solo. This also results in beating between partials of different frequencies and a corresponding ensemble sensation. Again, however, the final product is less than perfect.

In our view, the difficulties encountered in simulating ensemble sound reflect not only the richness of the ensemble perception, but also the limitations of past approaches. Indeed, a number of important questions implicit in these approaches have yet to be addressed directly. For example, is the summation of multiple voices capable even in principle of producing a completely convincing ensemble sound? If so, what sort of complexity is required in the individual voices? And how does this relate to the variable time-delay technique? These are questions which require a more analytical approach.

4.2 Experimental method

In our research, we viewed the ensemble sensation as a timbral attribute which could be investigated in the same way that Grey [1975] investigated the timbres of individual instruments. We therefore adhered closely to Grey's methodology throughout our study. In particular, we used the tracking phase vocoder and the additive model to analyze both solo and ensemble sounds; to represent them in a perceptually meaningful framework; to equalize pitch, loudness, and duration while modifying other perceptual features; and to resynthesize different versions of the sounds for perceptual evaluation.

Because our investigation was the first to examine this aspect of timbre, we found it necessary to impose a number of limitations on our analysis. First, we worked entirely with monaural recording and reproduction; as a result, we virtually eliminated spatial cues. This is standard practice in the analysis of solo timbres, but perhaps more questionable in the analysis of ensembles. However, the analysis of spatial cues is an entirely separate, and largely unexplored, area of research.

In addition, we restricted our investigation primarily to the identification of perceptual cues. A logical extension of this research would be to apply this knowledge to the development of superior techniques for ensemble simulation. However, the detailed development of these techniques was beyond the scope of this investigation.

A third (and unanticipated) limitation was in the quality of the sound examples available to us. This arose from our dependence on volunteers for both recording and listening sessions. We had no trouble obtaining excellent solo sounds, but we experienced considerable difficulty in finding comparable ensemble sounds. This restricted our analysis, but not severely.

Another practical limitation was in the relative informality of our listening procedure. This informality arose primarily because the diversity of the required perceptual judgements made formulation of a standardized test quite difficult. Instead, we presented listeners with unidentified sounds and simply solicited comments. For the relatively crude discriminations which were frequently required, this method provided more than adequate consistency, both among different listeners and over repeated trials with a single listener. However, any further investigation of the trends identified in this research would definitely require a rigorous and formalized evaluation procedure. In addition, it would utilize professional ensembles and examples of considerably greater duration.

A final restriction was in the variety of instruments which we could investigate. Since the violin is the single instrument most widely associated with a distinctive ensemble timbre, we concentrated almost exclusively on the violin. However, it seems reasonable to assume that most of our results hold for other instruments as well.

In the sections to follow, we present results obtained with ensemble violin sounds from three distinct sources:

1) A recording at $7\frac{1}{2}$ inches per second with quarter track tape of four violins in a moderately reverberant room. This was our best ensemble source due both to the quality of the playing and to the variety of the musical material.

2) A recording at 15 inches per second with half track tape of ten violins in a moderately reverberant hall, but with significant ambient noise. The recording took place at a rehearsal of a local college orchestra; unfortunately, the quality of the playing was frequently unacceptable.

3) Commercial phonograph recordings of violin concerti. The problem here was in finding musical selections in which the solo violin and the violin section (in unison) played the same material, and each in the absence of any extraneous sounds. We did find several appropriate examples, but used them only for preliminary testing.

In each of the above recording sessions, we obtained samples of solo violin for each example played by the ensemble. In addition, we recorded solo violinists in two entirely separate sessions:

1) Direct-to-disk (!) digitization of various solo violinists playing in a soundproof box adjacent to our computer.

2) A recording at 15 inches per second with half track tape of a solo violin in an anechoic chamber. We also obtained samples of solo trumpet in the same session.

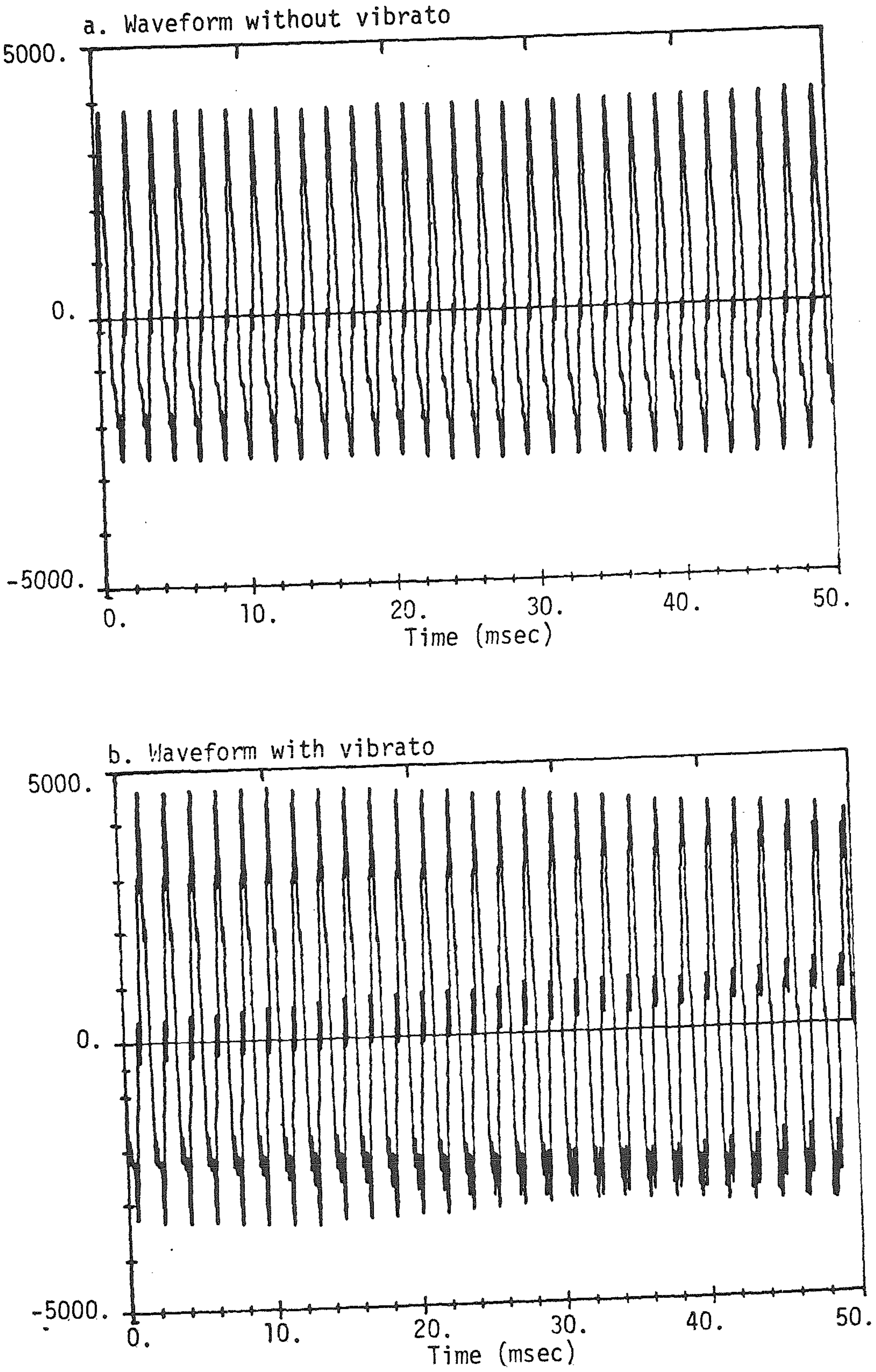
The complete sound recording and reproduction sequence is shown in Figure 3 and was described in Section 2.2.

4.3 Violin solo

In order to compare a violin ensemble to a solo violin, we first must establish the characteristics of the solo. Several investigators have addressed this issue, but none quite sufficiently for our purposes. Fletcher [1965; 1967] performed detailed analyses of the tones produced by a solo violinist in an anechoic chamber, and made important discoveries about the effects of vibrato. However, his investigation was limited by the fact that his analysis-synthesis techniques were entirely analog. Grey [1975] used the heterodyne filter to obtain a more exact representation of a violin tone, but he considered only a single tone of very limited duration (300 msec) and with no vibrato. More recently, Miller [1981] investigated violin vibrato, but using measurements directly from the violin rather than from the recorded tone.

The displacement-versus-time of the violin string at the point of contact with the bow is a roughly sawtooth function due to the alternate sticking and slipping of the bow hairs on the string. However, these oscillations undergo considerable filtering by the violin body so that the recorded waveform is rarely a simple sawtooth. This waveform varies with the particular instrument being played, with the bowing technique, and with the precise pitch of the tone. If these conditions are held constant, then the waveform is quite reproducible; but this seldom occurs in actual musical examples.

The vast majority of violin tones are produced with a periodic pitch variation known as *vibrato* . A typical vibrato introduces a roughly sinusoidal pitch variation of ±1 % at a frequency of 5 to 7 Hz. However, it also introduces a periodic amplitude modulation which is different for each partial. Consequently, the spectrum of the violin waveform is constantly varying throughout any single period of the vibrato. This variation can be seen both in the waveform (Figure


a. Waveform without vibrato

b. Waveform with vibrato

Figure 29. Waveforms of a solo violin played with and without vibrato. The tone is C#5 (555 Hz).

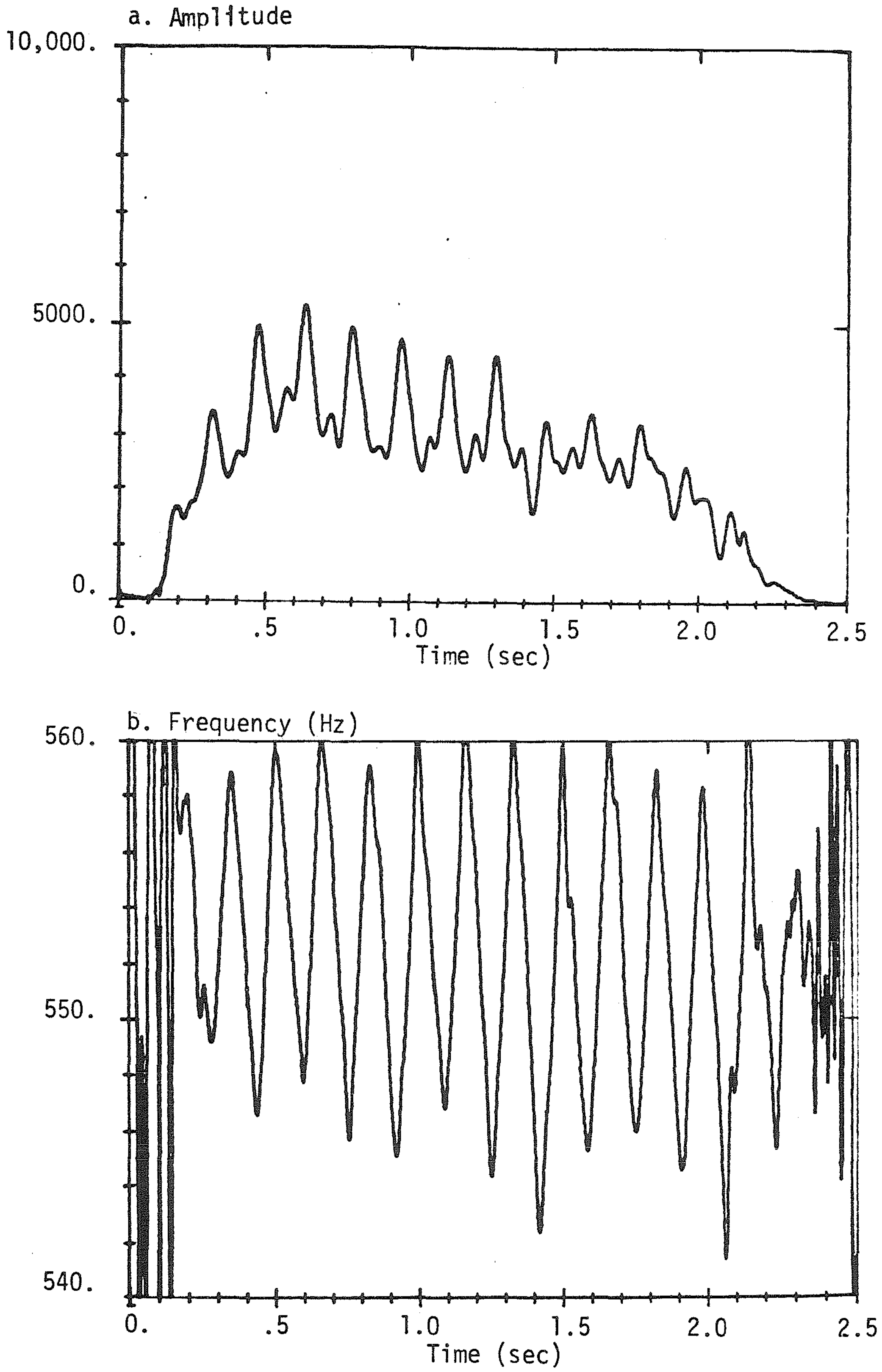


Figure 30. Amplitude and frequency of the fundamental of the tone in Figure 29b.

29) and in the individual partials (Figure 30). This phenomenon was first noted by Fletcher [1965; 1967], and was shown to be indispensable for convincing simulations of violin sounds.

Fletcher conjectured that the amplitude modulation was a consequence of individual partials being swept in and out of nearby body resonances. (Room resonances are ruled out because the effect occurs even in recordings from an anechoic chamber.) We were able to lend further support to this hypothesis by showing that the particular bow and finger motions associated with vibrato are neither necessary nor sufficient for producing this effect. For example, Figure 31 shows the amplitude variation of the fundamental which results from simply sliding the index finger of the left hand down the string to produce a *glissando* . This motion affects the bowing and damping of the string very differently from the vibrato; nevertheless, the differences in amplitude at 550 Hz and 560 Hz are comparable to the differences observed with a ±5 Hz vibrato at 555 Hz.

In contast, Figure 32 shows the variation in the waveform as the amplitude is increased at constant frequency. It is evident that the strength of the higher partials does change, but not nearly enough to link the amplitude modulations during vibrato to mere variations in bowing. In addition we note that, in the case of vibrato, the higher partials frequently exhibit a more complex amplitude modulation which is consistent with being swept through several adjacent resonances. We therefore find it very plausible that these amplitude modulations do arise entirely as a consequence of body resonances.

