Center for Computer Research in Music and Acoustics May 1975

Department of Music
Report No. STAN-M-3

ON THE SEGMENTATION AND ANALYSIS OF CONTINUOUS MUSICAL SOUND BY DIGITAL COMPUTER

by

James A. Moorer

Research sponsored by

Advanced Research Projects Agency
ARPA Order No. 2494
and
National Science Foundation

DEPARTMENT OF MUSIC
Stanford University

Center for Computer Research in Music and Acoustics May 1975

Department of Music
Report No. STAN-M-3

ON THE SEGMENTATION AND ANALYSIS OF CONTINUOUS MUSICAL SOUND BY DIGITAL COMPUTER

by

James A. Moorer

ABSTRACT

An examination of the problem of producing a written score from a piece of polyphonic music has been done with the result that a program to accomplish this end for a restricted class of input samples has been written and debugged. The program uses bandpass filtering to extract individual harmonics of each instrument, and then infers the notes from this harmonic data. Two examples are presented here that show the viability of the system given the restrictions on the music under analysis.

This research was supported by the Advanced Research Projects Agency of the Department of Defense under Contract DAHC 15-73-C-0435 and the National Science Foundation under Contract NSF DCR 75-00694 . The views and conclusions contained in this document are those of the author(s) and should not be interpreted as necessarily representing the official policies, either expressed or implied, of Stanford University, ARPA, NSF, or the U. S. Government.

Reproduced in the U.S.A. Available from the National Technical Information Service, Springfield, Virginia 22151.

ON THE SEGMENTATION AND ANALYSIS OF CONTINUOUS MUSICAL SOUND BY DIGITAL COMPUTER

(ABSTRACT)

James Anderson Moorer, Ph.D.

Stanford University, 1975

The problem addressed by this dissertation is that of the transcription of musical sound by computer. A piece of polyphonic musical sound is digitized and stored in the computer. A completely automatic procedure then takes the digitized waveform and produces a written manuscript which describes in classical musical notation what notes were played. We do not attempt to identify the instruments involved. The program does not need to know what instruments were playing.

It would appear that it is quite difficult to achieve human performance in taking musical dictation. To simplify the task, certain restrictions have been placed on the problem: (1) The pieces must have no more than two independent voices. (2) Vibrato and glissando must not be present. (3) Notes must be no shorter than 80 milliseconds. (4) The fundamental frequency of a note must not coincide with a harmonic of a simultaneously sounding note of a different frequency. The first three conditions are not inherent limitations in the procedures, but were done simply for convenience. The last condition would seem to require more study to determine the cues that human listeners use to distinguish, for example, notes at unison or octaves. Numerous other lesser restrictions were also imposed on the music to be analysed.

The method used for this analysis is a directed bank of sharp-cutoff bandpass filters. First, a pitch detector is used to determine the harmony of the piece at each point in time. Using the harmony information, the frequencies of a band of bandpass filters is determined so as to assure that every harmonic of every instrument will pass through at least one of the filters. The output of each filter is processed by a pitch detector and an energy detector. This gives power and frequency information as functions of time. Each power and frequency function pair is rated as to its quality. The rating takes into account the constancy of the frequency function, the smoothness of the power function, and several other measurements on the functions. This rating is used to eliminate spurious traces and null filter outputs.

Notes are then inferred from groups of power and frequency function pairs that occur simultaneously with frequencies that are harmonically related. Notes with higher overall ratings are preferred over other note hypotheses. The melodies are then grouped by separating the notes into the higher voice and the lower voice. Voice crossings are not tracked. For the final manuscripting, Professor Leland Smith's MSS program was used. The analysis program produces directly input for the manuscripting program, thus the entire procedure is automated.

In addition to the above described system, many other techniques were examined for their utility in this task. Each technique that was explored is described and analysed, with a description of why it was not found useful for this task.

One interesting observation is that there is considerably more activity in a piece of music than is perceived by the listner. This is especially common with stringed instruments, because the strings that are not being manipulated invariably resonate and produce sounds independently which are generally not heard due to aural masking. This indicates that perhaps we should use more perceptually-based techniques to help determine what would actually be heard in a piece of music, rather than determine exactly what is there, although detailed descriptions of the contents of the piece may be useful for other purposes, such as music education or musicology.

In general, the system works tolerably well on the restricted class of musical sound. Examples are shown which demonstrate the viability of the system for different instruments and musical styles. Since the procedure is extremely costly in terms of computer time, only a limited number of examples could be processed. These examples are discussed with a description of how the system could be improved and how the restrictions might be eliminated by better processing techniques.

Approved for publication:

By _____

For Major Department

By _____

Dean of Graduate Studies

ACKNOWLEDGEMENTS

This work would, of course, not have been possible without the generous support of the Stanford Artificial Intelligence Laboratory, especially Messrs Lester Earnest and John McCarthy. The support and encouragement of the computer music group has been especially appreciated: Messrs John Chowning, John Grey, and Loren Rush have been helpful at every step of the way. Thanks must be given to Leland Smith for writing and explaining his marvelous manuscripting program in all its many-splendored glory. Great thanks must be given to my dog, who showed limitless patience and understanding through the last few years.

TABLE OF CONTENTS

INTRODUCTION

STATEMENT OF THE PROBLEM

The problem addressed by this dissertation is the machine perception of polyphonic music. We seek to play a piece of music into the computer via an analog-to-digital converter and have the computer return an abbreviated score of the piece. In order to simplify the task, certain restrictions have been placed on the goals. First, we do not require the computer to identify the instruments involved. Second, we do not allow glissandi, fast trills, or exceptionally fast notes (less than 100 milliseconds duration). Third, the class of instruments that we will accept is limited to a subset of the orchestral instruments which excludes drums, gongs, cymbals, and other instruments with inharmonic overtones. Fourth, vibrato must be non-existent or very limited. Fifth, the program will only be expected to track a small number of independent voices (two at most). Sixth and last, we must disallow notes such that the fundamental of one note is at the same frequency as a harmonic of another note. This rules out notes at octaves, at twelfths, and many other intervals. Some of these restrictions represent inherent limitations in the methods used and some merely represent restrictions for the sake of economy. A discussion of each restriction will accompany its intoduction.

In performing this task, there are some things that we may require of the computer that we would not require of a human. One is that the pitches be identified with the actual note relative to the equal tempered scale based on A4 being 440 Hz. This would require the skill of "absolute pitch" which is somewhat rare even among trained musicians. Conversely, there are some things which people do quite well that we cannot at this time reasonably ask the computer to do, such as identify the instruments involved. The reasons why this is a difficult problem will be treated later.

A computerized musical scribe probably has its greatest application in the field of Ethnomusicology, where often hundreds of hours of recorded ethnic music are commonly transcribed by hand. A more long term application is in the field of computer music, where we might expect the computer to be able to perceive music as well as play it, thus taking its cues from the musicians (or other computers?) with whom (which?) it is playing.

ON MUSIC ANALYSIS

Music may be analysed for any number of purposes. There is analysis of a score for form, motifs, harmony, style, etc. These may be termed *high-level analyses* because they deal with concepts which are not rigorously defined, nor are they generally amenable to direct mathematical analysis. These analysis techniques are commonly taught to undergraduate music students as regular curriculum subjects. Some attempt has been made to use the computer to do high-level analysis from scores which have been typed in by hand [Hiller 1966, 1967; Jackson 1967; Winograd 1968] with some success. Perhaps the greatest contribution of the computer has been to the ethnomusicologist who seeks to classify the intervals or frequencies-of-occurrence of motifs.

Analysis of the acoustic waveform itself has been done for the purpose of gaining insight into the physics of music-related hardware (instruments, concert halls, musicians), for the purpose of simulation of musical tones (a musical "vocoder"), for gaining insight into human perception of musical sound, and finally, for the purpose of detecting and tracking the pitch of a single-voiced piece. This analysis might be termed *low* to *intermediate-level analysis* because it deals with musical sound on an acoustical level rather than on the level represented by the score of the piece.

It is, of course, an impossible task to recreate exactly the score that produced a given piece of music. When we listen to a piece of music, we cannot tell that a given note duration represents a quarter note, a half note, or whatever. The composer is free to introduce factors of two in the notation at will, and the conventions in this respect have changed over the years. Also, the amount of voice doubling on a particular line is often quite difficult for people to determine. Sometimes, a precisely played octave will not be recognized as such.

It is our intention to finesse these difficulties by restricting the range of pieces that will be accepted. With some restrictions in effect, the problem is manageable.

WHAT IS MUSICAL SOUND?

INSTRUMENTS, OVERTONES, AND A MODEL OF INSTRUMENT WAVEFORMS

Our model of music will consist of the sound pressure wave created by a finite number of instruments that play notes which begin at some time, have a finite duration, and are nearly periodic in that interval. For our purposes, an instrument will be defined as something which produces nearly periodic sound pressure waves. A note will be defined entirely by its pitch, starting time, duration, and loudness. Before we proceed any further, some definitions are in order:

pitch - Pitch is a subjective quality of sound that is not necessarily dependent upon the existence of a sinusoid at that frequency. A discussion of pitch perception is offered in later sections (see section entitled *Music Perception*), so please accept for now that "pitch" means what we commonly take it to mean, but "frequency" refers to the repetition rate of a perfectly periodic signal. "Frequency" is a physical quantity which can be measured objectively. "Pitch" is a perceptual phenomenon.

harmonic - A perfectly periodic waveform can be decomposed by Fourier's sine and cosine series into a sum of sinusoids whose frequencies are integral multiples of some base frequency, which is called the "fundamental" frequency of the sound. These sinusoids are described as "harmonically related" sinusoids, or more simply as "harmonics."

inharmonic - An adjective meaning "not harmonically related."

partial - Many waveforms are not periodic, but may nonetheless be represented by a sum of sinusoids that are not harmonics. The general term for the sinusoids which make up a waveform, be they harmonic or otherwise, is a "partial tone," or more simply, a "partial."

quasi-periodic - This term along with "nearly harmonic" applies to waveforms which are not perfectly periodic, but are very close. Stringed instruments show some inharmonicity due to effective shortening of the string at higher frequencies, but since the deviation is just a few percent, they are called "quasi-periodic."

half-step - This is the square root of a step, or the twelfth root of 2, which is 1.05946309. The half-step is the relation between the frequencies of notes which are played on adjacent keys on a piano keyboard. The half-step forms the basis of most Western music. This is also the basis of the *equal-tempered* scale, which is used throughout this thesis.

step - A "step" is a ratio of two frequencies which is defined as the sixth root of 2, or 1.12246205.

interval - The relation between the frequencies of two simultaneously sounding notes is called an "interval". We measure intervals in terms of steps, or half steps. The intervals consisting of integral numbers of half-steps have names and special meanings in most western music. If the frequency of one note is f_1 and the frequency of the other note is f_2, then the "distance" between those two notes in half-steps is simply $12*\log_2(f_2/f_1)$. This is the interval those two notes represent.

scale - A manner of subdividing a large interval, such as an octave, at definite points in order to provide a series of tones suitable for melodic or harmonic use. Two common divisions of the octave in Western music are the *major* and the *minor* scales, each of which divide the octave into eight notes (including the endpoints). If we number the notes of these scales from the lowest to the highest, the major scale has a half-step between the 3rd and 4th notes, and between the 7th and 8th notes, and whole steps between the other adjacent notes of the scale. The minor scale has half-steps between the 2nd and 3rd notes and between the 5th and 6th notes.

chord - Three or more notes sounding simultaneously. In more common usage, the intervals between adjacent notes is 3 or 4 half-steps (these intervals are called *minor* and *major* thirds, respectively). A more general term for the simultaneous sounding of three or more notes without regard for the intervals among them is a "cluster".

harmony - This is easy to confuse with "harmonic," but it refers to a subjective musical quality. When two or more instruments play different notes at the same time, we refer to the relation of the notes as the "harmony" of the music. To be more specific, this is actually the *vertical* harmony of the music. We may also define the *horizontal* harmony to be the relations among the chords as a progression in time. In this dissertation, we shall only be concerned with vertical harmony, although horizontal harmony is much more important musically.

Music instruments can be divided into many categories, but we shall only distinguish two: those that have nearly harmonic partials and those that do not. We shall be concerned here with only those instruments which have nearly harmonic partials. These instruments can be modeled as a sum of sinusoids with slowly-varying amplitudes and frequencies. The frequencies of these sinusoids are very close to integral multiples of the fundamental frequency of the note.

With the aid of the heterodyne filter (see section *Heterodyne Filter* in *Low-Level Techniques*), we can examine the behavior of the amplitudes and frequencies of notes played in isolation.

With these data available, we are in a position to test the validity of the model for describing the perceptually relevant attributes of music instrument tones. We can do this by resynthesizing the tones and comparing them with the original tones. We have done this for the following instruments:

violin, viola, cello, double bass, trumpet, trombone, French horn, baritone horn, oboe, English horn, bassoon, Bb clarinet, alto clarinet, bass clarinet, flute, alto flute, alto sax, soprano sax

The synthetic tones are very similar to the originals. When some white noise is added into the synthetic tones to simulate the effect of tape recorder hiss, most of the synthetic tones are extremely similar to the original. This affirms the validity of the model and of the heterodyne filter for representing this class of instruments. Although we have not done this test on every music instrument with nearly harmonic partials, we have no reason to believe that this model should not be adequate for representing all such instruments, including the human voice (possibly excepting frication).

That these instruments can be represented in this manner is somewhat curious, because some of the instruments exhibit inharmonicity. The heterodyne filter is not capable of detecting inharmonicity directly. It would appear that these effects show up as amplitude and frequency modulation on some harmonics. Since the sum of two sinusoids is identical to a single amplitude-modulated sinusoid, much of the effect of inharmonic partials seems to be captured in the detail of the amplitude and frequency contours for each harmonic.

A great body of work on music instrument tones in isolation is presented in a companion dissertation *An Exploration of Musical Timbre* by John M. Grey [1975]. The heterodyne filter was used to analyze a number of different instruments as a method of generating psychoacoustic stimuli for studying human perception of timbre. Figures 26 and 27 were taken from his work.

ON MUSICAL HARMONY

There is a well developed body of harmonic practice which is taught as an undergraduate music course [Piston 1941; Forte 1962]. This is generally referred to as "classical" or "traditional" harmony. Again, there is a difference between "vertical" and "horizontal" harmony. We shall only deal with "vertical" harmony, which does not take contextual information into account.

We shall discuss the mathematical implications of some aspects of harmony, notably the chord. The simplest chord is the triad. The triad consists of three notes sounding simultaneously. The most common triads are the major triad, and the minor triad. These are defined by the ratios of the frequencies of the notes in the triad. One simple form of the major triad in "root" position has the next higher note (which is called the "third" of the chord) located four half-steps higher than the lowest note, which is called the "root" of the chord. The third of the chord is so-called because it is the third note of a major scale which begins at the root. The highest note of the major triad is called the "fifth" of the chord and is located 3 half-steps higher than the third which makes a total of seven half-steps higher than the root. The "harmony" of a piece of music can be thought of (in an oversimplified manner) as the progression of chords in a piece of music.