An independent demonstration of the importance of body resonances was provided by Mathews and Kohut [1973]. They showed that an electrical resonance network could be used to greatly enhance the realism of electronically produced violin sounds. This idea has also been applied in some

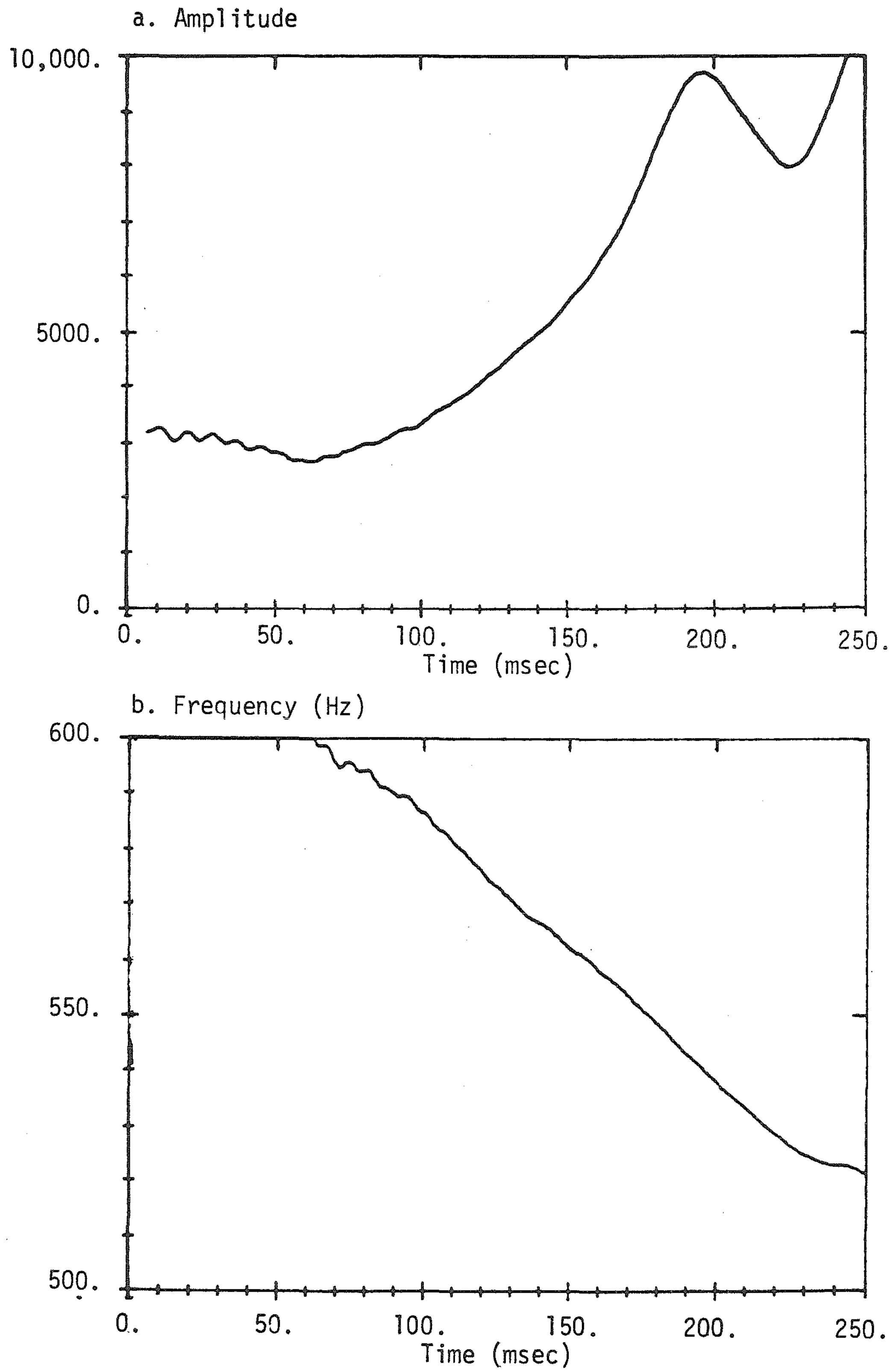
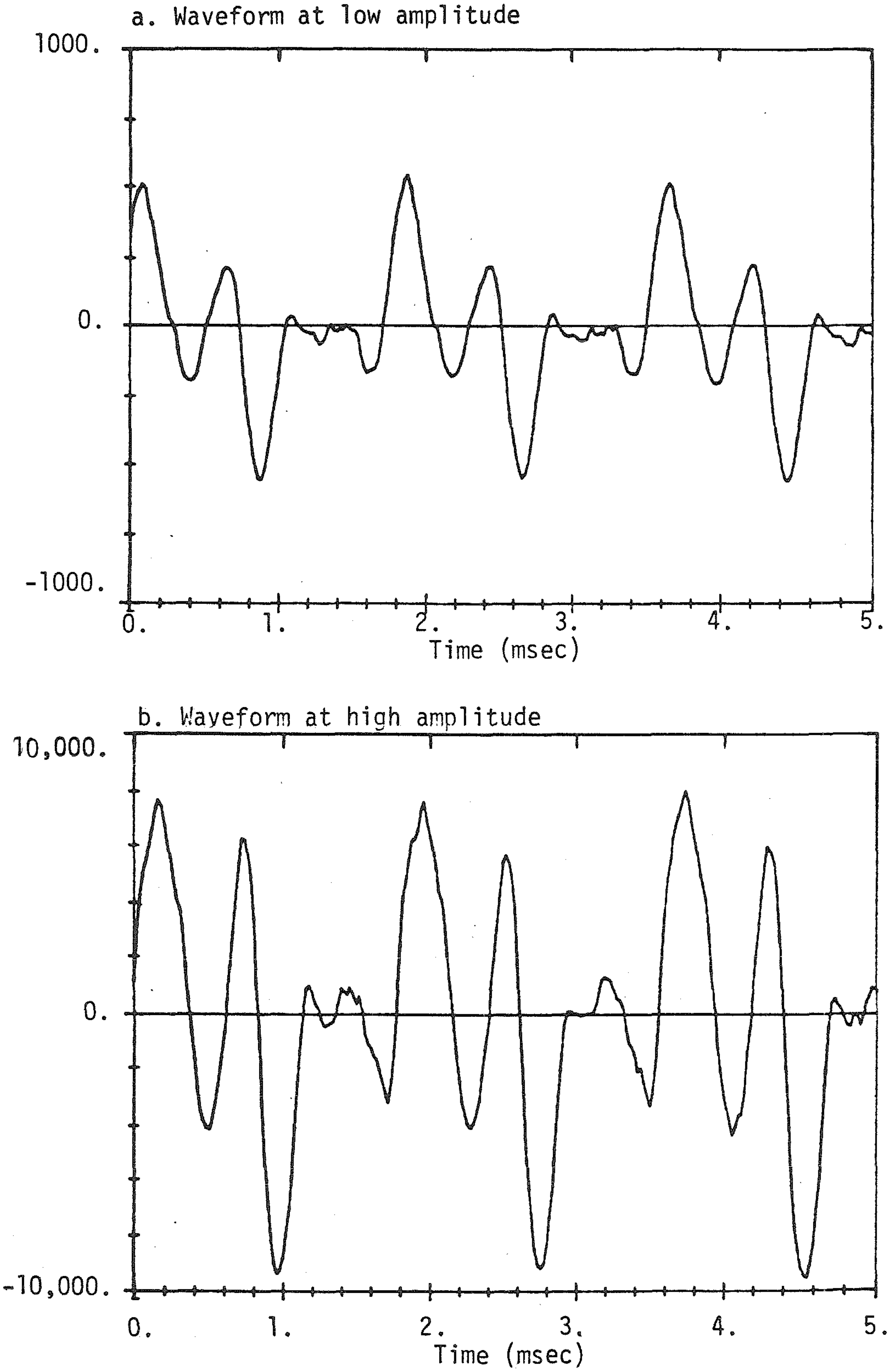


Figure 31. Amplitude variation of the fundamental for a downward glissando. The frequency variation is shown in (b.).



a. Waveform at low amplitude

b. Waveform at high amplitude

Figure 32. Variation of the waveform with amplitude.

recent digital music synthesizers. In our work, we found that a useful alternative could be developed using the additive model.

Following the approach of Grey, we approximated the amplitude and frequency estimates of each partial with five to ten line segments apiece. In this representation, we sought to capture average values rather than details of the modulation. (This is illustrated for the amplitudes of the first eight partials of a typical tone in Figure 33.) We were then able to independently impose pitch variation and amplitude variation for each partial. We found that very plausible simulations could be obtained by using a sinusoidal amplitude modulation of about 30% ; this modulation had the same periodicity as the vibrato, but its phase was chosen randomly for each partial.

In general, we found that the line segment approximations without modulation retained a vague violin-like character, but were not very realistic. We also found that *any* modulation produced a vast improvement in realism, but that both amplitude and frequency modulation were necessary for the ultimate effect. Even then, however, the attacks lacked the "crunch" which is associated with the bow noise during the attack.

We also made use of our improved analysis capabilities to examine the harmonicity of the violin partials. Fletcher reported that the violin partials were actually harmonic, at least within the limits of his analog measurement technique. However, Grey's analysis showed significant deviations from harmonic behavior throughout the entire tone. Further evidence of inharmonicity was obtained by Charbonneau [1981]; in attempting to simplify the additive model representation of cello tones, he found that the assumption of perfectly harmonic partials introduced a very slight but perceptible alteration in the sound.

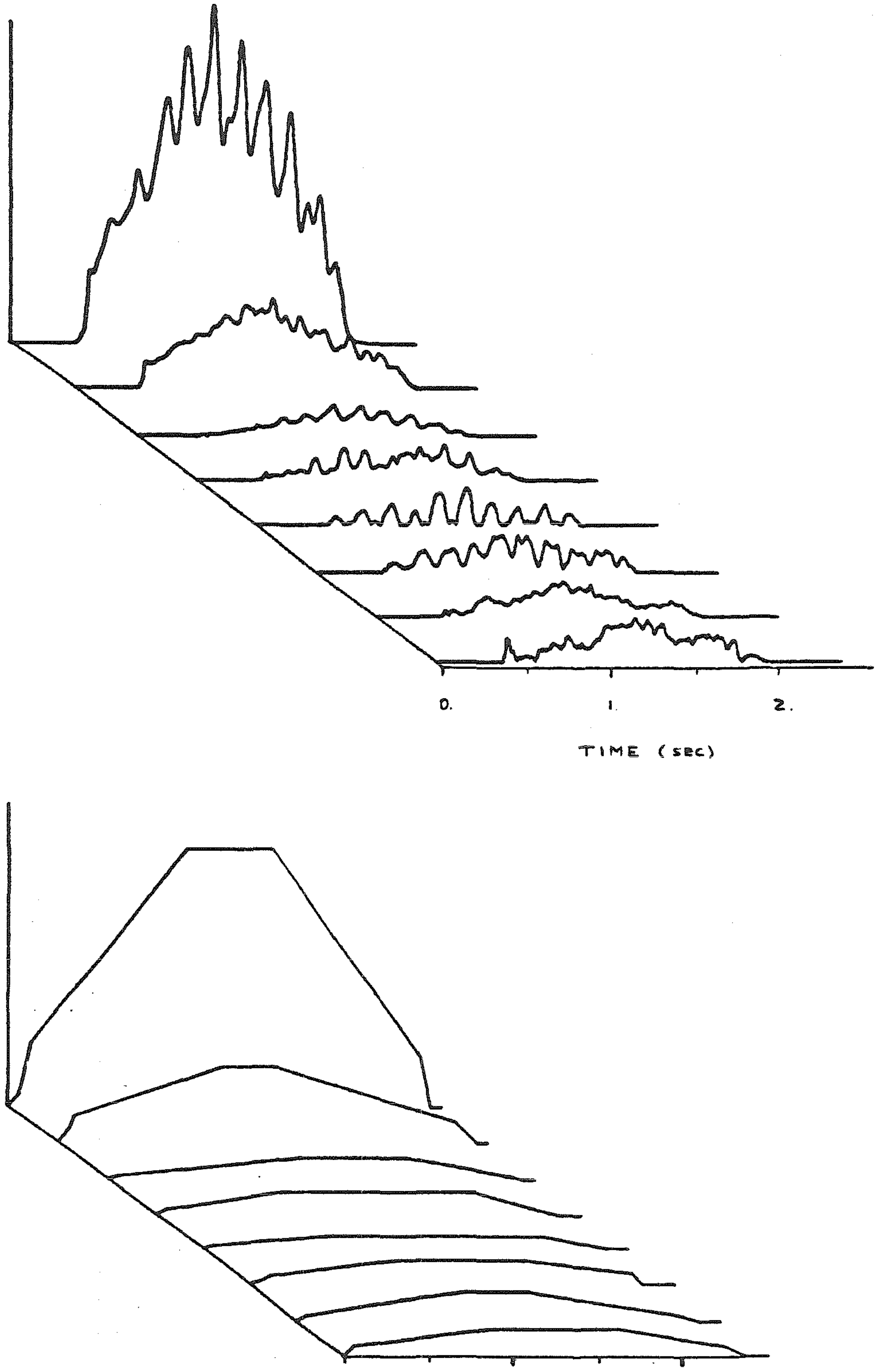


Figure 33. The line segment approximation for the amplitude of the first eight partials of a tone played with vibrato.

The tracking phase vocoder provided a means for resolving this issue because it automatically minimized the distortions in the estimated frequency. We found that the partials of the violin were, in fact, harmonic except possibly during the attack portion of the tone. This portion is typically accompanied by so much bow noise that attempting to define a single frequency is pointless. Since Charbonneau's tones were of extremely short duration (300 msec), it seems probable that his results were due entirely to slight differences in the attack.

Since ensemble sound arises from different instruments playing simultaneously, we also analyzed several different violins individually to determine the extent of the variations among them. We found that these variations were considerable, but still no greater than those of a single violin playing a number of different pitches.

Lastly, we used the tracking phase vocoder to analyze solo violins playing in reverberant rooms rather than in an anechoic chamber. At any point in a room, the reflections from various surfaces add together in such a way that some frequencies interfere constructively while others interfere destructively. This is generally thought of as *coloring* the recorded sound by imposing a characteristic weighting on the spectrum. However, for the violin, it is more appropriate to adopt a dynamic view in which each reflection has its own particular pitch corresponding to the particular phase of the vibrato cycle in which it originated. Consequently, the reverberation looks to the phase vocoder like several different violins playing at once. (For example, there are frequency spikes such as described in Chapter 3 for the case of two sine waves of slightly differing frequency.) How the ear is able to distinguish between the reverberation and the actual ensemble is an interesting problem.

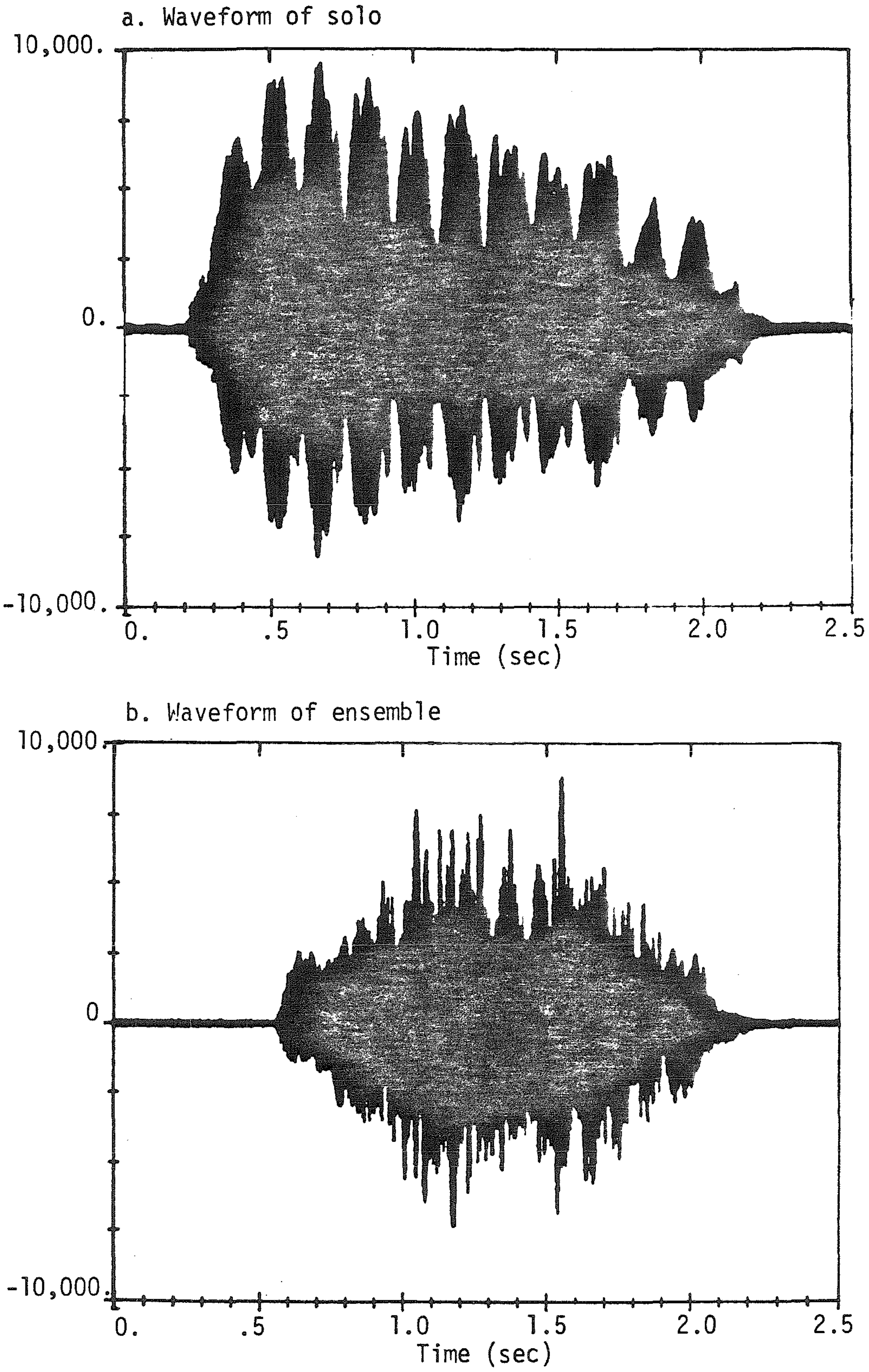
4.4 Violin ensemble

Our best example of violin ensemble came from a chamber group with only four violins. This was a smaller number than we would have preferred, but it proved to be quite sufficient; the ensemble sound differed markedly from the solo both perceptually and analytically. A detailed comparison of the solo sound with the ensemble is provided by Figures 34 thru 39. In this example the tone is F5 (690 Hz) played *mezzo forte* with normal vibrato; however, these figures illustrate features which were common to all our examples.

First, we note the presence of pronounced beating in the magnitude signals of Figures 36 and 38. This beating is not particularly regular; however, it clearly increases in frequency for the higher harmonics. Of course, this is just what we should expect; if the fundamental frequencies differ by Δf, then the 10^{th} harmonics differ by $10\Delta f$, and their beat frequency is proportionally higher. In addition, we note that the magnitude nulls occur at different points in time for different harmonics.

A second obvious difference between solo and ensemble can be seen in Figures 37 and 39; the composite frequency of the four violins shows almost no trace of the individual vibratos. (The more severe frequency spikes in these examples are suppressed for the sake of clarity.) Indeed, for the higher harmonics, it is difficult to make any sense at all of the frequency estimates provided by the phase vocoder. However, we note that since the individual violin waveforms are exactly harmonic, their sum must be also.

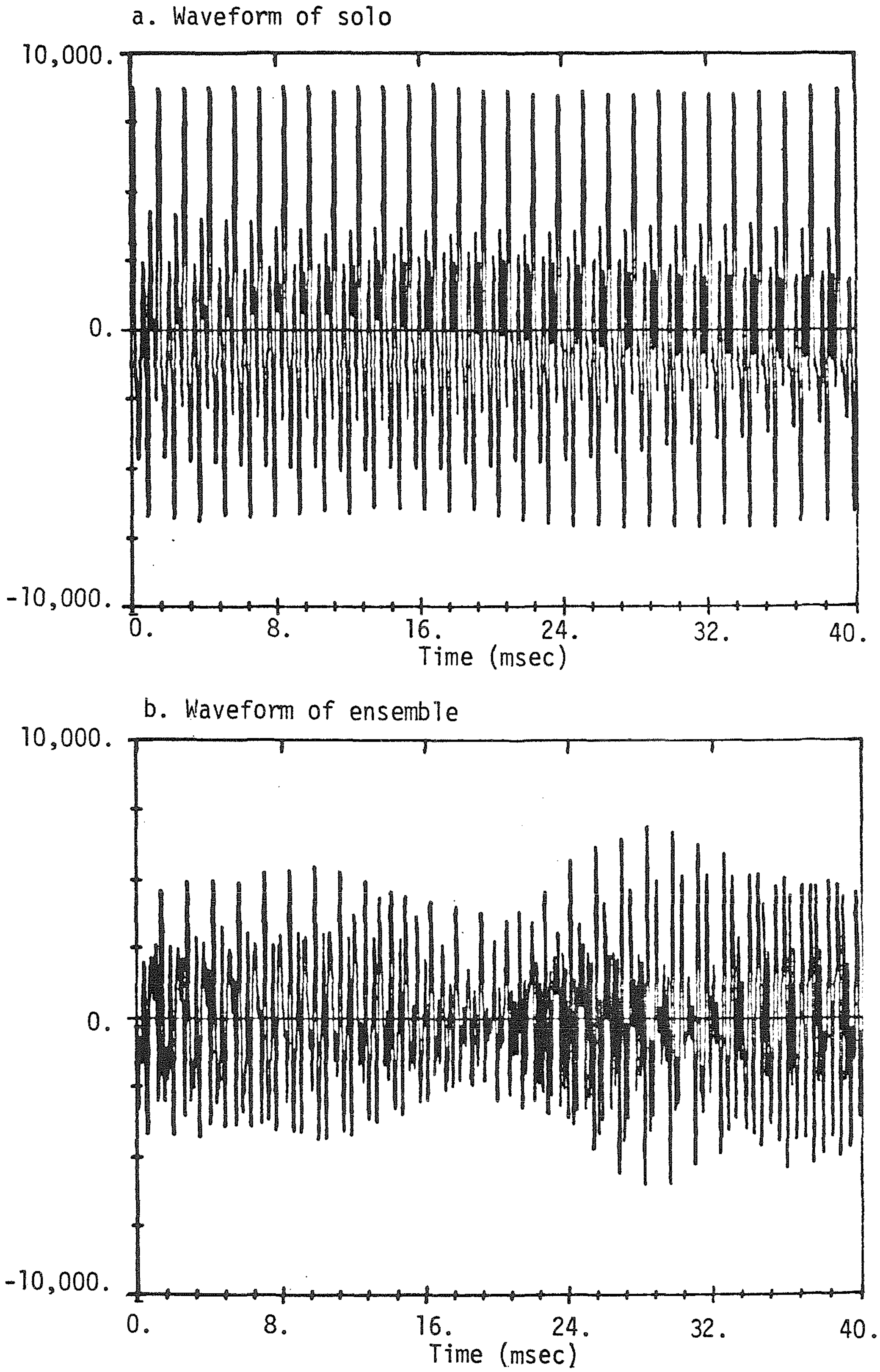
To test the importance of these two features, we performed a simple experiment. We took the additive model representations for several solo tones and for several ensemble tones, and equalized them in pairs for pitch, loudness, and duration. We then synthesized mixed versions in which the ensemble



a. Waveform of solo

b. Waveform of ensemble

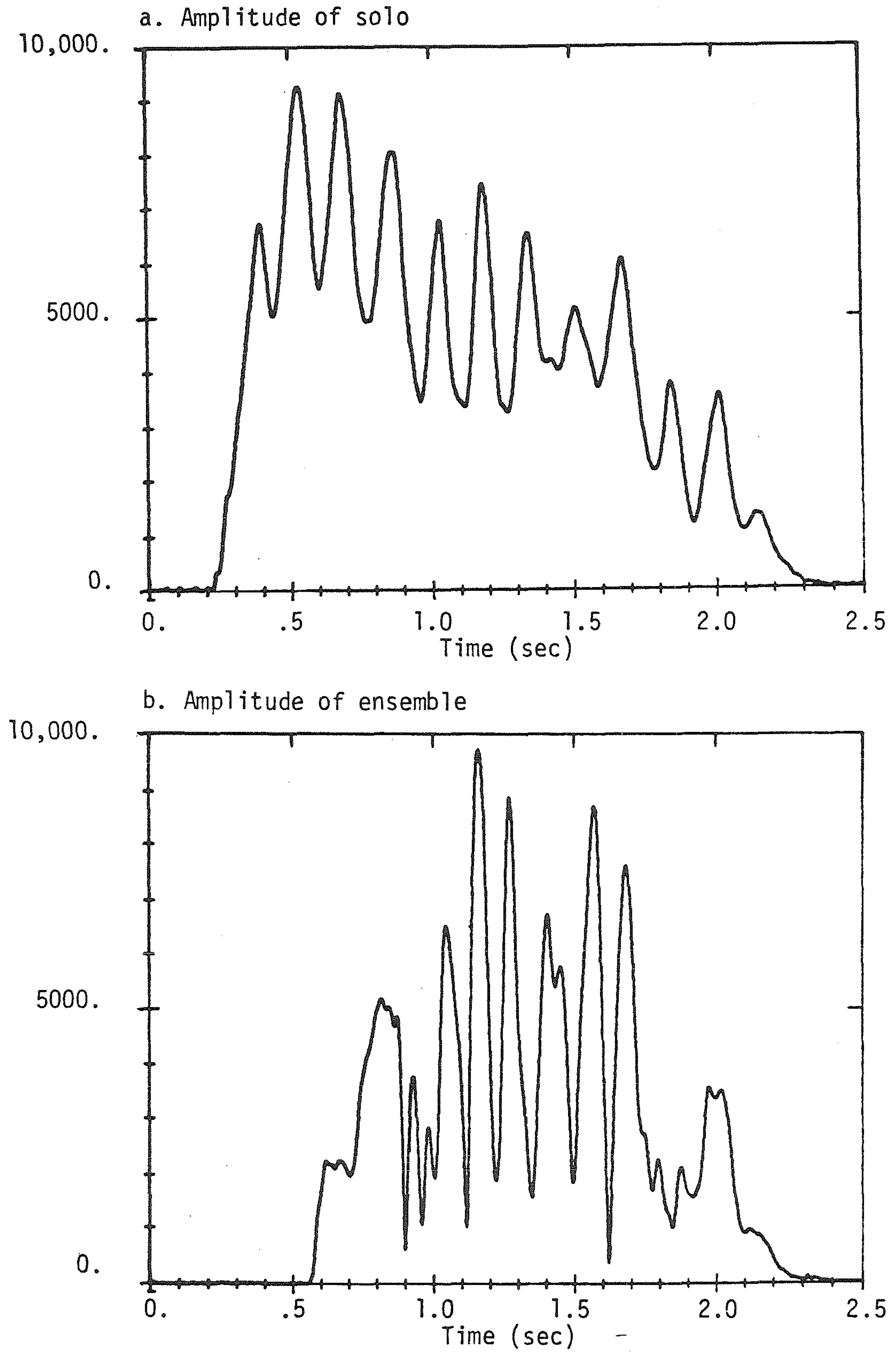
Figure 34. Comparison of solo and ensemble waveforms for F5 (690 Hz).



a. Waveform of solo

b. Waveform of ensemble

Figure 35. Comparison of waveforms in Figure 34 on expanded time scale.



a. Amplitude of solo

b. Amplitude of ensemble

Figure 36. Amplitude of the fundamental for the waveforms of Figure 34.