One of the things that makes music interesting is the fact that we may shuffle the notes of a chord up or down by some number of octaves and still have the same (in a certain sense) chord. There are names for many of the arrangements of notes that define a given chord. For instance, if the third is the lowest note in a chord, the fifth the next higher, and the root the highest, the chord is said to be in the "first inversion". Likewise, if the fifth is the lowest, the chord is in the "second inversion". This discussion is a bit oversimplified, in that the *inversion* of a chord depends only on the lowest sounding note. For instance, a chord can still be in root position if the third of the chord is raised an octave.

One might ask why a chord such as a major triad is so important in western music. Why wouldn't any combination of frequencies do? This question has as yet not been answered. It is not clear, for instance, whether the special nature of the major triad is "universal" or is a manifestation of cultural bias. Despite the complexity of the problem, several interesting observations have been made. One may observe that in the harmonic series for a particular frequency, the 4^{th}, 5^{th}, and 6^{th} harmonics of a note form a major triad. The 6^{th}, 7^{th}, and 9^{th} harmonics form a minor triad. (We should note here that this definition of the minor triad is not quite suitable for musical use, because the 7^{th} harmonic is actually somewhat lower in frequency than the usual definition. The interval between the 6^{th} and 7^{th} harmonics is about 2.67 half-steps, rather than the usual 3 half-steps). It might be more relevant to describe the minor triad in terms of the 4^{th}, 5^{th}, and 15^{th} harmonics. All unambiguous chords fall in the harmonic series somewhere. While we may speculate on mechanisms in the ear that makes listening to chords both natural and pleasant, it is more important to note that each chord can

be thought of as a manifestation of (harmonics of) a fundamental frequency which may well not be present. For each (unambiguous) chord, we can find a frequency whose harmonics will contain all the notes of the chord. The existence of this "fictitious fundamental" makes it possible to determine the harmony of a piece of musical sound without determining the notes that are being played. This can only be done when the harmony is unambiguous. Often composers use ambiguous chords to great advantage. It is also important to note that any interval consisting of an integral number of half-steps will imply one or more fictitious fundamentals. One does not need a full chord.

Methods for determining the harmony of a piece will be discussed in the section on low-level techniques, specifically, the autocorrelation and the optimum-comb.

OVERVIEW OF THE ANALYSIS SYSTEM

The musical scribe has been realized for a limited class of musical inputs. The system begins with the digitization of the waveform itself by an analog-to-digital converter, operating at 25,600 samples per second to a precision of 14 binary bits. The first processing step uses the optimum-comb method to determine the harmony of the piece. This step is not really necessary, but it greatly reduces the amount of computer time used by subsequent steps by reducing the number of possible notes that could be present at any given time. For music which contains notes which do not lie in perfect unambiguous harmonic relationship, more than one possible harmony will be generated by the programs.

The next phase of the analysis involves bandpass filtering the waveform at frequencies which represent the frequencies of all the harmonics of all the notes that might be present in the piece, given the results of the analysis of harmony from the above step. These filtered waveforms are processed to see if a sinusoid is present at or near the expected frequency. If one is found, its amplitude as a function of time is smoothed and approximated by a polynomial and recorded.

The last phase consists of looking at the results of detecting individual sinusoids and inferring what notes must have been present to produce those sinusoids. This last step is the least rigorous, the most heuristic, and the most sensitive link in the chain.

Except for the original digitization and the "beautification" of the final graphical output, the entire system is automatic and runs without human aid or intervention. This was a design criterion. Since the task of taking musical dictation is commonly taught at the freshman and sophomore levels in college, it seemed pointless to insert a human in the processing path when a person could do the entire task much more quickly. The only value the system might have is its ability to do the process all by itself.

In fact, the system computes the pitches of the notes much more accurately than a human could. This is as much a hindrance as it is a blessing when the final score is produced. The human being perceives the pitches to be members of the notes of the scale, even if some of the notes are mistuned. Humans will tolerate, even admire, large deviations from mathematically precise rhythm, yet can write down the original score despite the deviations. Computer synthesized music that does not have this built-in flexibility is often recognizable by the "inhuman" treatment of rhythm given by the mathematically precise rendering of a piece. It is quite difficult for the machine to infer what the original scoring was, based on a totally human performance. For this reason, the output scores can not be expected to be identical to the input score, but will reflect the modifications made by the performer.

For a piece of music that is only a single voice, the detection of pitch is a task which has been treated extensively by the speech understanding and recognition researchers. The topic treated in this thesis goes one step further in attempting to deal with more than one simultaneous voice. The only reason the present implementation is restricted to two voices is that the notes-at-

octaves problem does not appear to have a simple solution. It is not clear how people can distinguish notes whose harmonics overlap entirely.

OVERVIEW OF THIS THESIS

In organizing the thesis, many decisions had to be made concerning how much to include and where to include it. Rather than present just the program itself, a more complete description of the history of music analysis and a discussion of the relation of many common signal-processing techniques to musical sound is included, at the cost of including a large amount of detail on methods that were not included in the final realization. Since the failures can be as revealing as the successes, it is hoped that this additional information will be of use to future researchers who may avoid some duplication of effort.

Since there has been little effort to produce an automated musical scribe, no literature appears on the subject. The only effort known to the author is the Melograph, a special-purpose hardware device built by Inter-Ocean Systems of Santa Barbara. This device makes a graph of the pitch of the input waveform with time. This graph is in fact not a score, but is enough to get an idea of what was being played.

The historical review thus does not (can not) deal extensively with the exact problem at hand. There are, however, many analyses of music, musical instruments, and even musical sound, some of which have been done on the computer. If we temporarily widen our scope to include analysis for purpose of insight and analysis for the purpose of synthesis, then we have an abundance of material for discussion. This is, in fact, what was done. The historical review includes all analyses of musical sound by computer that we found, as well as a review of speech processing literature, a related subject.

While doing the research for this thesis, many techniques were discovered which were not directly useful for the musical scribe, but which had application in other areas of musical sound analysis. These techniques (the heterodyne filter especially) will be described, as well as a discussion of many of the techniques that were not found useful for any aspect of music processing for one reason or another. The latter were included so that future researchers will not spend too much time on known dead ends. To some extent, these are diversions from the subject at hand, but since they were part of the research done in the course of this thesis, it seems reasonable to expose them here.

The thesis is divided into four parts. The introduction (this section), a section on low-level techniques, a section on high-level techniques, and a critical review section.

In the introduction, we give background information as well as a detailed historical review. Readers not familiar with the characteristics of musical sounds may be interested in the section entitled *What is Musical Sound?* The historical review section is followed by a quick summary of pitch perception theory, which comes from the field of psychoacoustics.

The next section is on low-level techniques. These are the algorithms that operate directly on the digitized waveform. They are largely signal-processing techniques, adapted for this special

application. In order, we review the autocorrelation function and the optimum-comb technique. These are useful for periodicity detection and tracking. Their application to the detection of musical harmony is discussed. The heterodyne filter follows with a method for determining the amplitudes and frequencies of the harmonics of a single musical note. This technique has turned out to be very useful for music synthesis, for it can capture all the time-variant information in a musical tone. Next, we review the bandpass filter. Although it is a very old device, its application to musical sound has been little explored in the past. We show several graphs of applications of bandpass filtering to the extraction of a single harmonic from a polyphonic piece of musical sound. The bandpass filter forms the core of the musical transcription system.

In this section we also discuss several signal-processing techniques that were tried but were not found to be entirely useful for the current problem. These include the cepstrum, the discrete Fourier transform, and the linear predictor. The cepstrum and the linear predictor seem to be useful only in the monophonic case. The discrete Fourier transform assumes that the autocorrelation of the input signal is *stationary*. If the signal is changing either in amplitude or frequency, the transform is distorted. This means that any system based upon the discrete Fourier transform could never be extended to encompass vibrato or highly reverberant environments.

Next, we discuss the way we combine the various signal-processing routines to form a complete low-level package for musical transcription. Here we discuss the utility of determining the vertical harmony of the piece as a planning phase for setting up the frequencies of a band of bandpass filters. The filter output is processed with a pitch detector and an energy detector to produce power and frequency functions for the output of each filter. In the planning phase, we assure that every harmonic of every note will be passed by some filter.

The next section deals with intermediate-level techniques. Here we pass from the world of digital signal processing into the world of artificial intelligence. These techniques deal with making sense from the outputs of the bandpass filters, figuring out what notes were present in the input signal, and how best to print these for readability. To allow easy comparison of the filter outputs, we produce a rating of the quality of a given power-frequency function pair. If this rating is properly prepared, we can easily separate the spurious traces from the meaningful ones. We can then hypothesize the existence of notes from their harmonics. We then discuss some of the aspects of manuscripting.

The last section is a critical review of the system. We begin with some examples which show the viability of the system. We then discuss the weak points of the system with suggestions as to how they may be improved. This involves the development of adaptive pitch tracking filters as well as further research in other areas.

HISTORICAL REVIEW

EARLY ANALYSES

There have been many analyses done of music instrument tones, usually in order to gain insight into the physics of a specific instrument. It was not until the advent of electronics that music analysis on a quantitative basis became practical. One of the first examples we have is that of Backhaus [1927, 1932]. His system consisted of a narrow band-pass filter, using a carbon microphone and a 5 vacuum-tube amplifier, connected to a pen and drum recorder. The filter was tuned to the frequency of the harmonic of interest and the bandwidth was set to suppress adjacent harmonics. The drum assembly was brought up to speed by hand (turning a crank). Then all at once, the pen was lowered onto the paper, the threaded shaft that the drum turned on was stopped, leaving the drum to turn and screw itself down (by momentum) and thus cause the pen to leave a helical trace on the paper, and the musician played a single note on his instrument. The drum was apparently massive enough to keep its speed for quite a while. The resulting trace was taken to approximate the behavior of a single harmonic from the instrument. The process was repeated for many harmonics of many different notes. Needless to say, the process was cumbersome enough to prevent great volumes of data from being accumulated. The amplitude of the harmonic with time was then traced and plotted by hand. Since wire recording techniques were not yet perfected, the note had to be played again and again to get all the harmonics. We know now that no two notes are alike in fine structure, thus casting doubt on the details of the results, but the technique did work adequately on the steady-state portion of notes. His principal result was an analysis of violin resonances in an attempt to find out why the Stradivarius was so revered in the music world. This same theme recurs constantly throughout the literature.

The advance of the oscilliscope in the 40's brought about a new wave of research. The steady-state portion of a waveform could be photographedor drawn from the face of the cathode-ray tube, and then analyzed by calculating the Fourier sine and cosine series. The Fourier integrals were often computed by hand, until a mechanical device (the Henrici analyzer) was built to do just that. The operator would trace the curve with the stylus of the device and then just read off the amplitudes of the harmonics on the dials. Analyses of this sort are very common in the literature [Lehman 1964, Parker 1947, Saunders 1946, Fletcher *et al* 1962]. Saunders analysed wind instrument tones to try to determine if the wind instruments exhibited resonances like the string instruments do. He found no evidence of the existence of formants in the instruments he analysed (clarinet, oboe, English horn, French horn, and flute). Parker analysed the tones of wooden and metal clarinets using a mechanical embouchure, finding that there was little difference between wood and metal clarinet tones. Lehman analysed the bassoon in great detail, using the Kay sonagraph, a device consisting of a number of narrow band-pass filters and a recording system that produced bars on a roll of paper that became thicker in proportion to the energy output of each bandpass filter. He concluded that there is a strong formant between 440 and 500 Hz in the bassoon, accompanied by a weaker formant around 1220 and 1280 Hz.

COMPUTER ANALYSES

Let us jump immediately into the computer analysis of music instrument tones, leaving behind the large number of articles which were done without computers. One of the first computer-based analyses of music instrument tones was done by David Luce [1963]. Using the 709 at MIT, he digitized and analyzed tones from a large number of music instruments. Again, this was done for gaining insight into the behavior of the instrument and its possible perceptual implications. Since his analysis technique was the basis for several following works, including our own heterodyne filter, we will describe and analyze it in some detail.

LUCE

The object of Luce's method was to determine the amplitudes and frequencies of each of the harmonics of a tone as functions of time. These were plotted for further study. The method used was to approximate the integrals for the Fourier sine and cosine series by discrete summations. First, the fundamental frequency was determined by filtering the note itself to remove all harmonics except the fundamental. The fundamental was then digitized and the zero crossings were used to compute the frequency. This works in most cases, but sometimes gives errors-of-octave when the energy in the fundamental is very weak. In these cases, the pitch of the note was matched by hand with an oscillator and the waveform from the oscillator was used. This estimate of the fundamental frequency was used to divide up the waveform from the instrument roughly into separate periods. For each period, 24 equally spaced points were selected. Since the period of the signal was not necessarily a multiple of 24 points, linear interpolation was used to generate the values between the sample points. From these 24 points, the Fourier sine and cosine coefficients were generated. This is represented by the following formulae:

$$(1)\quad a_n(m) = \frac{1}{12}\sum_{L=1}^{24} s\left[(m-1)T_0 + \frac{LT_0}{24}\right] \sin\left(2\pi n \frac{L}{24}\right)$$

$$(2)\quad b_n(m) = \frac{1}{12}\sum_{L=1}^{24} s\left[(m-1)T_0 + \frac{LT_0}{24}\right] \cos\left(2\pi n \frac{L}{24}\right)$$

Where $s(t)$ is the input waveform,
T_0 is the period of the input waveform,
m is the number of the period under analysis,
L is the sample number within the period which is from 1 to 24,
and n is the harmonic number.

The result was one pair of coefficients for every period throughout the duration of the waveform. The pair of coefficients were converted to radial form and the magnitudes and

angles were then plotted. To test the validity of the analysis procedure, the magnitudes and angles were used to synthesize a tone. This tone was played through a digital-to-analog converter (DAC) and compared to the original tone. The first problem encountered was the fact that the magnitudes and phases that were sampled once per period lead to a discontinuous waveform. This is because at the beginning of each period, the phases and magnitudes were suddenly changed to the values for that period. If the parameters for this period were significantly different for the previous period, a discontinuity results. This is often the case during the attack and decay portions of a note. This was remedied in part by filtering (digitally) the waveform at a frequency higher than the frequency of the highest harmonic to remove spurious harmonic distortion. The results of listening tests were that the string family was well reproduced, but the brasses suffered a bit. The lowest octaves trumpet, trombone, tuba, and French horn were all noticeably different than the original notes. The notes sounded very rough. This was explained by the insufficiency of using 24 points per period. Since the brass tones have a pulse-like waveform, sometimes the pulse itself occurred between two selected points, thus reducing the magnitudes of the Forier components for that period. This hit-or-miss behavior created great jitter in the magnitudes as functions of time, thus contributing to a rough sound. Similar difficulties were encountered with the clarinet tone.

What we mean by "pulse-like" is that the waveform, in each period, has an initial strong maximum followed by activity of lesser amplitudes throughout the remainder of the period. This can occur if the harmonics of the waveform are all cosines, such that their maxima coincide and reinforce, producing one strong maximum per period.