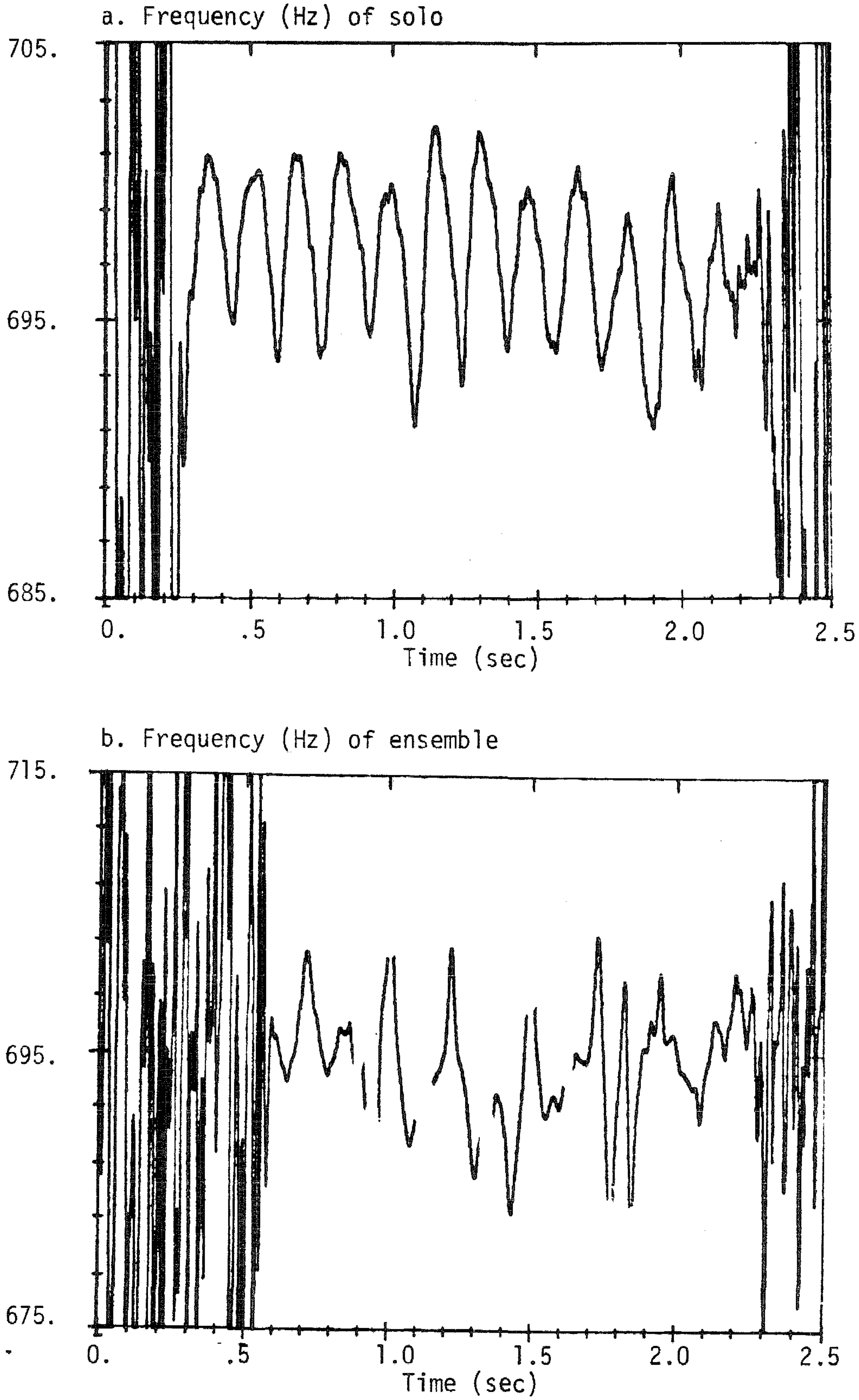
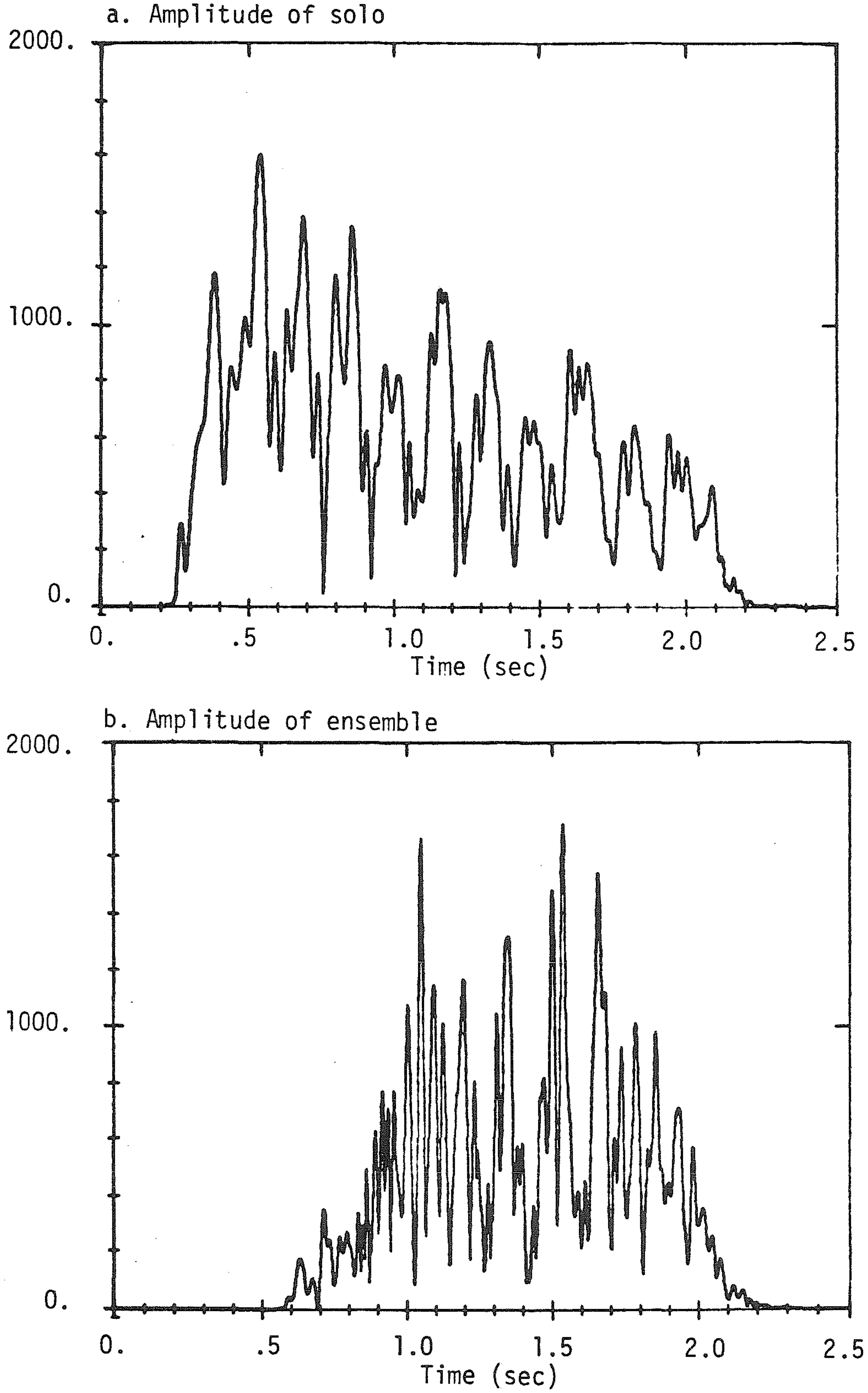


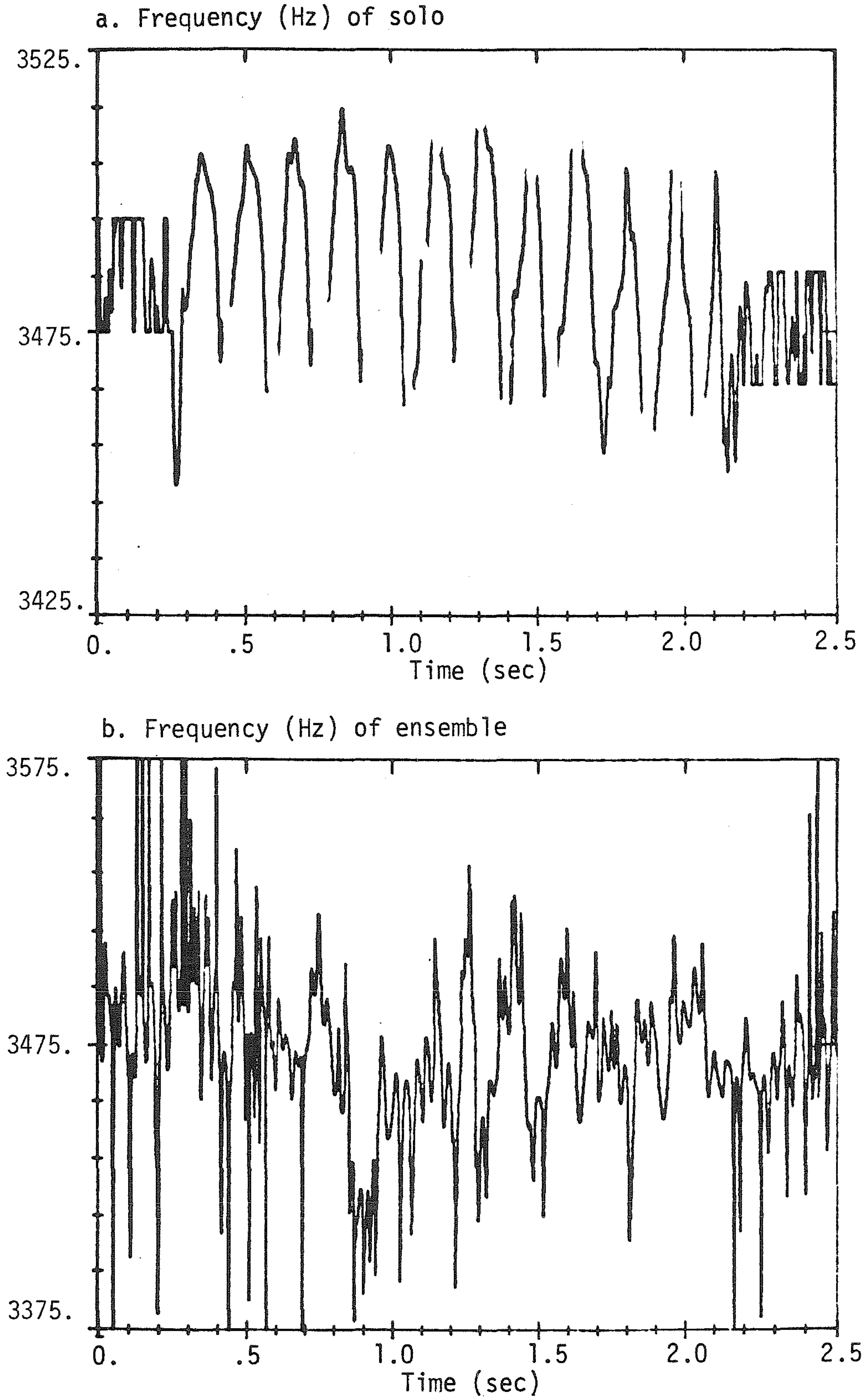
Figure 37. Frequency of the fundamental for the waveforms of Figure 34.



a. Amplitude of solo

b. Amplitude of ensemble

Figure 38. Amplitude of the fifth harmonic for the waveforms of Figure 34.



a. Frequency (Hz) of solo

b. Frequency (Hz) of ensemble

Figure 39. Frequency of the fifth harmonic for the waveforms of Figure 34.

amplitudes were paired with the solo frequencies and vice versa. Surprisingly, we found that the tones synthesized with the ensemble amplitude functions were indistinguishable from the true ensembles, regardless of their frequency variation. The tones synthesized with solo amplitude functions and ensemble frequency functions were distinguishable from the true ensembles, but they still retained a weak ensemble sensation. (However, we note that in this case, frequency spikes did not coincide with amplitude nulls.) Hence, we concluded that the amplitude variation of the ensemble was sufficient but not entirely necessary for the ensemble sensation.

4.5 The use of isolated tones

The comparisons of the preceeding section were based entirely upon the use of isolated tones. We viewed these comparisons as a valuable starting point for our investigation, but we also recognized their limitations. In this section, we discuss extensions of the above research to both shorter and longer time periods.

We first asked whether any particular portion of the tone was itself sufficient to produce an ensemble sensation. A number of early timbre investigations made sharp distinctions between the *attack* portion of the tone and the *steady state* and *decay* portions. Wedin and Goude [1972] in particular found that the attack portion was crucial in the determination of timbre; solo tones with the attacks deleted were surprisingly difficult to identify. We therefore wondered whether the steady state portion of the ensemble would still be identifiable as an ensemble.

To test this, we selected steady state portions of varying duration (.2 sec to 1 sec) from the equalized solo and ensemble tones and applied trapezoidal windows with ramp times ranging from 1 msec to 50 msec. (The attack time for the violin is typically in the range of 50 msec to 200 msec.) We found that the difference between solo and ensemble remained obvious throughout all examples. This is significant because it rules out the attack (and nonsimultaneous attacks in particular) as a necessary cue for the ensemble sensation. It can be reconciled with the findings of Wedin and Goude by assuming that the crucial element for timbral discrimination is not the attack specifically, but rather the presence of variation within the tone segment. In the solo violin waveform, this variation is provided by the vibrato; in the ensemble it is provided by the beating.

We also compared the solo and ensemble attacks; because of their duration, we found that the attacks themselves were sufficient to discriminate between solo and ensemble. In general, the ensemble attacks were more gradual than the solos. Furthermore, we found that imposing a very rapid attack on all partials of an ensemble tone in synchrony produced an initial but rapidly fading sensation of solo. Hence we concluded that, while the attack is not crucial to the ensemble sound, it cannot be entirely ignored.

The above experiments showed that the steady state portion of a single tone is sufficient to identify the ensemble sensation; however, such a time interval is scarcely sufficient to determine the quality of that sensation. For careful evaluations of quality, we would expect to use time intervals of at least several minutes. In this investigation, however, the limitations of memory and computing power made it difficult to work with time intervals of more than several seconds. Nevertheless, we did undertake a preliminary study of longer sound examples.