FREEDMAN

The next set of analysis programs were written by Morris David Freedman at the University of Illinois [1965, 1967, 1968]. In his system, music instrument tones are modeled by the following equation:

$$(3)\quad h_k(t) = \sum_{r=1}^{m} A_{rk}\, u(t-\tau_{rk})\,\{1-e^{-\alpha_{rk}(t-\tau_{rk})}\}$$

$$(4)\quad g(t) = \sum_{k=1}^{m} h_k(t)\, \sin[\omega_k(t-\tau_{1k})+\pi_k]$$

Where $u(t)$ is the unit step function,
k is the harmonic number,
ω_k is the radian frequency of the k^{th} harmonic,
π_k is the phase of the k^{th} harmonic,
τ_{1k} is the beginning time of the k^{th} harmonic,
$h_k(t)$ is the amplitude envelope of the k^{th} harmonic.
A_{rk} is the amplitude of the r^{th} component of the amplitude envelope of the k^{th} harmonic,
τ_{rk} is the beginning time of the r^{th} component of the amplitude envelope of the k^{th} harmonic,
α_{rk} is the time constant of the r^{th} component of the amplitude envelope of the k^{th} harmonic,
$g(t)$ is the signal that is to model the music instrument tone.

This is a sum of sinusoids, not necessarily harmonically related, with piecewise-constant frequencies. The amplitudes of the sinusoids are piecewise sums of exponentials and constants. For synthesis, linear interpolation was used to smoothly change from one frequency value to the next, thus eliminating Luce's problem of discontinuities. To get the parameters of the model from an actual music instrument tone, a three step process was used. The first step gets the phase differences of the harmonics and the average frequency of each harmonic. The second step determins the amplitudes and phases of each of the harmonics as functions of time, guided by the frequencies of the harmonics as computed in the first step. The second step can then be repeated with the new frequency data for a better approximation. This completed the analysis. The amplitude functions of the harmonics were examined for places of great change of slope and these places were taken to be the "breakpoints" for the piecewise-exponential amplitudes as shown above.

The first step of the analysis used what he called the "D-transform." It is defined as follows:

$$(5)\quad D(t,\omega) = \frac{1}{t}\int_{0}^{t} f(\tau)\, e^{-j\omega\tau}\, d\tau$$

Where $f(\tau)$ is the input waveform

This is a Fourier integral of a function that is limited in time to positive values less than t. The second and third steps of the analysis used what he called the "G-transform" which is defined as follows:

$$(6)\quad G(t,\omega) = \int_{t-T}^{t+T} f(\tau)\, e^{-j\omega\tau}\, d\tau$$

Where T is the period of the input waveform.

This is a Fourier integral over one period of the input waveform. This returns the quadrature components which can be used to derive the magnitudes and phases of the harmonics as functions of time. Freedman does not say how often the integral is evaluated, but we assume it is evaluated once per period of the input signal, as Luce did.

Again, the tones were synthesized using the data from the analysis. The trumpet and saxophone tones were judged to be nearly indistinguishable from the originals. The violin was judged the poorest, although it was judged as quite good. In each case, the synthetic tone showed the characteristic quality of the instrument. The violin sounded bowed and the flute sounded "breathy."

BEAUCHAMP, KEELER

Beauchamp, also at University of Illinois, built upon the work of Freedman by using only the G-transform, adding a filtering operation, and using piecewise linear functions to represent the amplitude functions [Beauchamp 1969]. The amplitude functions were filtered with a low-pass filter to remove a characteristic ripple in the functions that was at the frequency of the fundamental. He evaluated the functions "a few" times per period. The amplitude functions were then approximated with piecewise-linear functions. For synthesis, the frequencies (phases) of the harmonics were not varied with time. Just the initial phase angles were preserved. The frequency of the entire tone was allowed to vary in a piecewise-linear fashion, with the ratios between the frequencies of harmonics held constant, as with Luce and Freedman, but explicit and separate control over the frequencies of each of the harmonics was not used.

Since the publication of the above described paper, Beauchamp [personal communication, 1974] has applied the Fast Fourier Transform algorithm (FFT) to the evaluation of the G-transform. This is done by first reducing each period of the input signal to 64 points by linear interpolation, much like Luce, multiplying the signal by a Hamming "window" function [Blackman and Tukey 1959], and then taking the discrete Fourier transform of each period using the FFT algorithm for efficiency.

Keeler [1972] analyzed tones from organ pipes using techniques similar to Beauchamp's published method. He evaluated the Fourier integral numerically using quadratic

approximation by Simpson's rule and Lagrangian interpolation to improve the accuracy. In his method, the worst-case error in the amplitude estimate for a given harmonic was less than 1.25 percent. He was not concerned about the phase as a function of time and thus did not carry along that information. He did not attempt a synthesis of the tones from the analysis data.

THE MELOGRAPH

The computer analysis techniques described above were for the purpose of gaining insight into the properties of instruments or musical waveforms, and simulation of music instrument tones. We have still not described any method of transcribing a piece of music. This is because, to our knowledge, no such analysis has ever been done. The closest we have found is work in speech understanding and recognition, and a peculiar device called the Melograph.

The Melograph is a special-purpose piece of mostly analog hardware and a chart recording scheme which has two purposes. One function it can perform is that of a high-resolution spectrograph. It can simulate 100 bandpass filters and record the energy output of each on the graph. The second function is that of detecting, tracking, and graphing the fundamental frequency of an input waveform with time. It can only operate on a monophonic (one-voice) input signal in a relatively noise-free environment. It accomplishes this by realizing a band of 1/3 octave band-pass filters. The outputs of the filters are scanned every 4 milliseconds from lowest frequency to highest, searching for a maximum in the energy output of a particular filter relative to its neighbors. When the first maximum is found, the output of that filter is assumed to contain the fundamental of the tone. The zero crossings of the output of that filter are counted and that number is used to compute the pitch. This pitch is then plotted on the chart. Since there is no documentation on the operation of the device, this information was obtained by verbal contact. The device belongs to the Ethnomusicology department of the University of Los Angeles and is used for transcribing single-voiced ethnic music, usually human voice. The device was built by Inter-Ocean systems of Santa Barbara.

To comment on the operation of the Melograph, let us quote from an article by M.R. Schroeder [1970]:

> The oldest approach [to pitch detection] simply isolates the fundamental frequency of the signal by means of a low-pass or band-pass filter and then determines the frequency or period of the fundamental by means of measuring the rate of or the distance between axis crossings. Unfortunately, in many speech signals the fundamental is weak or even absent (as in most telephone signals).

In general, we cannot rely on the presence of the fundamental, or on the hope that the fundamental will be stronger than the second harmonic.

SPEECH TECHNIQUES

The research in speech understanding has contributed a great deal of work in pitch detection and system estimation. Since any musical scribe must detect the pitch of the incoming waveform, much of this may be useful. Let us describe some of these techniques in detail:

FOURIER METHODS

Our old standards, the Fourier transform and autocorrelation, were among the first to be tried [Harris and Weiss 1963]. These techniques were useful but had certain problems. In either the spectrum or the autocorrelation, there is a peak in the output at every multiple of the fundamental frequency (for autocorrelation, there is a peak at each multiple of the fundamental period). One could not just take the lowest peak because it is sometimes not there. Harris and Weiss developed a method of looking at several peaks in a row and forming an estimate of the fundamental frequency by averaging the contributions from the two strongest adjacent peaks. Rife and Vincent [1970], although not working directly with the pitch detection problem, developed a method of interpolating to get the position of the peak quite accurately by using weighting functions which had known effects on the transforms.

THE CEPSTRUM

With the advent of the cepstrum, probably first used by Bogert working on a suggestion by Tukey [Bogert, Healy, and Tukey, 1963], a new tool for speech research was opened up. Noll's classic article [1967] gave detailed instructions on the use of the cepstrum for the detection of fundamental frequency. This system had the advantage that the maximum of the cepstrum was often unique. When there was another peak, it was generally at twice the period of the fundamental, and rarely did it exceed the strength of the peak representing the fundamental. The cepstrum consists of the inverse Fourier transform of the log-magnitude Fourier transform of the input waveform. Since the autocorrelation is the inverse Fourier transform of the magnitude Fourier transform of the input waveform, the two processes are related. They both have time as the independent variable; they plot period rather than frequency. The theoretical basis of the method was developed in great detail by Oppenheim [1968, 1969], and Schafer [1969]. Roughly, the way it works in speech analysis is as follows: the speech waveform is taken to be the result of an excitation function (the glottal pulse) and a realizable filter (the vocal tract). It then follows that the log-magnitude Fourier transform of a segment of a speech waveform is the sum of the log-magnitude Fourier transforms of the glottal pulse waveform and the vocal tract impulse response. This being true, one can compute what the Fourier transform of this log-magnitude spectrum will be by superposition, since the signals add in the log-magnitude domain. Since the vocal tract is a filter, its frequency response is usually a broad, smooth curve with a small number of peaks (formants). The glottal pulse, however, is a nearly-periodic waveform which consequently has many harmonics. Its transform has a peak at the frequency of every harmonic. The transform is roughly periodic with a period equal to the fundamental frequency of the signal. If we take the transform of this quasi-periodic log-

magnitude spectrum, we would then expect to get a strong peak at the period representing the repetition rate in the frequency (or time) domain. When we take the transform of the log-magnitude frequency response of the vocal tract, however, we would expect to get something concentrated around the short periods, since the frequency response of the vocal tract is broad and slowly varying. This is, in fact, generally the case. The peak due to the periodicity of the glottal pulse tends to stand out from the activity due to the vocal tract. In fact, this separation of repetition from system response (excitation from filtering) was the basis of several ingenious techniques for removing echos [Schafer 1969] and for estimating the impulse response of the vocal tract. This estimation led to the development of the homomorphic vocoder [Oppenheim 1969, Miller 1973], where the cepstrum was used to determine the pitch of the speech signal as well as the impulse response. The signal could then be synthesized by convolving the derived impulse response with an impulse train at the original pitch. The impulse response was determined by eliminating the peak from the cepstrum and then inverting the process to yield a time series which was, in fact, an estimate of the impulse response of the filter. The peak was eliminated by simply setting the cepstrum to zero from the peak on, leaving only the short-time values of the cepstrum. Miller [1973] made extensive use of this technique to extract singing voice from orchestral background. Since the cepstrum just picked up whatever was loudest, there was quite a bit of error in the analysis which was subsequently corrected by hand. The cepstrum would just as happily track an orchestral instrument as the voice, if it happened to be dominant at the time. The result was synthesized with good results. The singing was highly intelligible and preserved well the character of the singer. One innovation in the synthesis is worth noting. Since the analysis is somewhat noisy, the impulse response estimate tended to vary from one estimate to the next. This produced some undesirable variation in the synthesis which sounded like roughness in the tone. This was eliminated by repeating each impulse response not just once, but five times with amplitudes which built up to a maximum and then f. This had the result of interpolating smoothly between one impulse response and the next and thus eliminated any roughness in the sound. Schafer's thesis gives an excellent review of homomorphic filter techniques.

THE LINEAR PREDICTOR

Another technique of system estimation which has been shown useful in pitch detection is the linear predictor [Itakura and Saito 1968, 1970, 1971; Markel 1972; Makhoul and Wolf 1972; Makhoul 1975; Boll 1973]. The idea here is to again model the signal as an excitation function, and a filter. We use the discrete analog of the Wiener-Hopf integral [Wiener 1947; Levinson 1947; Robinson 1967; Lee 1960] to estimate a non-recursive digital filter that approximates a filter which corresponds to the inverse of the filter that produced the sound. In other words, the filter we calculate has an anti-resonance everywhere the vocal tract has a resonance. If we filter the speech waveform with this filter that we have computed, the output will approach an impulse train. The better the estimation of the filter, the closer to an impulse train the output will be. This is because this filter, called an "inverse filter," tends to make the amplitudes of the

harmonics equal. Since the periodic signal with harmonics that all have the same amplitude is a pulse train, the output of the filter approaches the ideal pulse train. Pitch is then detected by calculating the distance between successive peaks of the inverse filtered speech waveform. Pitch can also be computed by taking the autocorrelation of the inverse filtered speech waveform. The largest peak in the autocorrelation is taken to represent the fundamental period. The theory behind this is that the reason the autocorrelation is not useful when directly applied to the speech waveform is widening of the autocorrelation peaks by the effect of the vocal tract. If the effect of the vocal tract is suppressed by filtering the waveform with the inverse filter, the peaks in the autocorrelation will be sharpened considerably. Since the speech waveform is constantly changing, the filter must be recomputed periodically. It is often done every 5 or 10 milliseconds.

The linear predictor can also be used, like the cepstrum, as a vocoder. Since the filter calculated by the predictor is an approximation to a filter whose inverse behaves like the vocal tract, the speech waveform can be synthesized by simply filtering a pulse train by the inverse of the filter produced by the predictor. Inverting the spectrum of a digital filter is a simple operation. Atal and Hanauer [1971] and later Markel and Gray [1974] programmed vocoders based on this principle and found them quite successful. A marvelous synthesis of the cepstrum and the linear predictor was done by Tribolet [1974], who joined the two methods to get an estimate of both the poles and the zeros of the filter. The linear predictor by itself is an all-pole model and is sometimes inadequate in the presence of a strong nasal zero. These topics are part of the larger field of system estimation. In this discipline, the object is to estimate the filter that could have produced the input signal in as much detail as possible with as little error and computation time as possible. Tribolet's thesis gives an excellent review of system estimation techniques. An excellent review and detailed analysis of the linear predictor is given by Makhoul and Wolf [1972]. Boll has also made significant contributions to the reduction of the compute time for the linear predictor [1973] by assuming that the filter which represents the vocal tract changes slowly with time. The estimate at this point in time can then be used to aid the computation of the estimate at the next point in time.

MISCELLANEOUS METHODS

Another method of pitch extraction that is also based on spectral flattening (making all the harmonics more alike in amplitude) was given by Sondhi [1968]. In his system, a band of bandpass filters are used to determine the spectral envelope. The speech waveform is then accentuated in frequencies where it is weakest. The resulting waveform has much more prominent peaks which can then be used to determine the fundamental frequency, either directly by measuring the distance between peaks, or by taking the largest peak in the autocorrelation. Sondhi also noted that the peaks in the autocorrelation can be enhanced by center clipping. This process uses an adaptive threshold to gate the signal through only when its magnitude exceeds the threshold. When the signal is passed, the threshold is subtracted

(added if the signal is negative) to prevent discontinuities in the waveform. The threshold is set to a fraction (such as .7) of the maximum amplitude in a given window. The center clipped waveform is then autocorrelated, and the strongest peak in the autocorrelation is taken to be the pitch period.

DIRECT WAVEFORM ANALYSIS

A series of pitch detectors have been devised which base their estimates directly on the speech waveform itself [Reddy 1966; Vicens 1969; Gold 1962; Gold and Rabiner 1969; Miller 1975]. Reddy used a three-step process based on measuring the significant maxima and minima of the speech waveform. The first step just detected the times when the speech waveform exceeded a certain fraction of the maximum of the waveform in a certain region. The second step determined the significant maxima and minima of the waveform, looking for places where a maximum and a minimum occur together. These two methods were related by three heuristic algorithms which matched the two pitch estimates, eliminated irregularities and filled "holes" in the pitch estimates. Gold and Rabiner made six measurements on the speech waveform, producing six different pitch period indications. A final stage of processing coordinated these six estimates to produce the final estimate. Two refinements were offered to improve the performance. Miller developed a technique which detects the "principal excursion" of the speech waveform for each period. This excursion is the large positive pulse which occurs after the glottal pulse. It is essentially the impulse response of the vocal tract. In most phonemes except nasals, this pulse is quite prominent. His method consists of integrating the waveform to locate the position of maximum positive area. The zero crossing preceeding this position is taken to be the beginning of the principal excursion. A series of heuristics is used to prune spurious and irregular zero crossings from the estimate.

All of the previous methods are based on the fact that the speech waveform is unique in many respects. It is this special behavior of the speech waveform that makes measurements on the waveform itself useful. These methods are somewhat sensitive to phase distortion. Miller's method, for instance, can be fooled by passing the speech waveform through an all-pass filter, which causes phase distortion that can eliminate the prominent peak in the signal. Excessive room reverberation, such as found in large concert halls, can also spoil the method, since reverberation causes great phase distortion. The method of Gold and Rabiner used a Lerner filter for bandpass filtering to preserve the phase relations as much as possible.

MUSIC PERCEPTION
or: A Child's Garden of Psychoacoustics

PITCH PERCEPTION

In trying to determine a method for analyzing musical sound, it would seem reasonable to look at what is known about how the ear does it, since we are trying to rival the ear's performance. As it turns out, many interesting observations have been made, but they raise many more questions than they answer. Let us review the existing literature in one particular area, the perception of the pitch of one voice. It seems impossible to cover all the interesting work in this area. We shall not attempt to do so here.

Our ear is presented with a musical tone. We perceive it as being at some pitch. What features of the waveform determine that pitch? What starts out sounding like such a simple problem turns out to be very complex.

In our naivite, we might first postulate something like Ohm's acoustical law [Ohm 1843]. Ohm suggested applying Fourier's theorem, such that each tone of a different pitch in a complex sound originates from the objective existence of a peak at that particular frequency in the Fourier analysis of the acoustic waveform. This would imply that the impression of pitch depends not only on the existence of a sinusoid at the fundamental frequency, but also that that sinusoid is of a stronger amplitude than any harmonics the tone may exhibit. Seebeck [1843] countered the theory of Ohm by determining the Fourier spectra of several of his previous observations [1841] and showing that in several cases, the sinusoid at the fundamental frequency was quite weak or even missing. A pitch at the hypothetical fundamental frequency was still perceived. Ohm [1844] and later Helmholtz [1863] declared Seebeck's observations to be invalid and the result of either illusion or faulty experimental technique.

We skip a half a century and pick up again with the work of Von Békésy [1928], who produced proof that the ear does a spectral analysis of some sort, where different frequencies excite responses from neurons originating in different places along the basilar membrane. As we progress along the membrane, the excitory frequency changes smoothly in a vaguely logarithmic manner.

With the coming of electronics, increasing evidence was gathered for the *case of the missing fundamental*, that indeed, a pitch could be perceived without the existence of any fundamental frequency at all. In fact, a group of higher harmonics can be heard collectively as a single, unified, percept. This percept is called the residue.

In an attempt to explain the phenomenon of the residue, one might observe that several adjacent harmonics added together produce a waveform which has a periodic modulation at the frequency corresponding to the difference of the harmonics. One might then hypothesize that either the ear detects the envelope of the incoming waveform, thus demodulating the signal

and extracting the frequency of the undulation, or perhaps the ear perceives the differences between the harmonics directly and infers the pitch from that. Figure 1 shows the waveform of a signal that has no fundamental frequency. It was produced by bandpass filtering a signal which has many harmonics. Notice the regular undulation that might imply some fundamental periodicity. Figure 2 shows the discrete Fourier transform of the waveform in figure 1, showing that it, indeed, has no fundamental. It also shows that the frequency of the undulation is roughly equal to the spacing of the harmonics in the Fourier transform. This undulation is a characteristic of a cluster of isolated harmonics.

Schouten [1940] in one experiment showed that neither of these could be the case. This was done by shifting the set of harmonics collectively by some amount. This makes the sinusoids no longer harmonically related, but it preserves the constant differences among them. In fact, one does perceive a change in the pitch of the residue even though the envelope of the waveform has not changed, nor has the differences of the frequencies of the sinusoids.

So. It is not the envelope, nor is it the differences among the harmonics. Well, what is it? De Boer [1956] did some revealing experiments which began the current trend in thinking on this question. If one takes a sinusoid of some frequency f, say 2000 Hz, and amplitude modulates it with some other frequency g, say 200 Hz, one gets three sinusoids of frequencies f−g, f, and f+g. As usual, these are heard as one percept of pitch g. A change in the carrier frequency, f, results in a proportional shift in perceived pitch. A more remarkable observation was that the pitch shifted downward when the modulating frequency, g, was raised! This effect was met with doubt up to incredulity. De Boer made the observation that these phenomena could be explained by hypothesizing that the ear detected the time difference between peaks of comparable amplitude. This is called the fine structure hypothesis, that the ear detects the details of the fine structure of the waveform and uses that data as the basis for pitch. Figure 3 shows the essence of this theory. We see a waveform which has a regular undulation. We have chosen an ambiguous case, where there are two separate maxima of equal amplitude, such that the time between the maximum of the previous undulation and this undulation can have one of two values. This theory predicts that the pitch will be ambiguous in this case.

Ritsma [1970] extended this theory a bit by showing that if pitch information is available along a large part of the basilar membrane at once (that is, if a tone has many harmonics), then the ear uses only the information from a narrow band. This band is positioned at about 3 to 5 times the pitch value. This is called the concept of dominance. Ritsma sums up the theory as follows:

> The sound is subjected to a spectral analysis on the basilar membrane. Because of the limited resolving power of the membrane, on each place of the membrane, a waveform is generated. According to the concept of dominance, only one region on the basilar membrane is dominant with respect to the perception of pitch. This region is roughly 4 times the pitch value. On the waveform generated in this dominant region, the ear performs an autocorrelation-like process determining the time interval between two pronounced positive peaks in the fine structure.

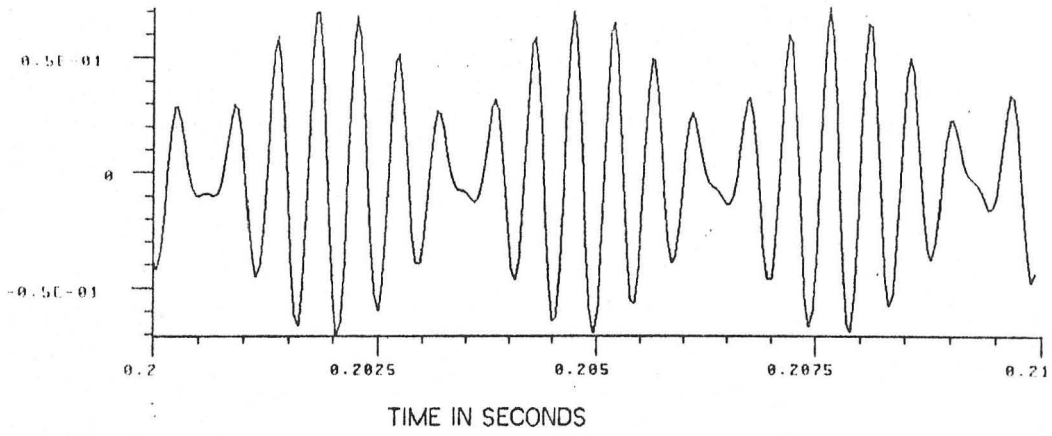


FIGURE 1. This waveform was produced by filtering the waveform of a guitar tone so as to select only a few of the upper harmonics. The note that was being played was roughly an E4 (332 Hz). The sixth and seventh harmonics were most prominant in this waveform, although many others are present to a lesser extent. It is clear that the waveform is periodic with a period of roughly 3 milliseconds, which corresponds to the frequency of the note.

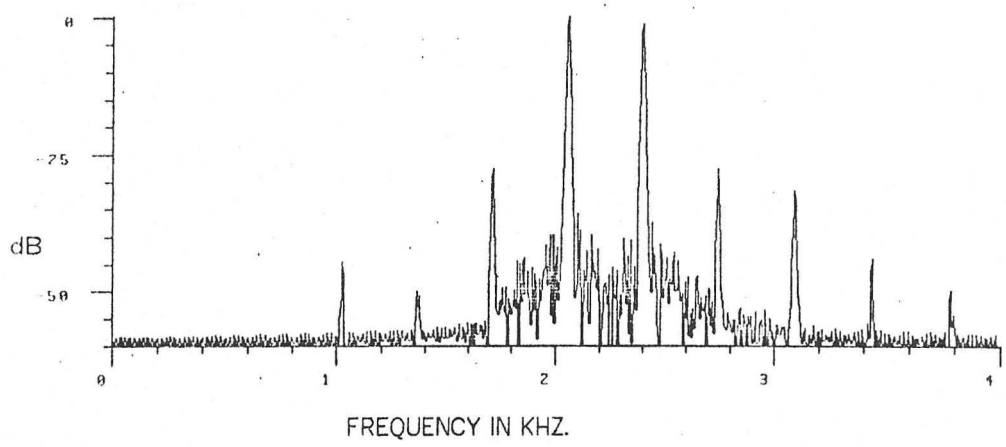


FIGURE 2. This is the discrete Fourier transform of the waveform in figure 1. As we can see, the first and second harmonics are entirely absent. Despite their absence, the waveform in figure 1 is quite periodic.

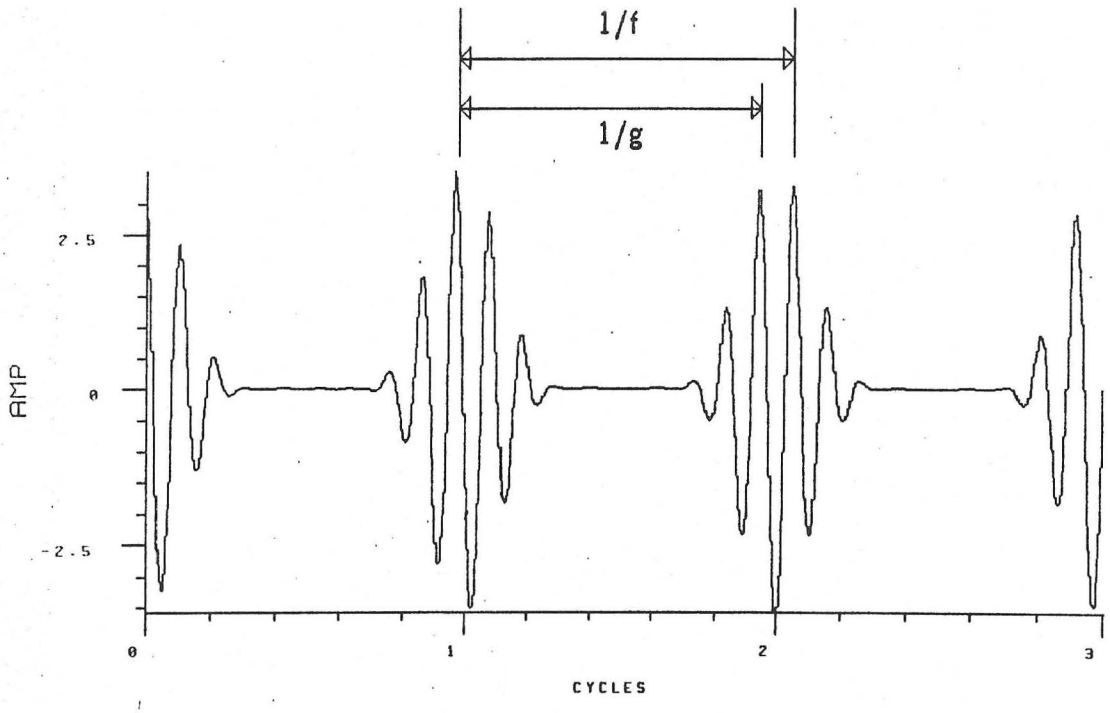


FIGURE 3. This illustrates one theory of pitch detection which is sometimes called the "fine-structure hypothesis". This theory states that the pitch is determined by measuring the time between the peaks in successive wave groups. In the case pictured above, the theory predicts a perceptual ambiguity in pitch, that some subjects would report f Hz. and some subjects would report g Hz as the pitch of this tone. This tone is inharmonic. As was pointed out by Wightman [1973], this theory is highly suspect because it depends on the phasing of the component sinusoids, whereas pitch perception does not seem to. The effect of phase change can be demonstrated simply by inverting the waveform. If we measure the distance between the negative peaks rather than the positive peaks, there is no longer any ambiguity in the pitch measurement.

This is what is called the *place versus period* controversy. The *place* advocates, of which Helmholtz and Ohm were members, attribute the perception of pitch to the position of maximum stimulation on the basilar membrane. The basilar membrane is known [Békésy 1934] to be frequency sensitive, with the frequency distributed motonically along the length of the membrane. The *period* advocates use the existence of the residue to show that there doesn't have to be any maximum at the place where pitch is perceived.

There is, again, evidence that the fine structure process is not the whole story. Smoorenburg did experiments with the perceived pitch of complexes consisting of two pure sinusoids. The problem is that given two tones at frequencies f_1 and f_2 ($f_1<f_2$), one not only hears the difference tone f_2-f_1, but one hears the combination tone $2f_1-f_2$, and it is louder than the difference tone. This effect can not be explained by any of the methods discussed so far. Hmmm! One explanation might be that there are nonlinearities in the ear that produce cross-frequencies. The problem is that although one can hear tones at frequencies $(n+1)f_1-nf_2$, one does not hear the corresponding higher tones at $(n+1)f_2-nf_1$. One can only wriggle out of this one by declaring that the nonlinearity must be frequency-selective, that it suppresses the higher sideband itself. Further work places more and more restrictions on the nonlinearity, such that it can only be considered as tentative, and the existence of the combination tones has yet to be explained satisfactorily.

Terhardt [1970] advanced De Boer's (and others') work and found small deviations in the pitch of the residue from what would be predicted by the fine-structure hypothesis. His conclusions imply that the ear itself transduces primary sensory data on the level of frequencies and amplitudes of the partials of a tone, and some higher level of processing is responsible for many of the funny effects, like the residue.

This was all fine and good until Wightman [1973, 1974] came along and showed that a change in the relative phases of the harmonics of a tone changes the fine structure drastically, but does not alter the perceived pitch. This essentially eliminates the fine-structure hypothesis. This can be seen in figure 3 by merely inverting the picture. This changes the fine structure entirely. For instance, there is no longer an ambiguity in the distance between maxima.