The extension of timbral research beyond the level of isolated tones is a fairly recent idea. A first step in this direction was taken by Grey in 1978. Grey

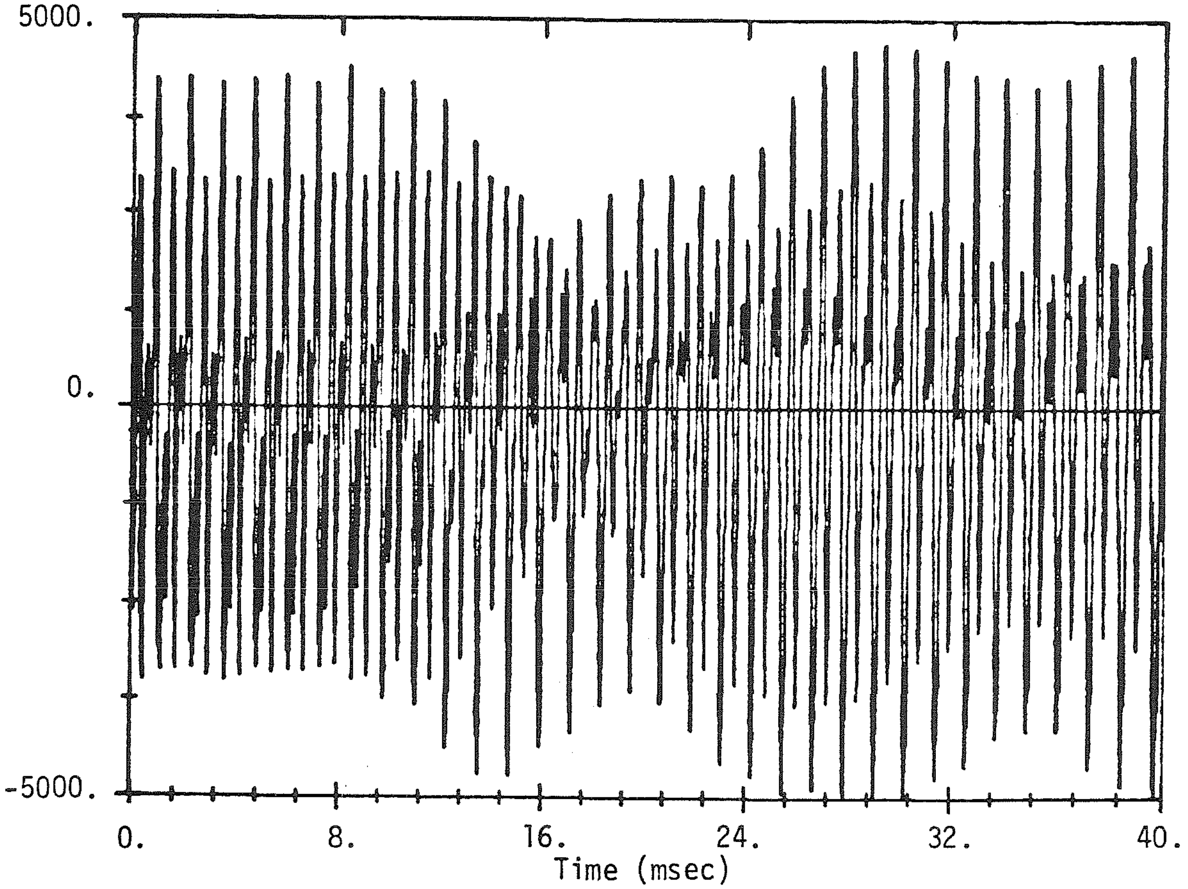
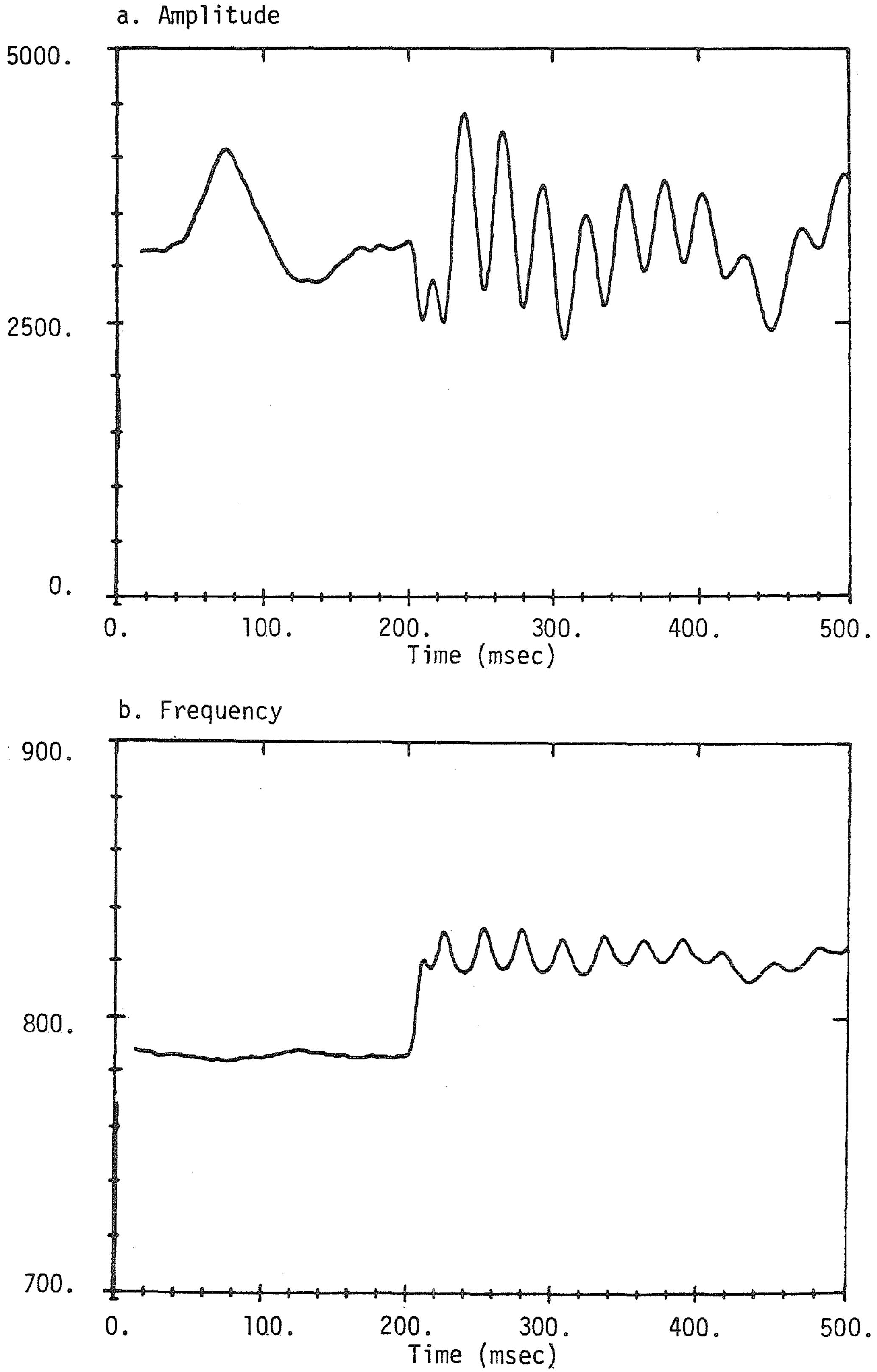


Figure 40. Waveform of violin during transition between two tones of different pitch.



a. Amplitude

b. Frequency

Figure 41. Amplitude and frequency of the fundamental for the transition of Figure 40.

noted that musical context could affect the perception of timbre in at least three important ways:

1) The opportunity to compare the spectra of a variety of pitches from the same instrument could lead to the creation of a composite spectral map.
2) The articulation employed in playing successive tones could provide crucial additional information.
3) The mode of listening might vary dramatically in different musical contexts.

In Grey's study, he simply resynthesized a particular solo tone at different pitches and concatenated the resulting tones to form musical patterns. He found that the nature of the patterns did affect the timbral discrimination. However, in avoiding the analysis-synthesis of connected tones, he excluded the first two of the above points from consideration.

In our work, we asked whether we could extend the analysis to include useful representations of connected tones. An example of a typical *legato* transition between two different pitches is shown in Figure 40. The phase vocoder output for the channel tracking the fundamental is shown in Figure 41. It can be seen that there is a slowly decaying amplitude and frequency modulation throughout the first quarter second of the new note. This is an indication that the room reverberation introduces a considerable overlap of the two tones. It is possible that transitions of this sort can be effectively simulated with simple overlapping, but we performed only preliminary investigations in this regard. A significant problem in such analyses was in obtaining reliable tracking for the higher harmonics during the transition.

4.6 The use of multiple voices

Perhaps the ultimate goal of musicians involved in ensemble simulation, is to discover a single mathematical operation which will transform any given solo waveform into a convincing ensemble. Implicit in their attempts, is the assumption that this can in fact be done. A weaker assumption, which is also widely shared, is that a summation of independent tones from a single instrument can produce a convincing ensemble. However, there are a number of conceivable reasons why this might not be so.

Instrumentalists in an ensemble each occupy a unique location in space. Consequently, their individual sounds are colored differently by reverberation. In addition, they play separate instruments which vary considerably in the details of their individual timbres. It is certainly possible that the combined effects of these differences might be audible. Furthermore, we know very little about the detailed interactions between the musicians in the context of the ensemble. This information could be obtained by placing separate microphones on each violin in the ensemble and correlating the different signals. But are such details really important in the final product?

To answer this question, we conducted a detailed test of the "summation of multiple voices" technique. We collected a variety of examples in which a single violinist repeated a note or phrase from four to ten times. At first, we simply added these together arithmetically. The result sounded very much like an ensemble, but there were obvious cues in the decay portion because not all repetitions had precisely the same length.

In a modified version of this experiment,, we used the additive model to equalize the repetitions for duration prior to addition. We then added the multiple individual voices together and equalized them for loudness and pitch with respect to the ensemble. We found that listeners could still discriminate between the different examples, because each was unavoidably different.

However, listeners were unable to consistently identify the true ensemble as such; that is, within the limits of our experimental method, the simulated ensemble was not identifiably inferior.

Interestingly enough, it was quite easy to discriminate the artificial ensemble analytically. This was because the spectrum of the simulated ensemble retained the irregular deviations of the solo instrument; for the ensemble these solo variations tended to cancel. This is (as described in Section 4.5) a cue which would not be evident for an isolated tone, but which would definitely be a factor when listening to an entire passage. An illustration of this effect is given in Figure 42.

We were also curious about the importance of synchrony in creating the simulated ensemble. The only known study of asynchrony in actual ensemble performances was carried out by Rasch [1979]. He found that asynchronies of 30 to 50 msec were not uncommon for string ensembles, and he offered a number of possible explanations as to why such differences are generally not perceived. We wondered, though, whether these differences could play a role in the ensemble sensation. Consequently, we constructed another set of simulated ensembles using the same technique as above, but varying the synchrony of the attacks. We found that the significant factor was not so much the synchrony itself, but rather the way in which the corresponding vibrato patterns happened to correlate. In cases where the vibratos themselves were in synchrony, the ensemble effect was poor regardless of the attacks.

4.7 Minimal cues for ensemble sensation

We next attempted to determine the minimum set of perceptual cues which would still provide the ensemble sensation. We already knew that neither the attack nor the composite frequency variation was important; hence we

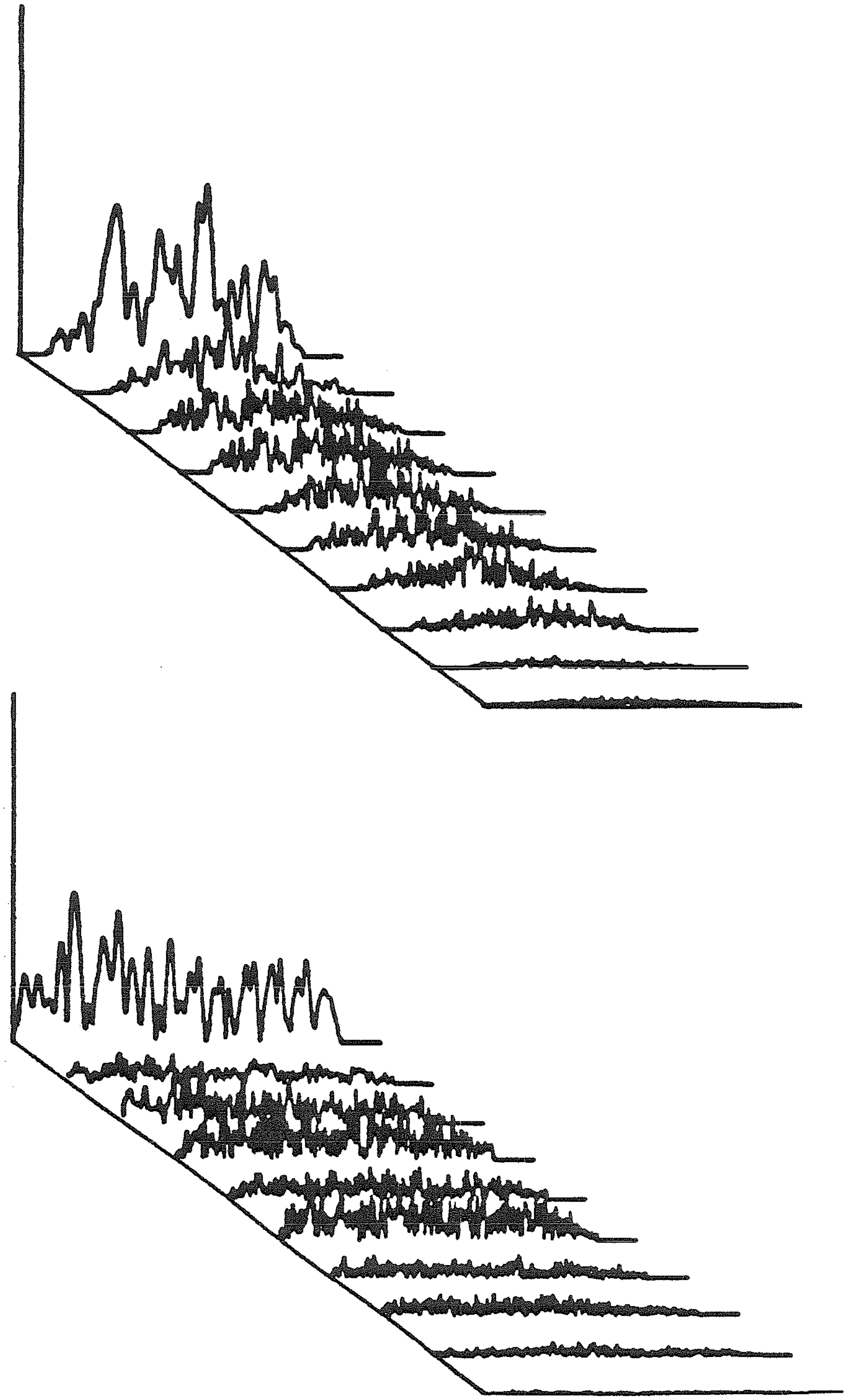


Figure 42. Comparison of amplitude for the first ten partials of an actual ensemble (above) and a summation of multiple solo voices (below).

constructed entirely artificial tones and experimented simply with amplitude modulation. We found that nothing at all could be done to an isolated sine wave which would result in an ensemble sensation. This is not surprising, because we feel intuitively that the ensemble sensation corresponds to a situation in which the ear is somehow overburdened, but in which there is nevertheless a certain structure. In the case of an isolated sine wave, this is apparently simply not possible.