There are any number of other effects which should be mentioned just to give one an idea of the complexity of the issue. One marvelous effect is that of repetition pitch. If one takes a signal (like white noise) and delays it by some amount (say, 10 ms) and adds it back into itself, a listener generally perceives a pitch at the frequency represented by the delay. If the original signal is passed through a bandpass filter, and its delayed repetition passed through another bandpass filter whose passband does not overlap that of the first filter, the sum of the two filtered waveforms does not produce any pitch effect [Bilsen 1970]. The point here is that this effect could not be due to comparing successive peaks in the waveform for repetition because there are not necessarily meaningful repeating peaks. This argues for a more gross, averaging sort of process, like autocorrelation. There is a dichotic repetition pitch also. The original can

be played into one ear and the delayed sound can be played into the other, thus producing a pitch. This could only be produced at the first place where the signals from different ears meet at the same place, where they can be compared. The first place this is done is in the cortex itself.

Another effect reported in the literature is that of the binaural residue [Houtsma and Goldstein 1972]. In this experiment, two higher harmonics are used to produce a perceived pitch at the frequency of the missing fundamental. The difference is that one harmonic is played into one ear and the other harmonic is played in the other ear. At low sound pressure levels, one indeed does get a residue phenomenon. Like the dichotic repetition pitch, this implies that some aspects of pitch formation are done at a high level of processing. Our informal listening tests have failed to confirm this effect.

Siebert [1970] calculated entirely from statistical arguments that human perception of pure sine tones was based on place rather than periodicity. His calculations show that not only would the frequency resolution be much more acute, but the form of the behavior as a function of the frequency of the tone would be different if time cues were used. It would, for one thing, be dependent upon the amplitude of the tone. Except in the limit (very loud or very soft), the resolution is independent of amplitude. Three more recent theories (Wightman, Goldstein [1973], Terhardt) go on to propose modified *place* theories. In these theses, the *place* of stimulation is transmitted to the brain, where some higher-level process pieces together the evidence and registers a pitch. Terhardt even shows a *learning* model which must undergo a training sequence to acquire effects like the residue. In none of these theories is the fundamental necessary for pitch perception. It is inferred from a sequence of harmonics. Both Goldstein and Terhardt present models that are essentially statistical in nature, leaning heavily toward decision-theoretic methodology. Wightman is still using a modified autocorrelation approach with reasonable results so far. None of the models is comprehensive enough to explain all the effects of pitch perception that have been noted, but they all show promise of being extendable. If implemented on the computer, Terhardt's model would require more than 10^6 words of memory just for the decision table.

In any case, it would appear that the current concensus is that the ear resolves separately each of the harmonics of a complex tone. The existence of and pitch of these harmonics is sent to the brain. The brain then examines them (and the immediate past, presumably) and decides what pitches are present. The theory to date is not detailed enough to directly code for the computer, but it is somewhat suggestive of promising directions for research.

It is not clear what the residue and combination tones have to do with music perception. Most music is polyphonic, which already implies that weak effects like residue and combination tones are of secondary importance.

There is a great deal more literature in psychoacoustics that deal with topics that are related to

music to one degree or another that will not be reviewed here. These include works on consonance and dissonance, timbre, cognitive (high-level) processing, and many others.

LOW-LEVEL TECHNIQUES

INTRODUCTION

The low-level techniques are those which operate directly on the digitized waveform. They belong largely to the realm of digital signal processing. The purpose of these techniques, in our application, is to determine what frequencies are present in the input waveform, how strong they are, and over what intervals in time they exist. This is, of course, a statement of the variables in our model of musical sound. We wish to determine how many sinusoids are present at any given time as well as what the slowly-varying amplitude and frequency functions are, as functions of time. Since we are not interested, for the moment, in identification of the instruments, nor are we interested here in synthesis of music instrument tones (synthesis will, however, be discussed briefly in the following sections), we do not need to determine these functions to great accuracy.

The routines group themselves into two broad categories: pitch detectors and harmonic extractors. The pitch detectors (more precisely, *periodicity* detectors) take a signal in and produce as output a list of what frequencies are present in the signal as a function of time. Pitch detectors work best when the signal is a single periodic waveform, but have some application in polyphonic sound. Although any number of techniques have been used as pitch detectors in the past [Gold 1962; Gold and Rabiner 1969; Moorer 1974; Miller 1975; Harris and Weiss 1963; Markel 1972; Noll 1967; Sondhi 1968; Reddy 1966], we will only deal with two autocorrelation-like methods: the optimum-comb method and the autocorrelation function. The reason is that these methods are more useful in the polyphonic case than any other common methods. The methods that use direct waveform measurement [Reddy 1966; Gold 1962; Miller 1975] are biased toward monophonic human speech. The spectral flattening methods [Markel 1972; Sondhi 1968] are based entirely on the assumption of monophony and have no application in polyphony. The spectral methods [Harris and Weiss 1963; Noll 1967] have various problems and will be discussed individually later.

The purpose of a harmonic extractor is to produce the waveform, or at least a model of the waveform, as a function of time, with all other simultaneous activity eliminated. We will discuss two such extractors: the heterodyne filter and bandpass filtering. The heterodyne filter is a *harmonic-based* technique, in that it requires that the input waveform be periodic. It then returns the amplitudes and phases of each of the harmonics as functions of time. Bandpass filtering has no such restriction, but has a problem with resolution of time-detail. There is a direct tradeoff between frequency resolution and time resolution with the bandpass filter. This is sort of the signal processing enthusiast's "Heisenberg principle" (or perhaps the signal processor's own personal albatross!).

And then there are all the methods that didn't work. These are, of course, far too numerous to detail in one lifetime, but three of the more important failures are discussed.

The techniques that were found useful are interesting in their own right, but they must be merged into a unified whole to accomplish anything. The last section of this chapter deals with the algorithms used to weave meaningful threads through the data.

METHODS FOUND TO BE USEFUL (AND WHY)

THE AUTOCORRELATION FUNCTION

INTRODUCTION

The autocorrelation function is one of the oldest and best understood signal-processing techniques. It is defined as follows:

$$(7)\quad A(\tau) = \int_{-\infty}^{\infty} F(t)F(t+\tau)\,dt$$

Where $F(t)$ is the input waveform at time t,
and τ is the lag time in seconds.

In the world of sampled-data, we do not have the function from the beginning of time to the end, nor do we have the function at all points. For sampled-data systems, there are several analogous functions we may use:

$$(8)\quad A_m = \sum_{n=-\infty}^{\infty} F_n F_{n+m}$$ *(discrete analog of (7))*

$$(9)\quad A_m = \sum_{n=0}^{N-m-1} F_n F_{n+m}$$ *("windowed" to N points)*

$$(10)\quad A_m = \sum_{n=0}^{N-1} F_n F_{(n+m) \bmod N}$$ *("cyclic" autocorrelation)*

$$(11)\quad A_m = \sum_{n=0}^{N-1} F_n F_{n+m}$$ *(covariance)*

Where F_n is the input waveform at the n^{th} sample, that is, at time nh
where h is the time between samples
and m is the lag index in samples, that is, the total lag time is mh

We shall use the definition of equation (11). To see what this does to a signal, let us calculate and observe its behavior on a pure sinusoid.

$$(12)\quad A_m = \sum_{n=0}^{N-1} B \sin(n\omega h+\phi)\; B \sin[(n+m)\omega h+\phi]$$

Where B is the amplitude of the sinusoid
ω is the radian frequency of the sinusoid
ϕ is the phase of the sinusoid

And by the magic of the summation calculus we get:

$$(13)\quad A_m = \frac{1}{2} B^2 \left\{ N \cos(m\omega h) - \frac{\sin(N\omega h)}{\sin(\omega h)} \cos[m\omega h+(N-1)\omega h+2\phi] \right\}$$

This is plotted in figure 4 for certain values of the parameters. By equation (13), we can see that A_m is periodic with period $\Delta m = 2\pi/\omega h$. It has maxima and minima that recur with that period. As a function of m, it is, in fact, a perfect sinusoid. This can be seen because it is the sum of two sinusoids of the same period ($2\pi/\omega h$) with differing but constant phases and amplitudes. The result is another sinusoid.

Since the autocorrelation is not linear, superposition does not apply. We cannot generalize by inspection. We can, however, compute the autocorrelation of a perfectly periodic waveform of arbitrary spectral content.

$$(14)\quad A_m = \sum_{n=0}^{N-1} \left[\sum_{j=1}^{L} B_j \sin(nj\omega h+\phi_j) \right] \left[\sum_{k=1}^{L} B_k \sin(nk\omega h+\phi_k) \right]$$

Where n is the harmonic number,
B_n is the amplitude of the n^{th} harmonic,
ω is the radian fundamental frequency of the waveform,
ϕ_n is the phase of the n^{th} harmonic.

Which comes out to the following:

$$(15)\quad A_m = \frac{1}{2} \sum_{j=1}^{L} \sum_{k=1}^{L} B_j B_k \left\{ \cos\left[mk\omega h+\phi_k-\phi_j+\frac{N-1}{2}\right] \frac{\sin\left[\frac{N(k-j)\omega h}{2}\right]}{\sin\left[\frac{(k-j)\omega h}{2}\right]} \right.$$

$$\left. - \cos\left[mk\omega h+\phi_k+\phi_j+\frac{N-1}{2}\right] \frac{\sin\left[\frac{N(k+j)\omega h}{2}\right]}{\sin\left[\frac{(k+j)\omega h}{2}\right]} \right\}$$

This expression is plotted in figure 5 for several values of the variables involved.

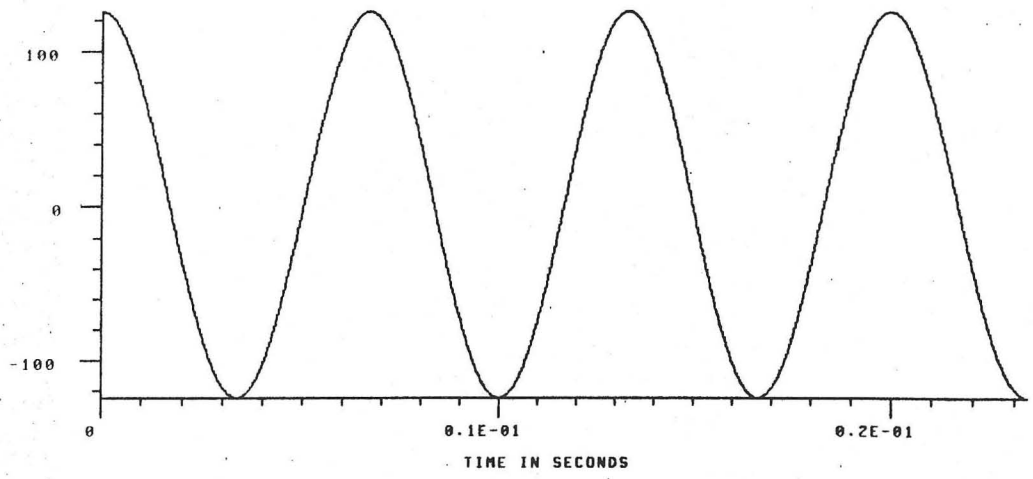


FIGURE 4. This is the autocorrelation of a pure sinusoid. The result is, as we would expect, a pure sinusoid with a maximum at integral multiples of the period.

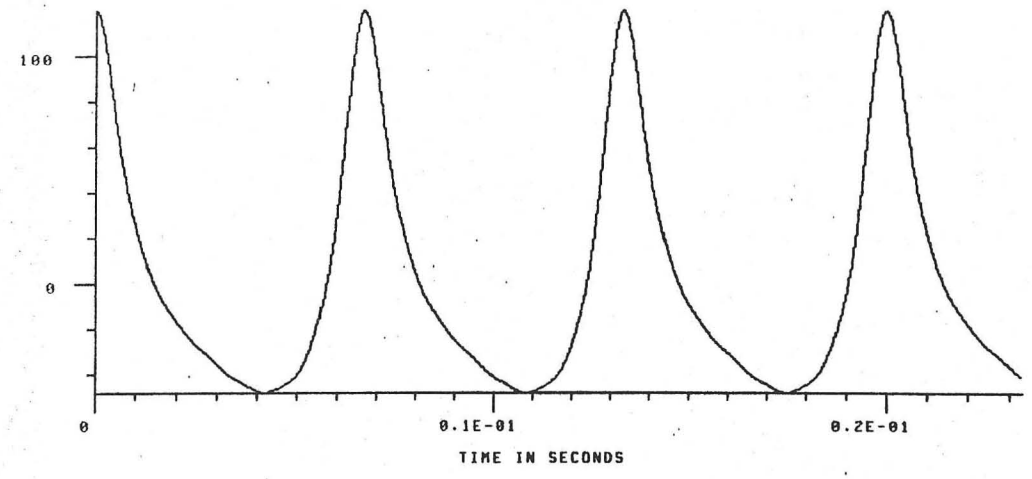


FIGURE 5. This is the autocorrelation of a periodic signal with 5 harmonics. As we see, the result is also periodic, although the harmonic amplitudes are entirely different from those of the input waveform.

Again, it is periodic in m with period $\Delta m = 2\pi/\omega h$. Again, the maxima and minima recur with that period. While this result is no longer a pure sinusoid, it is a harmonic series, and is thus periodic.

It is interesting also to observe the results when a waveform with missing harmonics is applied. Figure 6 shows the autocorrelation of a waveform with only three harmonics, numbers 5, 6, and 7. The autocorrelation is still periodic with a period equal to the period of the missing fundamental frequency. Figure 7 shows the autocorrelation of a waveform with harmonics 2, 3, 4, 6, 8, 9, and 10 present. This is what you might get if two notes were present at 300 Hz and 450 Hz, an interval of a perfect fifth.

Two instruments playing at perfect fifths will produce an autocorrelation with a period equal to that of a fictitious "fundamental" period.

With this theoretical base, let us see what this function does with actual music waveforms.

USAGE

We see in figure 8 the waveform of a trumpet playing an G4, roughly 392 Hz. This waveform and the next were taken from a recording of Ravel's orchestration of Mussorgsky's *Tableaux D'une Exposition*. This is the first note of the piece. We can easily see that the period is near 2.5 milliseconds. What small deviation exists is due to inaccuracies in the rotational speed of the turntable. In figure 9 we see equation (11) evaluated for 3.5 periods of the input waveform. We see that the output is periodic also with period of about 2.5 milliseconds.

In figure 10 the waveform of the first brass chord of the piece. This is a G-minor triad. The note, G, corresponds to a frequency of about 98 Hertz, which is slightly over 10 milliseconds in period. The evaluation of equation (11) for this waveform is shown in figure 11. The greatest maximum is clearly at about 10 milliseconds. This demonstrates the principle of determining the harmony of a piece of music without determining what notes are being played at any given time.