Interestingly, we found that simply adding in higher partials also failed to result in an ensemble sensation. Rather, as the variation in the individual partials approached the level which was associated with ensemble sound in our other examples, the perceptual effect became increasingly less unified. Instead of hearing a single tone of complex timbre, listeners reported a number of simultaneous sine waves of different frequency.

This difficulty in obtaining fusion with certain artificial kinds of sounds has been noted before [McAdams, 1980]. The solution is to impose a common envelope and vibrato on the individual partials. After a great deal of experimentation, we settled on a linear attack and decay of 20 msec duration and a spectrum which rolled off inversely with frequency. Even then, we found that anywhere from four to eight partials were required before the sound could be identified as an ensemble.

We also considered a variety of amplitude modulations. For example, a very simple amplitude modulation which has been employed in some recent simulation techniques is a beating whose frequency is independent of the frequency of the partial. We found that this technique produced only a very marginal ensemble effect. In general, we found that the more rapid beating of the higher partials in the actual ensemble was essential to prevent the ear from detecting the amplitude modulation explicitly. On the other hand, rapid beating of the lower partials simply made the ensemble sound badly out of tune.

We also succeeded in creating a somewhat weaker ensemble sensation using only frequency modulation. We did this by synthesizing a tone in which the phase of the vibrato sinusoid was different for each partial; consequently, none of the partials were instantaneously harmonic, but all were harmonic on the average. We found that the ensemble sensation persisted even when we used tones with missing partials so that there was only one partial per critical bandwidth. This suggested that the comparison of frequencies was being carried out at some higher level of processing.

4.8 The number of instruments in the ensemble

Another aspect of the ensemble sensation which we attempted to investigate was the number of instruments in the ensemble. Unfortunately, our recorded examples provided us with only two different sizes of ensemble: four and ten. Furthermore, the recording environments for these two examples were significantly different. Consequently, we could not reliably attribute differences in the perceived timbre to differences in the number of instruments; indeed, listeners who heard these examples varied widely in their estimates of the number of instruments involved.

To circumvent this problem, we constructed artificial ensembles by adding together varying numbers of independent solos as in Section 4.6. However, these solos - having all been produced by the same instrumentalist - all had nearly identical vibratos. As a result, the crucial factor in the ensemble sound was not the number of solos, but rather the relative phasing of the vibrato in the individual solos. Only solos with significantly different instantaneous frequencies appeared to contribute to the ensemble sensation.

As an added refinement, we used the line segment approximation technique of Section 4.3 to synthesize solos with different vibrato rates and then added these together. However, we still failed to observe any perceptual correlation with the number of solo voices. We therefore turned to the examination of artificial tones such as those of Section 4.7; but rather than applying an artificial amplitude modulation, we constructed independent voices and added them together. All of the voices in these experiments had a 20 msec linear attack and decay and consisted of eight harmonic partials with intensities inversely proportional to frequency. This allowed us to explicitly evaluate the effect of known solo frequency variations on the amplitude of the sum.

We first experimented with solo voices of constant amplitude and frequency. We found that the most acceptable ensemble sounds resulted when solo pitches were all within one or two percent of a given average value. Individual pitch deviations of more than two percent produced a distinct out-of-tune sensation in the ensemble. These percentages were fairly independent of the actual pitch (ie.,the important factor was the ratio of pitch deviation to average pitch rather than the actual pitch deviation in Hz).

To determine the effect of varying the number of solo voices, we allowed the solo pitches to be chosen randomly from within ±1.5 percent of a specified average value; we then added the solos together to produce various size ensembles. These ensembles were automatically equalized for pitch (250 Hz) and duration (1 sec), and then iteratively adjusted to equalize the perceived loudness. This last operation was necessary because differences in loudness translated to major discrepancies in perceived timbre.

We found by listening that there was a significant difference between one, two, and three voices, but very little difference between three and ten voices. We then analyzed selected harmonics to see if these observations could be explained in terms of the corresponding amplitude modulations. As expected, we found that two-voice ensembles had a very regular amplitude modulation; however, the addition of a third voice completely disrupted this regularity. Furthermore, the addition of voices beyond the third was virtually indetectable (Figure 43). Hence, we concluded that any timbral differences between large and small ensembles must be more related to differences in loudness (or to differences in the constituent solo timbres) than to any inherent feature of ensemble sound.

Lastly, we used the above technique to investigate the effect of individual vibratos on the composite amplitude modulation. For this experiment, we used a sinusoidal vibrato and allowed the rate to be chosen randomly between 5 Hz and 7 Hz for each voice. We found that the vibrato did introduce a recognizable distortion in the amplitude modulation (Figure 44), but that this distortion was not actually audible. This was somewhat surprising because the case of no-vibrato versus vibrato was easily distinguishable in our recorded ensemble examples. To improve the realism of our simulations, we imposed individual amplitude modulations on each harmonic as described in Section 4.3. However, the simulations with vibrato were still indistinguishable from those without. The explanation for this is still unknown.

4.9 Simulation of ensemble sounds

The development of an improved technique for ensemble simulation was beyond the scope of this research; nevertheless, we did undertake some preliminary investigations. The most popular existing choir effect generator appears to be the analog variable-time-delay circuit which is used in the Moog and many other electronic synthesizers. The sounds produced by this technique

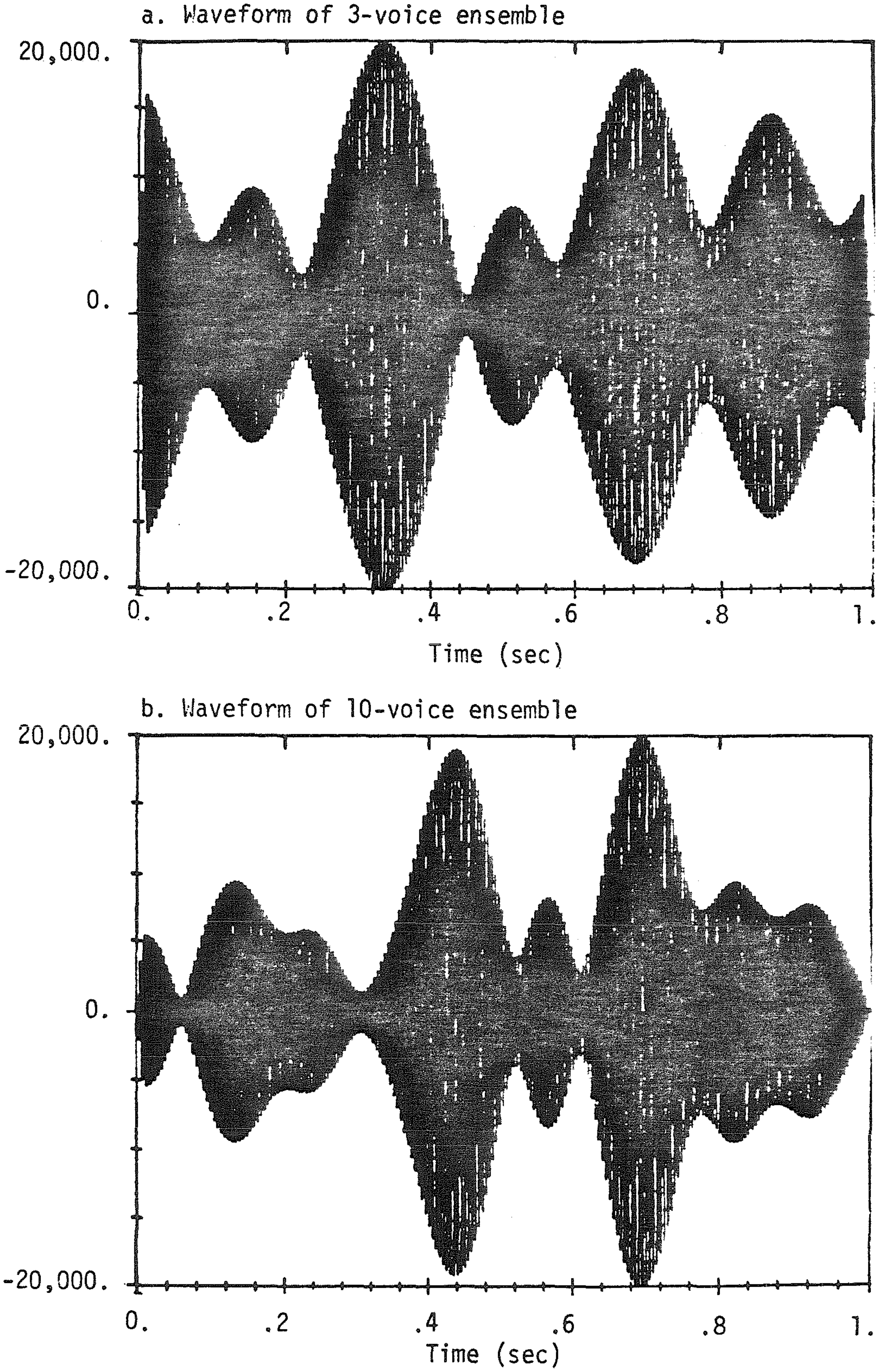
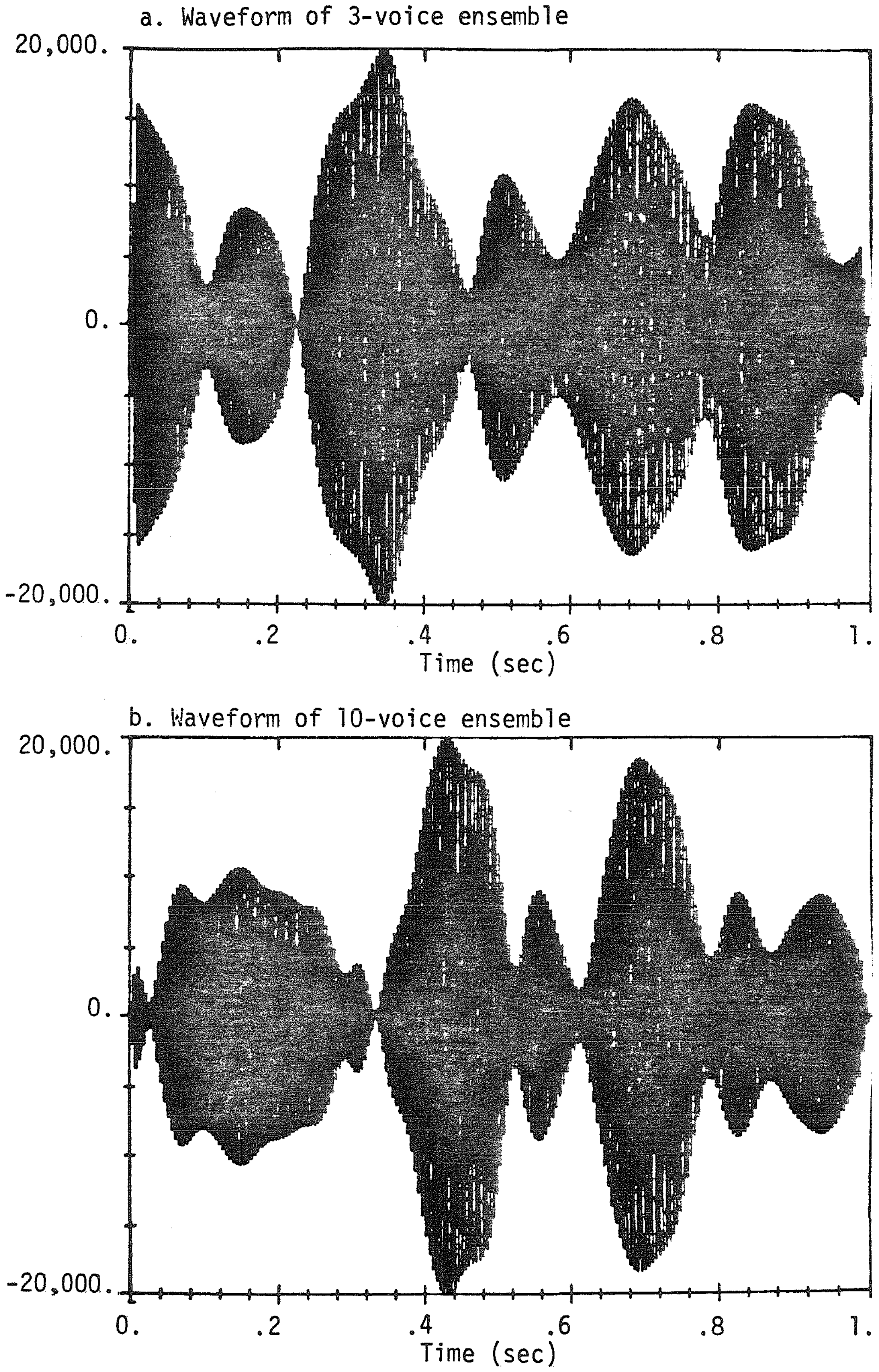


Figure 43. Waveforms of simulated ensembles. (For clarity, each voice consists of only one harmonic.)



a. Waveform of 3-voice ensemble

b. Waveform of 10-voice ensemble

Figure 44. Waveforms of the simulated ensembles of Figure 43 with sinusoidal vibrato of 1% on each solo voice.

have a definite ensemble timbre, yet they retain a distinctly artificial character. However, this could be due simply to the very artificial nature of the Moog solo sounds. Our first test was therefore to apply the Moog choir effect algorithm to real violin sounds.

To determine the algorithm, we used the tracking phase vocoder to analyze recorded examples. We assumed that the simulated ensemble $x(t)$ was the sum of a known solo $x_1(t)$ and a variably time-delayed version $x_2(t)$ for which the time delay $\Delta(t)$ was the unknown to be determined. The phase vocoder frequency estimate for the fundamental was

$$f(t) = 518 + 3\cos(2\pi 6t) + \cos(2\pi \frac{1}{2} t)$$

We took this to be the average of the solo frequency of 518 Hz and the variably time-delayed frequency of

$$f_2(t) = 518 + 6\cos(2\pi 6t) + 2\cos(2\pi \frac{1}{2} t)$$

We then had

$$\varphi_2(t) = \int f_2(t)\, dt$$

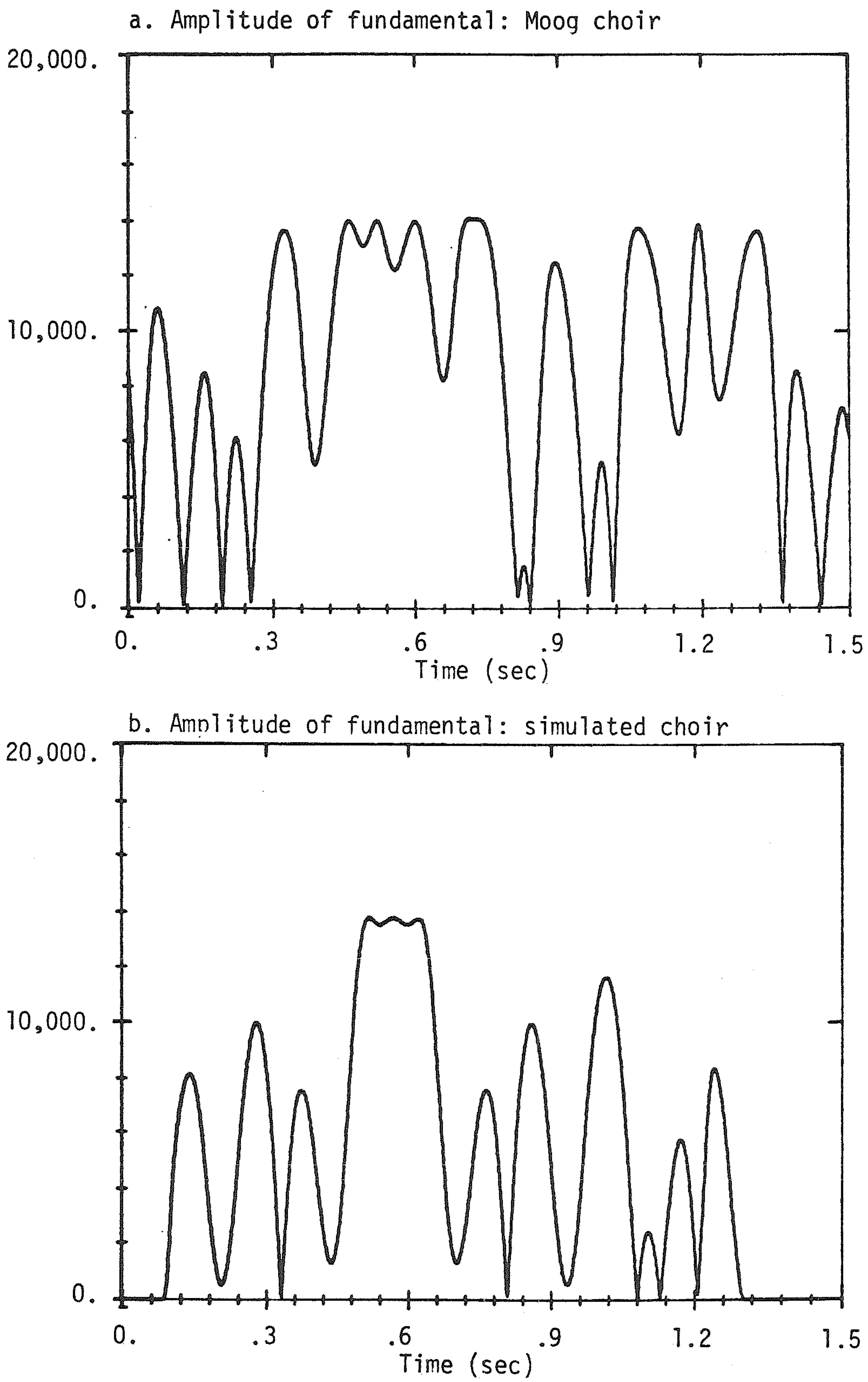
$$\varphi_2(t) = 518\, t + \frac{1}{2\pi} \sin(2\pi 6t) + \frac{4}{2\pi} \sin(2\pi \frac{1}{2} t)$$

Furthermore, we had

$$x_2(t) = \sin(2\pi 518(t + \Delta(t)))$$
$$x_2(t) = \sin(2\pi 518t + 2\pi \varphi_2(t))$$

Hence, the variable time delay $\Delta(t)$ was given simply by $\frac{\varphi_2(t)}{518}$.

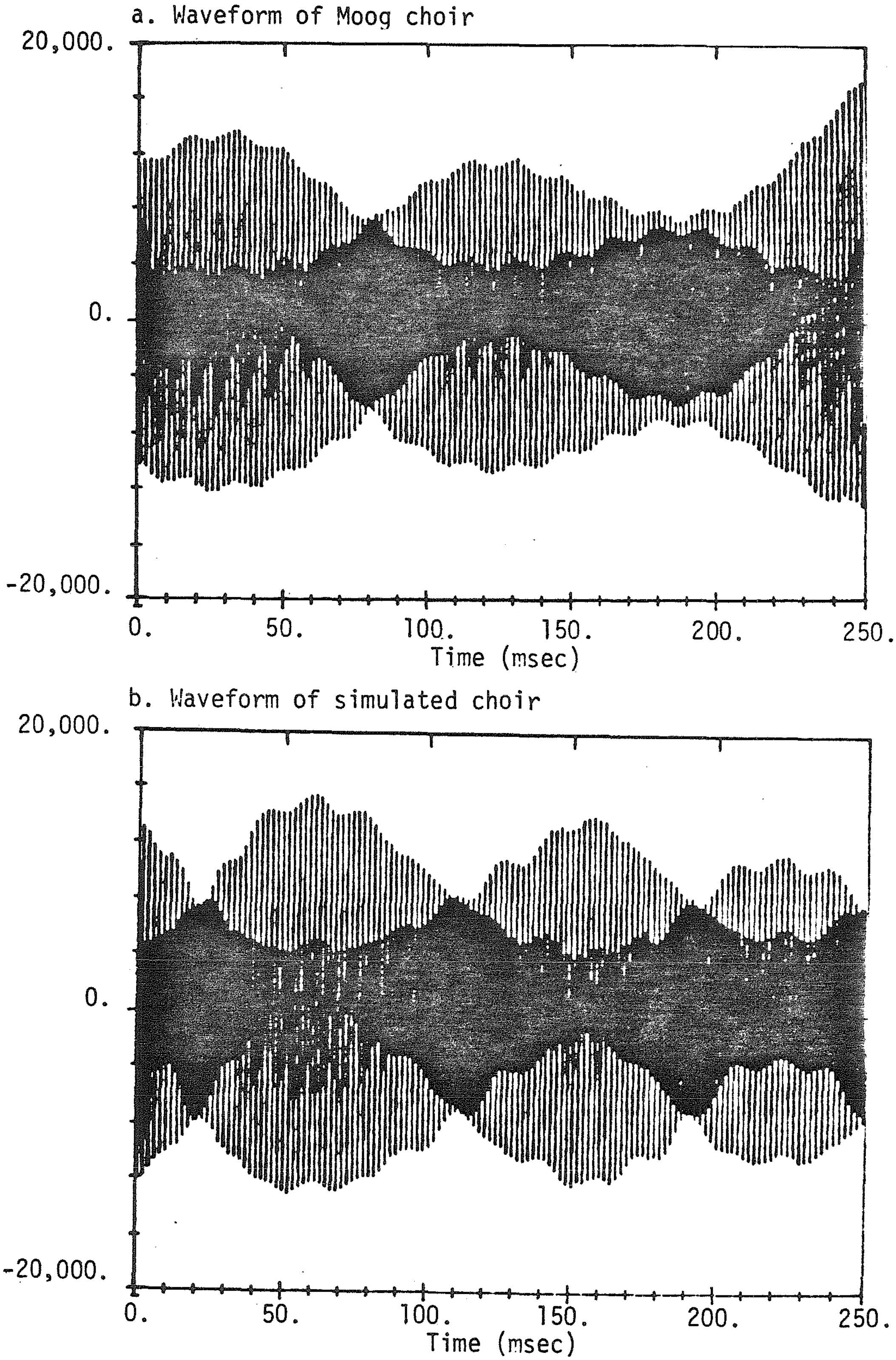
We found that ensemble simulations using this $\Delta(t)$ were extremely similar to the Moog examples both perceptually and analytically (Figures 45 and 46). However, applying this algorithm to actual violin waveforms produced highly variably results. More research is needed to determine the explanation for this, but it seems possible that a more randomly varying time delay and/or an



a. Amplitude of fundamental: Moog choir

b. Amplitude of fundamental: simulated choir

Figure 46. Phase vocoder amplitude estimates for the fundamentals of the waveforms in Figure 45.


a. Waveform of Moog choir

b. Waveform of simulated choir

Figure 45. Comparison of Moog choir waveform with simulated choir waveform.

additional independently varying time delay may be required.

A novel method of ensemble simulation suggested by the results in this report is to use the additive synthesis technique with appropriate amplitude modulation automatically included. In applications where high quality is essential, this could provide an attractive alternative to the summation of multiple independent voices. However, the variable time delay technique is still more efficient when a solo voice is already available without computation.

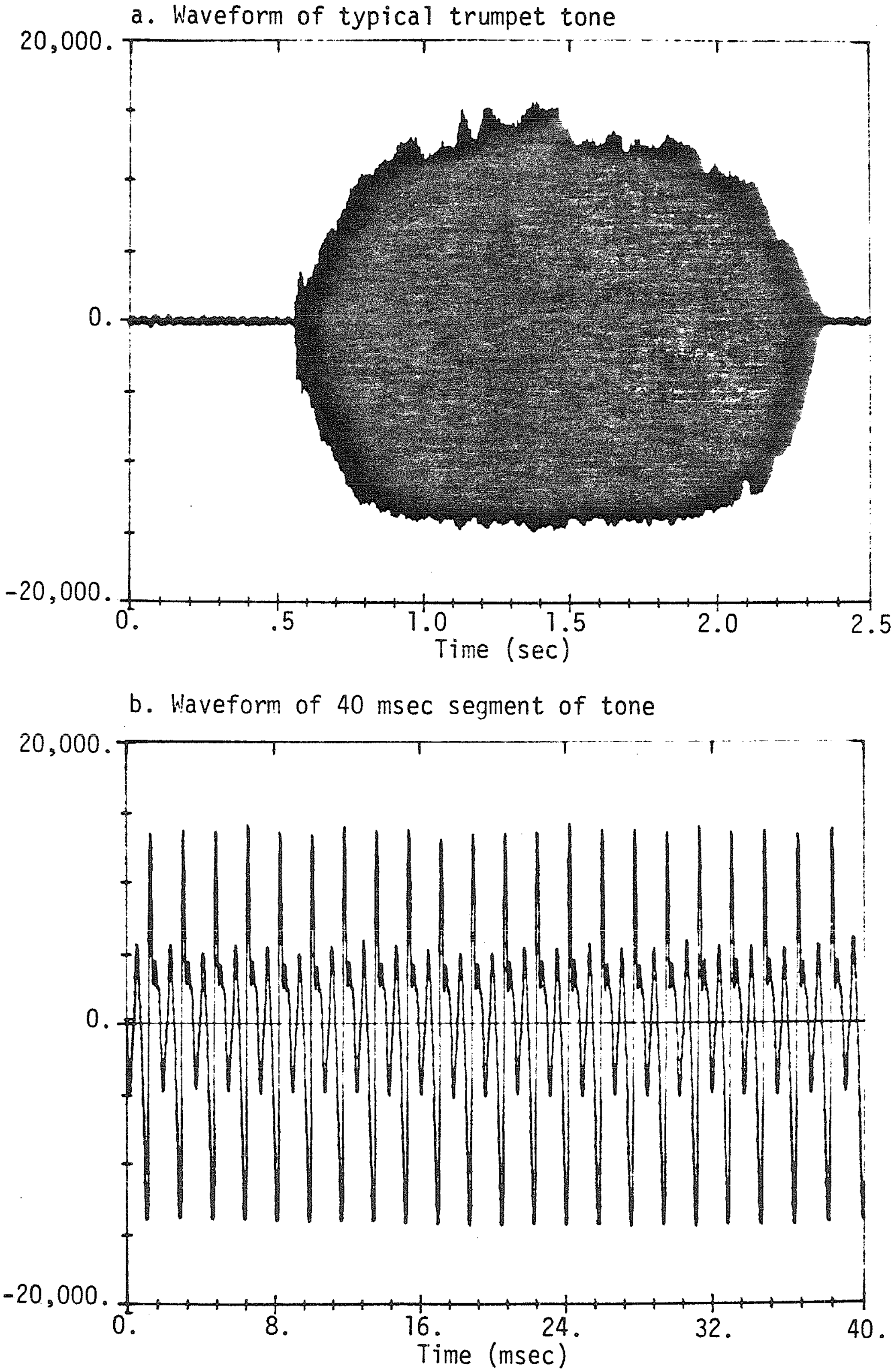
4.10 Conclusions

The above results demonstrate the usefulness of applying an analytical approach to the problem of ensemble sound. In particular, we showed that the ensemble sensation results when there are at least four to eight partials, and when the amplitudes (and, to a lesser extent, the frequencies) of each partial vary in an uncorrelated manner so that the overall average values are still approximately those of the solo waveform. For the amplitude modulation, this variation is a beating which is proportional to the average frequency of the partial. However, several important questions remain unanswered.

First, are these results valid for instruments other than the violin? In our view, it is quite likely that they are; in fact, we have performed preliminary experiments with solo trumpet (Figures 47 and 48) which fully support this assumption. The trumpet is not typically played with vibrato, but there is still a sufficient random frequency variation in a solo tone so that independent solos can be added to produce an ensemble sensation.

Secondly, can these results be employed to create an improved ensemble simulation technique as suggested in Section 4.9? Again, it seems very likely that they can; but an appropriate test would require a professional ensemble, coherent musical examples, and a formal evaluation procedure.

A number of interesting extensions are also possible. For example, an intriguing feature of ensemble sounds is that they manage to retain, to a surprising degree, the distinctive timbre of the underlying solo. Is this based simply on gross spectral differences? An even more interesting question involves ensembles of different instruments playing in unison. Under what conditions can the individual timbres be identified, and when do they fuse into an entirely new timbre? These are questions to be answered by future investigations.



a. Waveform of typical trumpet tone

b. Waveform of 40 msec segment of tone

Figure 47. Waveform of a typical trumpet tone. The tone is C#5 (555 Hz).