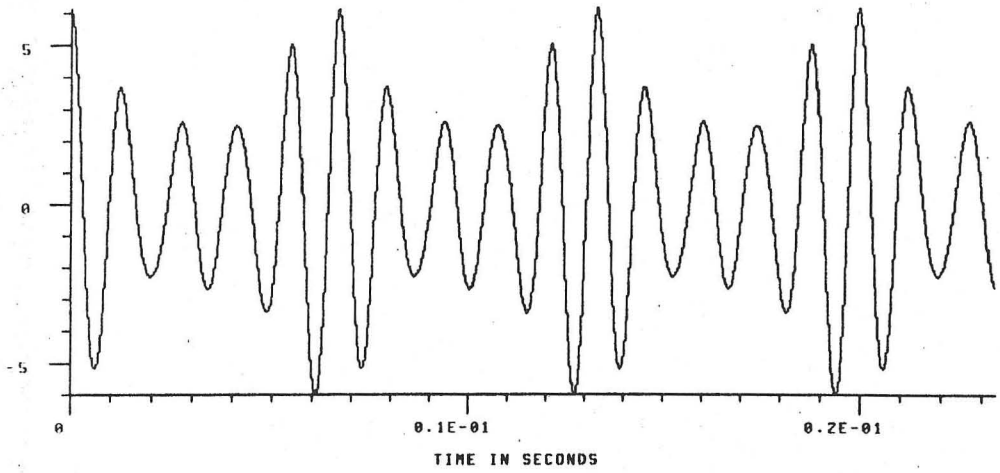


FIGURE 6. The autocorrelation of a periodic signal with only three harmonics: the 5th, 6th, and 7th. The autocorrelation is periodic with a period equal to the missing fundamental of the waveform.

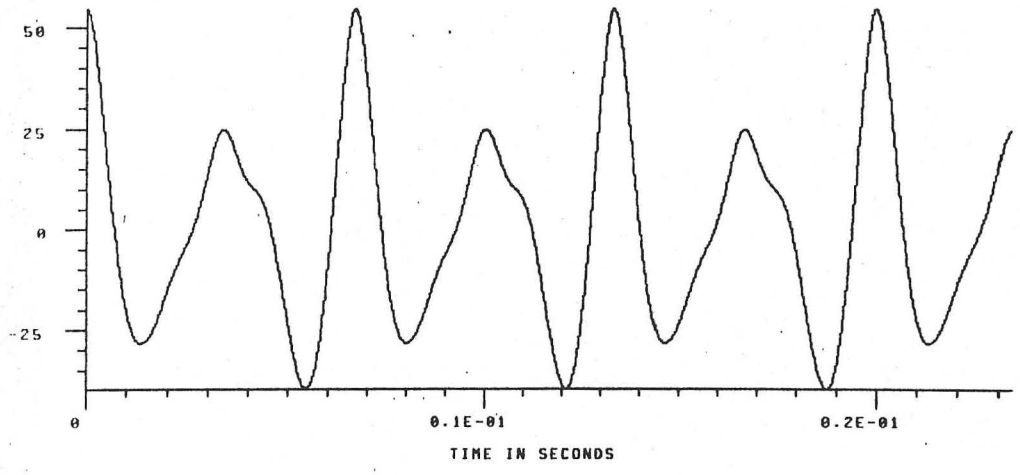


FIGURE 7. The autocorrelation of a periodic signal with only harmonics 2, 3, 4, 6, 8, 9, and 10 present. This is what would occur, for instance, if two tones at 300 Hz and 450 Hz were present simultaneously. This represents the musical interval of the perfect fifth. Any two tones at this interval will produce a periodicity in the autocorrelation equal to an implied fundamental period of the composit waveform.

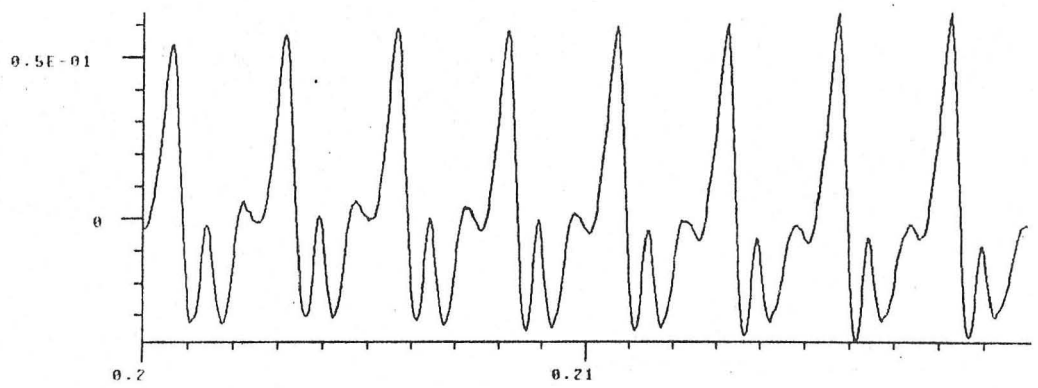


FIGURE 8. A segment of the waveform of a solo trumpet in a highly reverberant environment. This was taken from a recording of Tableau D'une Exposition.

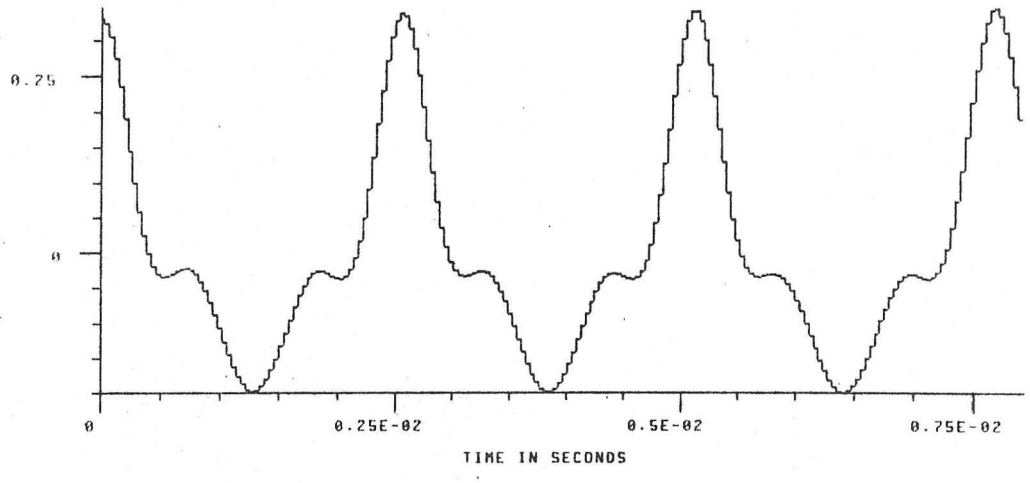


FIGURE 9. The autocorrelation of the waveform shown in figure 8. As we would expect, it is periodic with the same period as the input waveform.

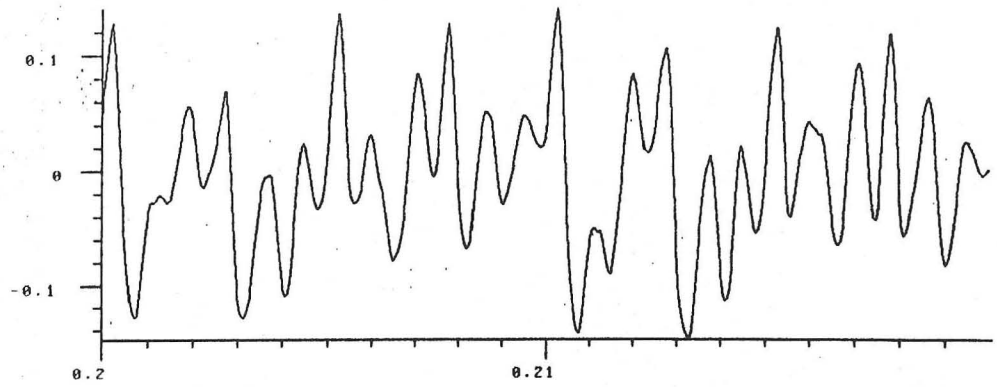


Figure 10. A segment from a recording of a brass choir. This is a root-position G-minor chord taken from a recording of Tableau D'une Exposition.

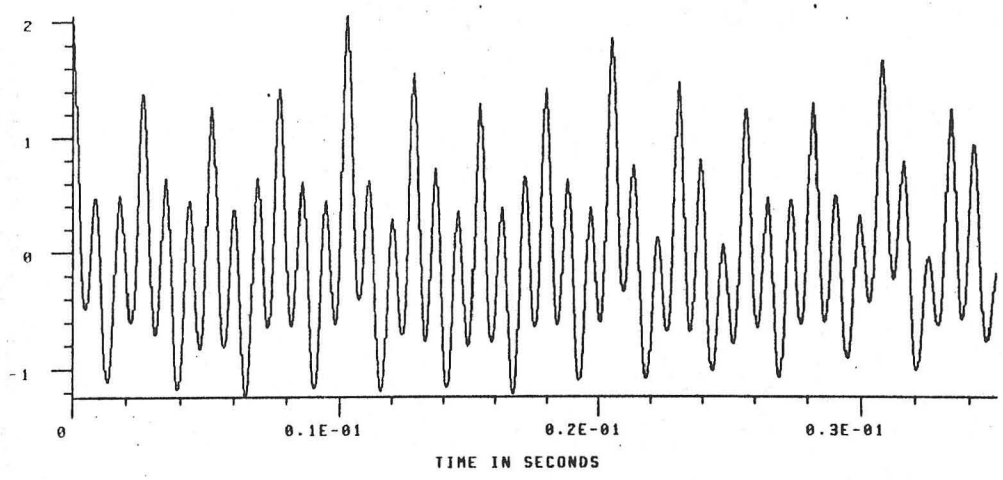


Figure 11. The autocorrelation of the waveform shown in figure 10. It has maxima at multiples of 98 Hz, representing the low G2 root note.

THE COMB FILTER

DEFINITION AND ANALYSIS

Another function that is closely related to the autocorrelation function is the magnitude of the output of a comb filter whose delay is swept over some range of interest. This was discussed by Moorer [1974] and by [Ross *et al* 1974].

A comb filter is defined by the following difference equation:

(16) $Y_n = X_n - X_{n-m}$

Where X_n is the n^{th} sample of the input waveform,
and Y_n is the n^{th} sample of the output waveform

There are, in fact three other things that are called comb filters. The first is produced by changing the subtraction to an addition. The other two are formed by delaying and differencing the output rather than the input. We will only discuss the form shown in equation (16).

It is easy to show that the magnitude-frequency response of the comb filter as defined above is

(17) $\{\sin^2(m\omega h) + [1-\cos(m\omega h)]^2\}^{1/2}$

This comb filter has a zero of transmission at frequencies which are integral multiples of $1/mh$ Hertz. Thus, if the input waveform is a stationary signal consisting of nothing but frequencies which are multiples of $1/mh$ Hertz, the steady-state output of the filter will be identically zero.

What we do is to sum the magnitude of the output of the filter for some number of points, say k points. The minima in this sum represents periodicities present in the input waveform. This sum may be written in the following manner:

(18) $\sum_{i=0}^{k-1} |X_{n+i} - X_{n+i-m}|$

This is related to the autocorrelation function as defined in equation (11). In fact, it is approximated by the following function [Ross *et al* 1974]:

(19) $(A_0 - A_m)^{1/2}$

Where A_m is defined by equation (11). This shows that where A_m has a maximum, equation (18) will show a minimum. Computationally, equation (18) is easier to compute than equation (11) because it involves only additions, no multiplication or division.

USE FOR DETERMINATION OF HARMONY

A program was written using the comb filter as the fundamental technique for the purpose of determining the harmony of a piece of music. Figure 12 shows a display of the results of this program when applied to the first brass choir in *Tableau D'une Exposition.* The graph shows time in milliseconds on the horizontal axis and frequency (actually, inverse period) on the vertical axis. The vertical axis is period in seconds, but it is labeled in frequency. This places the highest frequency (smallest period) nearest the origin and the lowest frequency (largest period) is at the top. The heavy squiggly roughly horizontal lines represent minima in the evaluation of equation (11). The equation was evaluated every 10 milliseconds throughout the excerpt. The minima in adjacent time slices which were extremely close in frequency were linked into lists. The beginning of each list is denoted on the figure by a vertical stroke. The long, light horizontal and vertical lines were placed there by hand as a guide to interpretation of the figure. The vertical lines denote the places where the chords change, as determined by hand (by ear?) by the author. The horizontal lines point out some selected frequencies. The names of the chords have been placed above the graph as a guide to interpreting the data. One attribute which is used by subsequent programs but is not shown here is the depth of the minimum. Many of the traces are weak and will be subsequently ignored.

One of the interesting features is that the first G minor triad produces a strong trace on the low G natural, but the second G minor triad produces a strong trace on the low Bb. This is because on the second G minor triad, the Bb is doubled in the trumpets, giving it much more strength. The score of the first few bars of the piece is shown in figure 13 for reference.

One thing to notice is how the traces often continue to run on after the chord has changed. This is because the recording was made in an extremely reverberant environment. The tones continued to ring long after the chord changed.

There are many other traces for each chord than just the root of the chord. These other traces are subharmonics of the notes in the chord. They are clear to see in figure 14 as all the other minima. One must remember that any periodic component of the waveform will produce some kind of minimum in equation (11). The minima get deep when the periods are rational multiples of one another. Then their subharmonics will coincide to produce a deep minimum.

To demonstrate both the power and the limitations of this method for determining harmony, 9 test chords were synthesized and processed. The first was a C-major triad in root position. The results are shown in figure 15. We see a strong minimum at slightly over 15 milliseconds, which is somewhat over 64 Hertz, which is about C2. This is as if the notes of the chord were the 4^{th}, 5^{th}, and 6^{th} harmonics of C2.

When we add an A4 to the chord, the chord becomes ambiguous. It is the superposition of a C triad and an A-minor triad. This chord is usually referred to as an A-minor *seventh* chord. A *major seventh* chord produces unambiguous deep minimum, because the major seventh chord represents the 4^{th}, 5^{th}, 6^{th}, and 7^{th} harmonics of the root (even though the 7^{th} harmonic is lower in frequency than is commonly used in the major seventh chord).