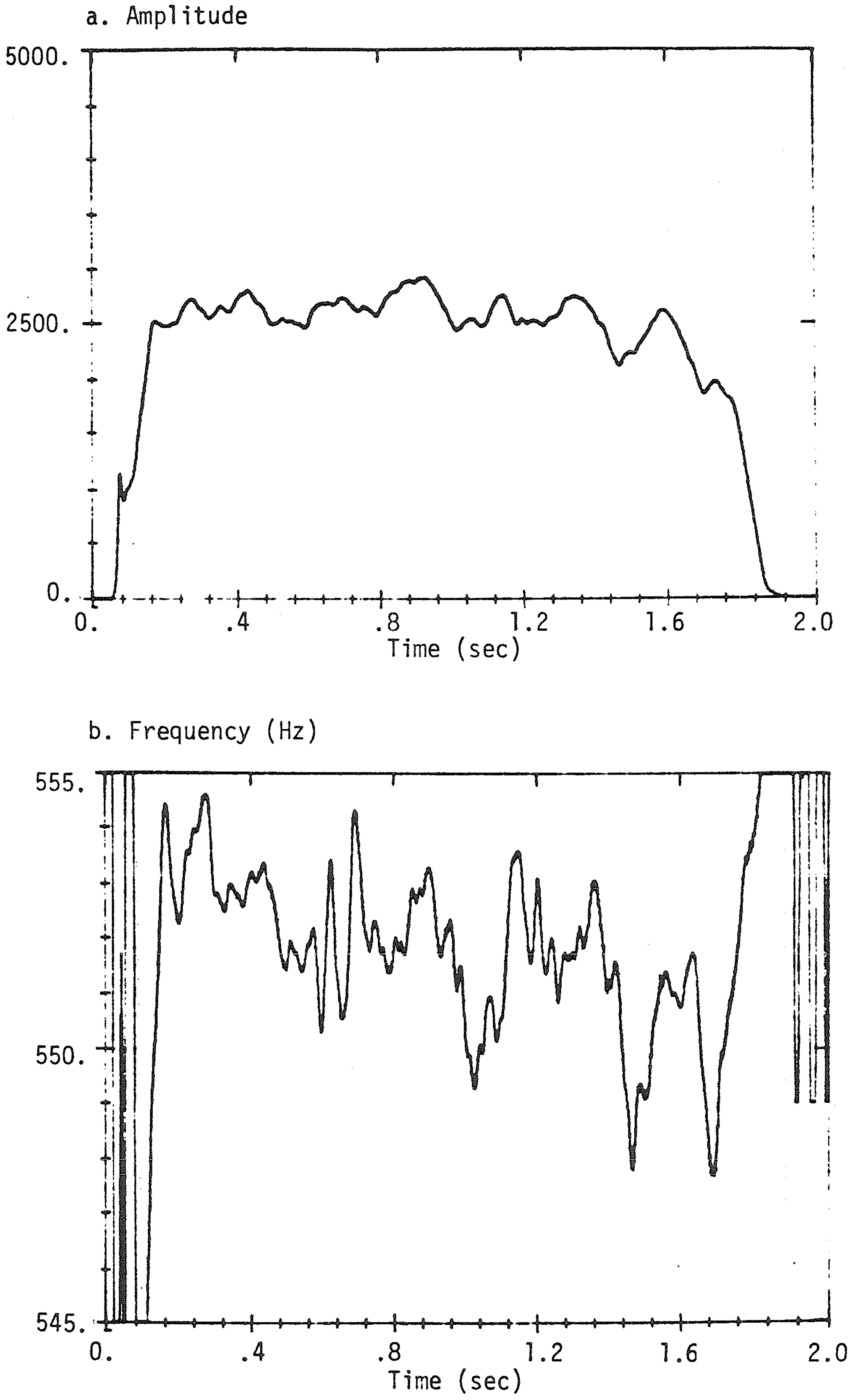


Figure 48. Amplitude and frequency of the fundamental of the trumpet tone of Figure 47.

REFERENCES

Allen, J.B., and Rabiner, L.R., "A Unified Theory of Short-Time Spectral Analysis and Synthesis," *Proceedings of the I.E.E.E.*, 65(11): 1558-1564 (1977)

Arfib, D., "Digital Synthesis of Complex Spectra by Means of Multiplication of Nonlinear Distorted Sine Waves," *J. Audio Eng. Soc.*, 27(10): 757-768 (1979)

Backus, J., *The Acoustical Foundations of Music*, Norton & Co., New York (1969)

Beauchamp, J.W., "A Computer System for Time-Variant Harmonic Analysis and Synthesis of Musical Tones" in *Music by Computers*, H. Von Foerster and J.W. Beauchamp, eds., Wiley, New York (1969)

Beauchamp, J.W., "Synthesis by Spectral Amplitude and Brightness Matching of Analyzed Musical Instrument Tones," *J. Audio Eng. Soc.*, 30(6): 396-406 (1982)

Blesser, B.A., "Digitization of Audio: A Comprehensive Examination of Theory, Implementation, and Current Practice," *J. Audio Eng. Soc.*, 26(10): 739-771 (1978)

Blesser, B.A., "Perceptual Issues in Digital Processing of Music" in *Proceedings of the I.E.E.E. International Conference on Acoustics, Speech, and Signal Processing 1981*, Atlanta, Georgia, March 28 - April 1 (1981)

Cahn, C.R., "Combined Digital Phase and Amplitude Modulation Communication Systems," *I.R.E. Trans. Comm. Sys.*, CS-8: 150-155 (1960)

Charbonneau, G.R., "Timbre and the Perceptual Effects of Three Types of Data Reduction," *Computer Music Journal*, 5(2): 10-19 (1981)

Chowning, J.M., "The Synthesis of Complex Audio Spectra by Means of Frequency Modulation," *J. Audio Eng. Soc.*, 21(7): 526-534 (1973)

Flanagan, J.L. and Golden, R.M., "Phase Vocoder," *Bell System Technical Journal*, 45: 1493-1509 (1966)

Flanagan, J.L., "Parametric Coding of Speech Spectra," *J. Acoust. Soc. Am.*, 68(2): 412-419 (1980)

Flanagan, J.L., and Christensen, S.W., "Computer Studies on Parametric Coding of Speech Spectra," *J. Acoust. Soc. Am.*, 68(2): 420-430 (1980)

Fletcher, H., Blackham, E.D., and Stratton, R., "Quality of Piano Tones," *J. Acoust. Soc. Am.*, 34(6): 749-761 (1962)

Fletcher, H., Blackham, E.D., and Christensen, D.A., "Quality of Organ Tones," *J. Acoust. Soc. Am.*, 35(3): 314-325 (1963)

Fletcher, H., Blackham, E.D., and Geertsen, O.N., "Quality of Violin, Viola, and Cello Tones," *J. Acoust. Soc. Am.*, 37(5): 851-863 (1965)

Fletcher, H., and Sanders, L., "Quality of Violin Vibrato Tones," *J. Acoust. Soc. Am.*, 41(6): 1534-1544 (1967)

Freedman, M.D., "A Technique for Analysis of Musical Instrument Tones," Ph.D. Thesis, University of Illinois, Urbana, Illinois (1965)

Grey, J.M., "An Exploration of Musical Timbre," Ph.D. Thesis, Stanford University, Stanford, California (1975)

Grey, J.M., "Multidimensional Perceptual Scaling of Musical Timbres," *J. Acoust. Soc. Am.*, 61(5): 1270-1277 (1977)

Grey, J.M., and Moorer, J.A., "Perceptual Evaluations of Synthesized Musical Instrument Tones," *J. Acoust. Soc. Am.*, 62(3): 454-462 (1977)

Grey, J.M., and Gordon, J.W., "Perceptual Effects of Spectral Modifications on Musical Timbres," *J. Acoust. Soc. Am.*, 63(5): 1493-1500 (1978)

Grey, J.M., "Timbre Discrimination in Musical Patterns," *J. Acoust. Soc. Am.*, 64(2): 467-472 (1978)

Harris, F.J., "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform," *Proceedings of the I.E.E.E.*, 65(1): 51-83 (1978)

Helmholtz, H.L.F., *On the Sensations of Tone*, Dover, New York (1954)

Justice, J.H., "Analytic Signal Processing in Music Computation," *I.E.E.E. Trans. Acoustics, Speech, and Signal Processing*, ASSP-27(6): 670-684 (1979)

Keeler, J.S., "Piecewise-Periodic Analysis of Almost-Periodic Sounds and Musical Transients," *I.E.E.E. Trans. Audio and Electroacoustics*, AU-20(5): 338-344 (1972)

Lansky, P., and Steiglitz, K., "Synthesis of Timbral Families by Warped Linear Prediction" in *Proceedings I.E.E.E. International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, Georgia, March 28 - April 1 (1981)

Le Brun, M., "Digital Waveshaping Synthesis," *J. Audio Eng. Soc.*, 27(4): 251-265 (1979)

Le Caine, H., "Electronic Music," *Proceedings of the I.R.E.*, 44: 457-478 (1956)

Lindsay, P., and Norman, D. *Human Information Processing*, Academic Press, New York (1972)

Luce, D.A., "Physical Correlates of Nonpercussive Musical Instrument Tones," Ph.D. Thesis, M.I.T., Cambridge, Mass. (1963)

Malah, D., "Time-Domain Algorithms for Harmonic Bandwidth Reduction and Time Scaling of Speech Signals," *I.E.E.E. Trans. Acoustics, Speech, and Signal Processing*, ASSP-27(2): 121-131 (1979)

Mathews, M.V., "Analysis of Musical Instrument Tones," *Physics Today*, 22(2): 23-30 (1969)

Mathews, M.V., and Kohut, J., "Electronic Simulation of Violin Resonances," *J. Acoust. Soc. Am.*, 53(6): 1620-1626 (1973)

McAdams, S., "The Effect of Spectral Fusion on the Perception of Pitch for Complex Tones," Abstract in *J. Acoust. Soc. Am.*, 68(S1): 109 (1980)

McClellan, J.H., Parks, T.W., and Rabiner, L.R., "A Computer Program for Designing Optimum FIR Linear Phase Digital Filters," *I. I.E.E.E. Trans. Audio and Electroacoustics*, AU-21(6): 506-526 (1973)

Miller, J.E., "Measurements of Violin Vibrato," Abstract in *J. Acoust. Soc. Am.*, 70(S1): 23 (1981)

Moorer, J.A., "The Optimum Comb Method of Pitch Period Analysis of Continuous Digitized Speech," *I.E.E.E. Trans. Acoustics, Speech, and Signal Processing*, ASSP-22(5): 330-338 (1974)

Moorer, J.A., "On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer," Ph.D. Thesis, Stanford University, Stanford, California (1975)

Moorer, J.A., "The Synthesis of Complex Audio Spectra by Means of Discrete Summation Formulae," *J. Audio Eng. Soc.*, 24(9): 717-727 (1976)

Moorer, J.A., "Signal Processing Aspects of Computer Music: A Survey," *Proceedings of the I.E.E.E.*, 65(8): 1108-1137 (1977)

Moorer, J.A., "The Use of the Phase Vocoder in Computer Music Applications," *J. Audio Eng. Soc.*, 26(1/2): 42-45 (1978)

Moorer, J.A., "The Use of Linear Prediction of Speech in Computer Music Applications," *J. Audio Eng. Soc.*, 27(3): 134-140 (1979)

Petersen, T.L., "Dynamic Sound Processing" in *Proceedings A.C.M. Computer Science Conf. 1976*, Anaheim, California, February 10 - 12 (1976)

Petersen, T.L., and Boll, S.F., "Critical Band Analysis-Synthesis" in *Proceedings I.E.E.E. International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, Georgia, March 28 - April 1 (1981)

Piszczalski, M., and Galler, B., "Automatic Music Transcription," *Computer Music Journal*, 24-31 (1977)

Piszczalski, M., Galler, B., Bossemeyer, R., and Hatamian, M., "Performed Music: Analysis, Synthesis, and Display by Computer," *J. Audio Eng. Soc.*, 29(1/2): 38-46 (1981)

Portnoff, M.R., "Implementation of the Digital Phase Vocoder Using the Fast

Fourier Transform," *I.E.E.E. Trans. Acoustics, Speech, and Signal Processing*, ASSP-24(3): 243-248 (1976)

Portnoff, M.R., "Time-scale Modification of Speech Based on Short-Time Fourier Analysis," Ph.D. Thesis, M.I.T., Cambridge, Mass. (1978)

Portnoff, M.R., "Time-Frequency Representation of Digital Signals and Systems Based on Short-Time Fourier Analysis," *I.E.E.E. Trans. Acoustics, Speech, and Signal Processing*, ASSP-28(1): 55-69 (1980)

Rabiner, L.R., and Gold, B., *Theory and Application of Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, New Jersey (1975)

Rabiner, L.R., and Schafer, R.W., *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, New Jersey (1978)

Rasch, R.A., "Synchronization in Performed Ensemble Music," *Acustica*, 43: 121-131 (1979)

Risset, J.C., *Computer Study of Trumpet Tones*, Bell Telephone Laboratories, Murray Hill, New Jersey (1966)

Risset, J.C., *Musical Acoustics*, Centre Georges Pompidou Rapports IRCAM No. 8, Paris (1978)

Saunders, S., "Improved FM Audio Synthesis Methods for Real-Time Digital Music Generation," *Computer Music Journal* 53-55 (1977)

Schafer, R.W., "Echo Removal by Discrete Generalized Linear Filtering," Ph.D. Thesis, M.I.T., Cambridge, Mass. (1969)

Schafer, R.W., and Rabiner, L.R., "Design and Simulation of a Speech Analysis-Synthesis System Based on Short-Time Fourier Analysis," *I.E.E.E. Trans. Audio and Electroacoustics*, AU-21(3): 165-174 (1973)

Schroeder, M.R., "Digital Simulation of Sound Transmission in Reverberant Spaces," *J. Acoust. Soc. Am.*, 47(2): 424-431 (1970)

Schwartz, M., Bennett, W., and Stein, S., *Communication Systems and*

Techniques, McGraw-Hill, New York (1966)

Strawn, J., "Approximation and Syntactic Analysis and Synthesis of Amplitude and Frequency Functions for Digital Sound Synthesis," *Computer Music Journal*, 4(3): 3-24 (1980)

Tucker, W.H., and Bates, R.H.T., "A Pitch Estimation Algorithm for Speech and Music," *I.E.E.E. Trans. Acoustics, Speech, and Signal Processing*, ASSP-26(6): 597-604 (1978)

Wedin, L., and Goude, G., "Dimension Analysis of the Perception of Instrumental Timbre," *Scand. J. Psychol.*, 13: 228-240 (1972)

Youngberg, J.E., *A Constant Percentage Bandwidth Transform for Acoustic Signal Processing*, University of Utah, Salt Lake City, Utah (1980)