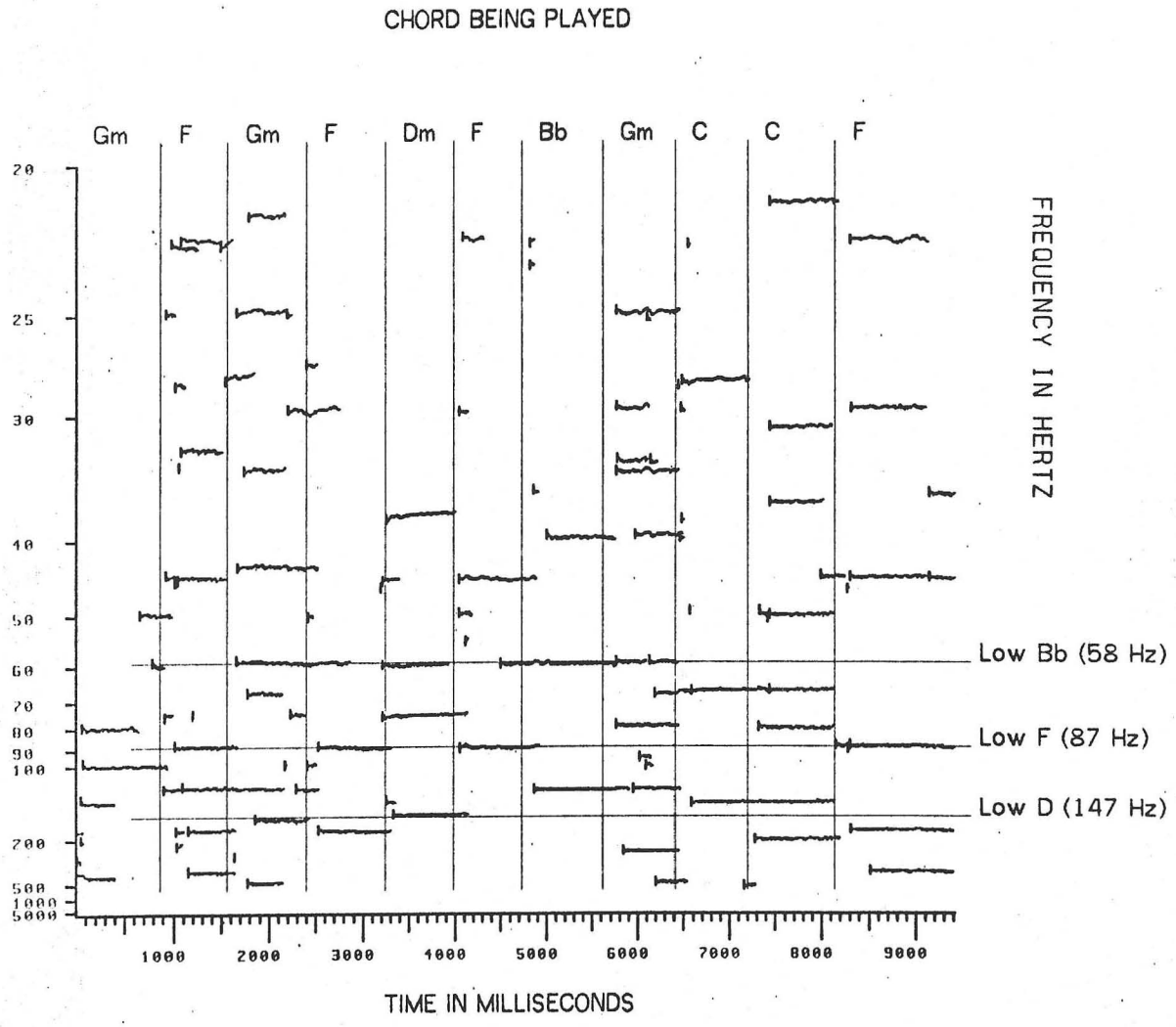


FIGURE 12. This shows the output of the optimum-comb pitch detector when applied to the first brass choir in Tableaux D'une Exposition. The minima in adjacent time slots have been linked together into lists. There is a vertical stroke at the beginning of each list. The horizontal axis is time in milliseconds. The vertical axis is period, but is labeled in frequency. This means that the labelings in frequency are not equally spaced and the highest frequency (smallest period) is at the origin. Naturally, the scale goes asymptotic at zero period (infinite frequency). To help in evaluating the results, light vertical bars have been placed at the places where the chords change. The chord names have been printed at the top of the figure. The light horizontal bars denote some important frequencies for comparison. The strongest traces seem to occur when notes are doubled in the orchestration. Compare this plot to figure 13 which shows the score of the first part of the piece.

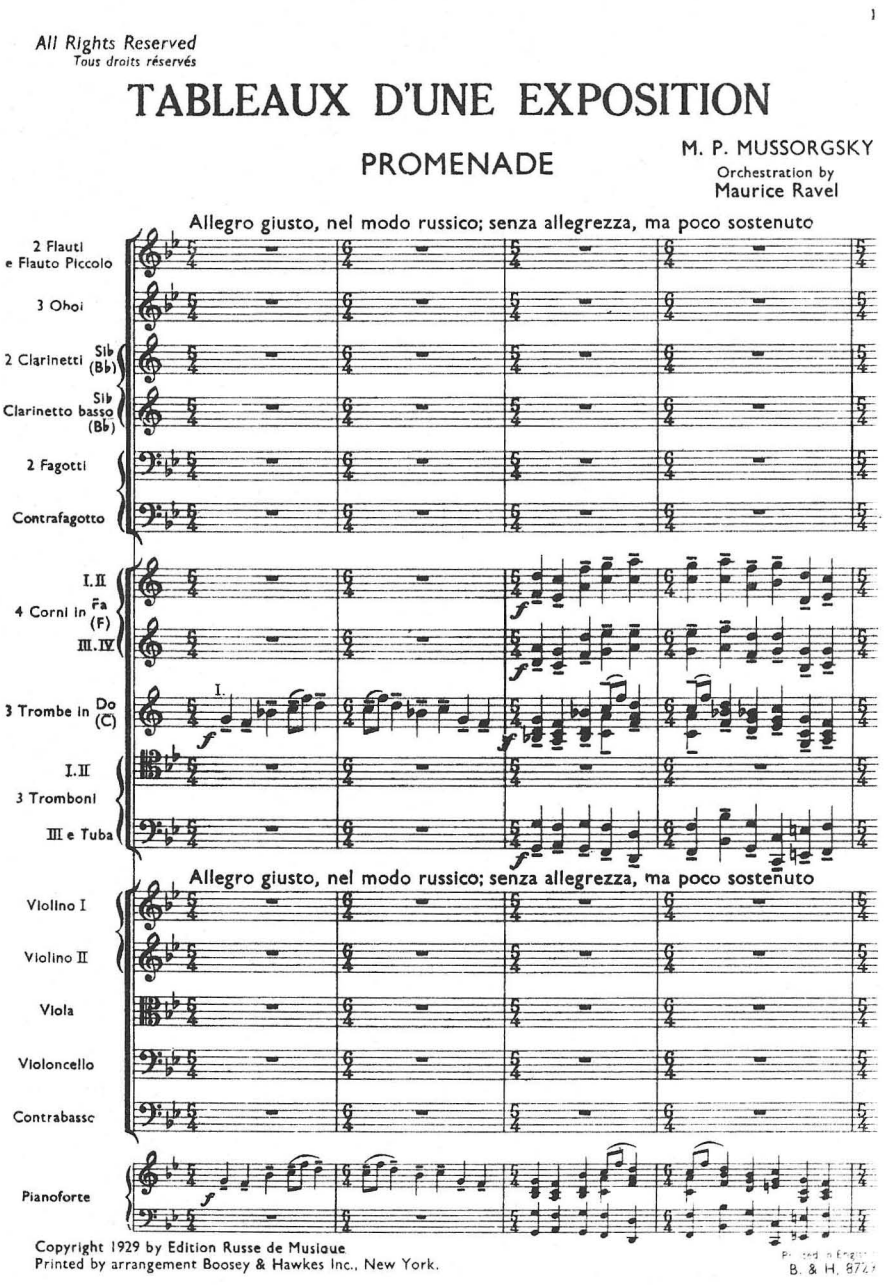


FIGURE 13. This is the first page of Ravel's orchestration of Mussorgsky's Tableaux D'une Exposition. The original piano score is shown at the bottom. This is from the Boosey & Hawkes pocket edition, 1929.

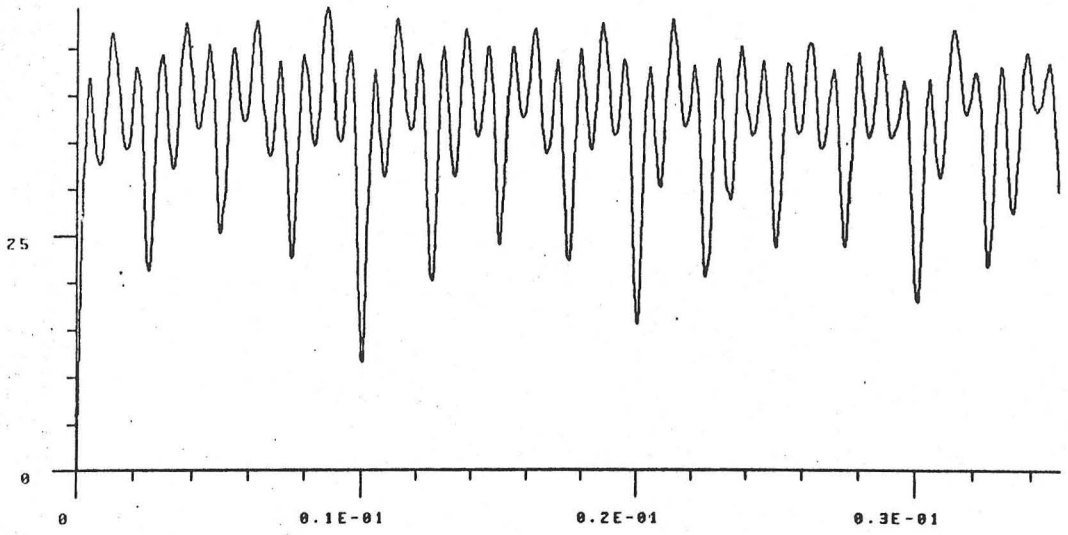
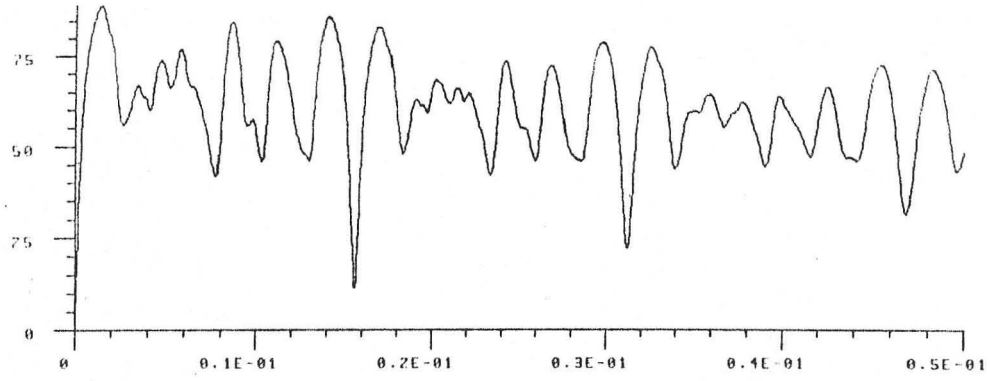
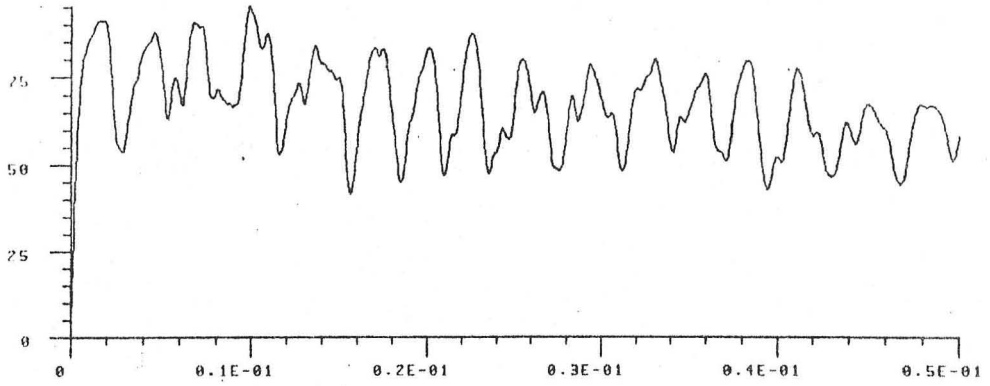


FIGURE 14. This is the results of applying the optimum-comb to the first chord of the brass choir in Tableaux D'une Exposition. The chord is a G-minor. The principal minima are subharmonics of G2 (about 98 Hz.).



TIME IN SECONDS

FIGURE 15. Equation (18) applied to C-major chord in root position. The notes in the chord are C4, E4, and G4. We see a distinct minimum at 15.5 milliseconds, which is C2.

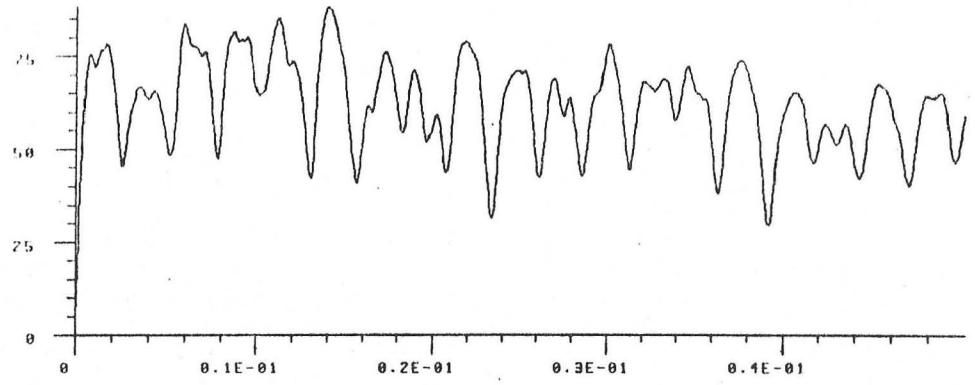


TIME IN SECONDS

FIGURE 16. Equation (18) applied to a C-major-sixth chord in root position. The notes in the chord are C4, E4, G4, and A4. Since this chord is ambiguous, no strong minimum occurs This chord is usually called an A-minor-seventh, in which case this chord is in the first inversion.

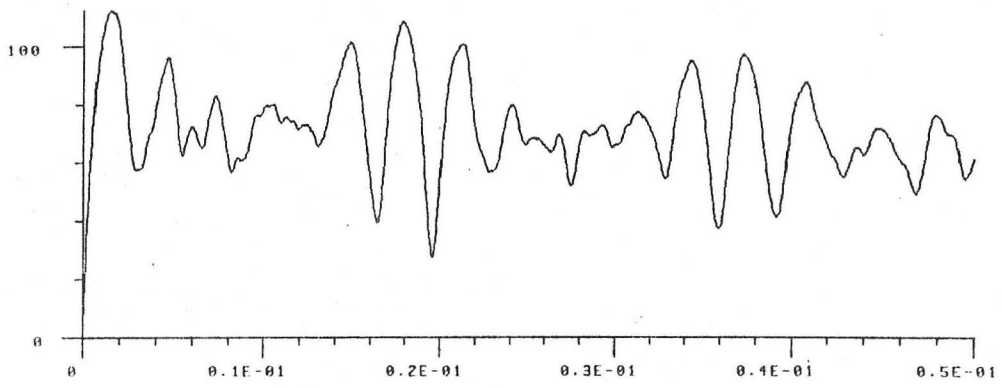
The minor seventh chord does not have such a clean correspondence to the harmonic series. The minima in the comb filter output for ambiguous chords are subharmonics of the notes of the chord. This is shown in figure 16. When we apply the formula to a C-minor triad, we get two strong minima. One is at F1, which makes the notes of the chord the 6^{th}, 7^{th}, and 9^{th} harmonics. The other is at Ab0, which makes the notes of the chord the 10^{th}, 12^{th}, and 15^{th} harmonics. This is shown in figure 17. In figure 18 we see the results from a C-diminished chord. The strong minimum is at Ab1, which makes the notes the 5^{th}, 6^{th}, and 7^{th} harmonics. In figure 19 we see the results from the famous diminished-seventh chord. This is one of the most ambiguous chords in common usage. As we might expect, there is no strong minimum. Figure 20 reports the results for a C-augmented chord. There is a minimum at F0, which makes the notes the 12^{th}, the 15^{th}, and the 19^{th} harmonics. Now we have 3 simpler examples. Figure 21 shows the results from a C-major-nineth chord, figure 22 is for a C-major triad in first inversion, and figure 23 if for a C-major triad in second inversion. These three all show strong minima at C2.

Thus we see that the comb filter can be used to detect and identify any unambiguous chord with reasonable accuracy.



TIME IN SECONDS

FIGURE 17. Equation (18) applied to C-minor chord in root position. The notes in the chord are C4, Eb4, and G4. There are two strong minima. One at slightly over 23 milliseconds, or 43 Hertz. 43 Hertz is F1. There is another minimum at 39 milliseconds, which is Ab0



TIME IN SECONDS

FIGURE 18. Equation (18) applied to a C-diminished chord in root position. The notes in the chord are C4, Eb4, and Gb4. The strong minimum is slightly over 19 milliseconds, or about 52 Hz, which is Ab1.

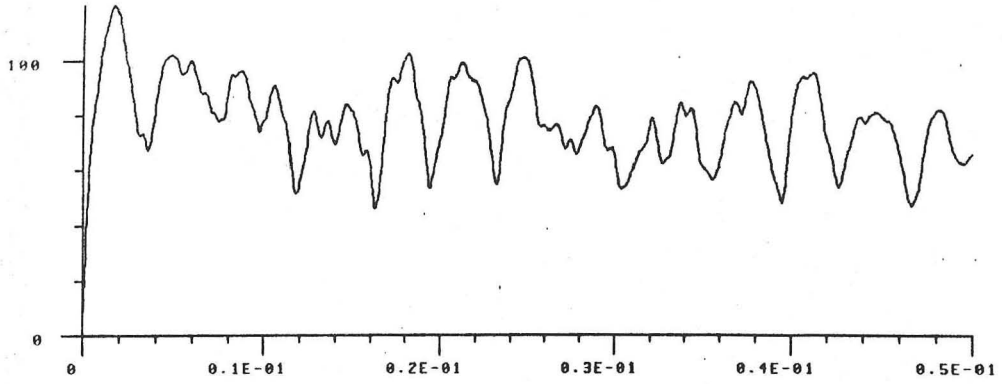


FIGURE 19. Equation (18) applied to C-dimished-seventh chord. The notes in the chord are C4, Eb4, Gb4, and A4. There are no strong minima because this chord is highly ambiguous.

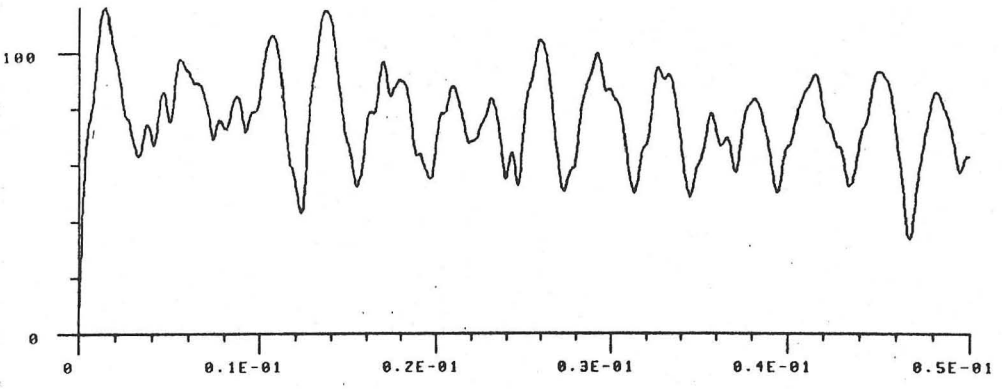
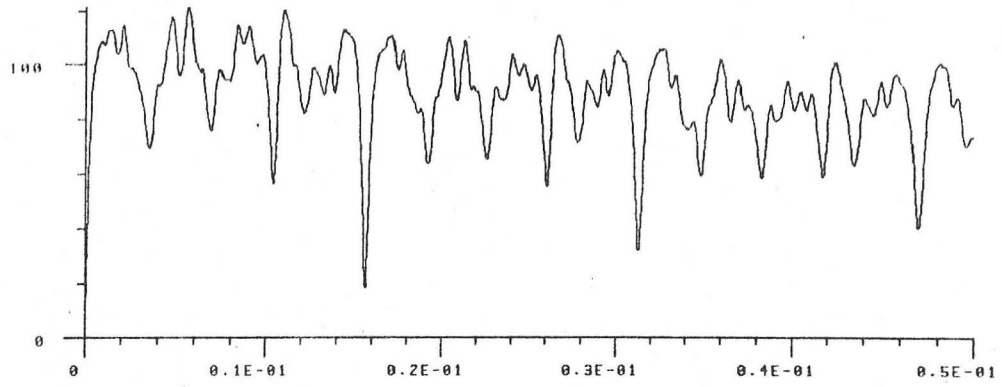
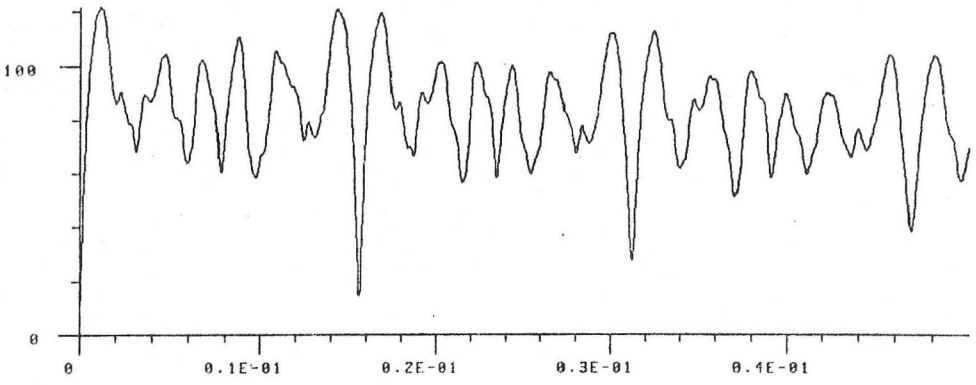


FIGURE 20. Equation (18) applied to a C-augmented chord in root position. The notes in the chord are C4, E4, and G#4. The strong minimum is slightly over 46 milliseconds, or about 22 Hz, which is F0.



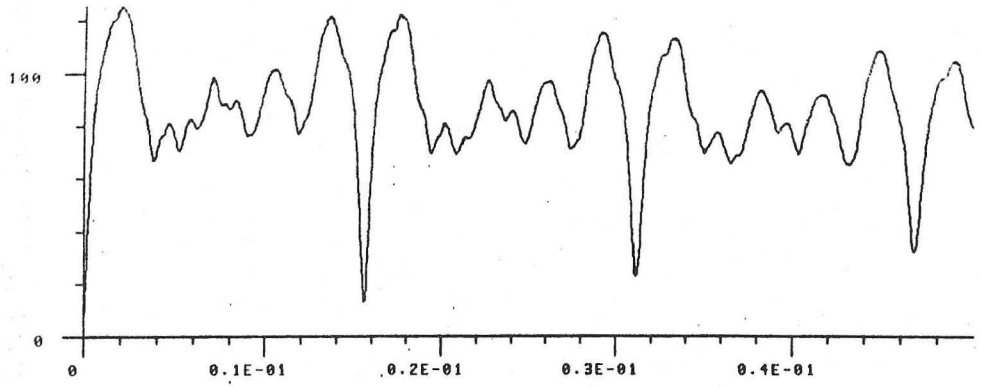
TIME IN SECONDS

FIGURE 21. Equation (18) applied to C-major-ninth chord in root position. The notes in the chord are C4, E4, G4, and D5. This chord, like the C-major chord, has a strong minimum at 15.5 milliseconds, or 64.5 Hertz, which is C2. The traditional definition of the ninth chord includes the seventh degree, which in this case would be B♭4. It is ommitted here to help separate the effects of the D5, although its inclusion would not greatly perturb the plot nor disturb the location of the minimum.



TIME IN SECONDS

FIGURE 22. Equation (18) applied to a C-major chord in the first inversion. The notes in the chord are E4, G4, and C5. The strong minimum is again at 15.5 milliseconds



TIME IN SECONDS

FIGURE 23. Equation (18) applied to C-major chord in second inversion. The notes in the chord are G3, C4, and E4. This chord, like the C-major chord, has a strong minimum at 15.5 milliseconds

THE HETERODYNE FILTER

INTRODUCTION

This tool is an adaptation of the discrete Fourier transform, hereafter abbreviated DFT. The heterodyne filter is used as a filter or operator. It takes a function of time as input and gives many functions of time as output. It is used to determine the amplitude and frequency functions which make up nearly-periodic waveforms. More directly, we represent such waveforms as follows:

$$(20)\quad F_\alpha = \sum_{n=1}^{M} A_{n\alpha} \sin(n\omega\alpha h + \theta_{n\alpha})$$

Where F_α is the signal at time αh,

h is the time between consecutive samples,

ω is the radian fundamental frequency of the note,

n is the harmonic number,

$A_{n\alpha}$ is the amplitude of harmonic n at time αh,

M is the summation interval in samples. For best results, this must be set to the number of samples in one period of F_α, or the closest integer to $2\pi/(h\omega)$.

$\theta_{n\alpha}$ is the phase of harmonic n at time αh.

This models the waveform as a sum of sinusoids with time-varying amplitudes and phases. We must insist that the amplitudes and phases vary slowly with time, or the analysis procedure does not give correct results.

This is not a Fourier series representation, although it looks similar. The Fourier series demands that the sinusoids be perfectly harmonic and of constant amplitude. If we allow the amplitudes or phases to vary, the sinusoids are no longer orthogonal by summation over one period, thus the sinusoids do not constitute a Fourier series. We mention this fact because this means that the tone can not be resynthesized by use of the fast Fourier transform algorithm. To resynthesize the tone from $A_{n\alpha}$, $\theta_{n\alpha}$, and ω, we must evaluate M sinusoids for every point in time.

The heterodyne filter has its main use in analysis for the purpose of insight into music instrument physics and for resynthesis of the instrument tone. It could be used for analysis of music that formed unambiguous chords at every point, that had no notes outside of the chord. This is the case with very little music, thus making the filter of little use to the musical scribe. One would be hard put to find any such music outside of harmony textbooks.

METHOD AND ANALYSIS

The method is defined as follows:

$$(21)\quad a_{n\alpha} = \sum_{i=\alpha}^{\alpha+N-1} F_i \sin(n\omega_\theta ih+\phi_\theta)$$

$$(22)\quad b_{n\alpha} = \sum_{i=\alpha}^{\alpha+N-1} F_i \cos(n\omega_\theta ih+\phi_\theta)$$

$$(23)\quad A_{n\alpha} = (a_{n\alpha}^2 + b_{n\alpha}^2)^{1/2}$$

$$(24)\quad \theta_{n\alpha} = \operatorname{atan}(a_{n\alpha}/b_{n\alpha})$$

Where ω_θ is the radian frequency of analysis,

ϕ_θ is the phase of analysis,

n is the harmonic number,

F_i is the input waveform at time ih

N is the nearest integral number of samples in one period of the input waveform.

The initial phase angle, ϕ_θ, is included for generality. The method is independent of this phase angle.

The summations are taken over one period of the input waveform. Since N must be an integer, we can not analyse for an arbitrary frequency whose period may not be an integral number of samples. We must settle for taking the nearest integer. Having chosen the number of samples in the summation, we must then set ω_θ to $2\pi/Nh$. If this is not done, a very strange kind of inaccuracy sets in. We will show an example of this presently.

We apply equations (21) through (24) to the digitized waveform of a single note of constant frequency for each harmonic of the waveform. This produces two output waveforms for each harmonic. The waveform represented by $A_{n\alpha}$ in equation (23) corresponds to the amplitude of the harmonic as a function of time. The waveform represented by $\theta_{n\alpha}$ in equation (24) corresponds to the phase of the harmonic as a function of time. We may convert this to frequency by taking the slope of the function at each point in time. This may be done with a band-limited differentiator [Kaiser 1963, 1966].

To better understand what the heterodyne filter does, we may examine its output when a pure sinusoid is applied. The heterodyne filter is a nonlinear filter, so the principle of superposition does not apply. Equations (21) and (22), however, are linear. The transformation to equations (23) and (24) does not change certain principles. If a signal is annihilated entirely by equations (21) and (22), it will not be present in the outputs of either equations (23) and (24). Signals greatly suppressed in equations (21) and (22) will be greatly suppressed in equations (23) and (24).

If we apply a pure sinusoid of frequency ω, we may compute the output of the heterodyne filter exactly by means of the summation calculus [Hamming 1962].

$$(25)\quad A_{n\alpha} = \frac{1}{4N^2}\sin^2\left(\frac{\omega N h}{2}\right)\left\{\csc^2\left[\frac{(\omega+n\omega_\theta)h}{2}\right]+\csc^2\left[\frac{(\omega-n\omega_\theta)h}{2}\right]\right.$$

$$\left.+\frac{2\cos[n\omega_\theta h-2\phi_\theta]}{\sin\left[\frac{(\omega+n\omega_\theta)h}{2}\right]\sin\left[\frac{(\omega-n\omega_\theta)h}{2}\right]}\right\}$$

The expression for the phase is not included here because it is so complex as to be almost meaningless. Equation (25) is plotted in figure 24a. The frequency of analysis was the 5th harmonic of 500 Hz. We can see that the response is identically zero for all multiples of 500 Hz except the 5th.

It is interesting to compute the limit of the exact expressions for the response to a pure sinusoid. If we define $\Delta\omega$ to be $(\omega-n\omega_\theta)$, the limits may be computed as shown in equations (26) and (27).

$$(26)\quad \lim_{\omega\to n\omega_\theta} A_{n\alpha} = \frac{1}{4N^2}\{\theta+N^2+\theta\} = \frac{1}{4}$$

$$(27)\quad \lim_{\omega\to n\omega_\theta} \frac{a_{n\alpha}}{b_{n\alpha}} = \frac{\sin\left\{2n\omega_\theta h\left[\frac{N-1}{2}+\alpha\right]\right\}+N\sin\left\{\Delta\omega h\left[\frac{N-1}{2}+\alpha\right]\right\}}{\cos\left\{2n\omega_\theta h\left[\frac{N-1}{2}+\alpha\right]\right\}+N\cos\left\{\Delta\omega h\left[\frac{N-1}{2}+\alpha\right]\right\}}$$

The first important point is that the results are, in the limit, not dependent upon the absolute phase of the input sinusoid. Also, the magnitude of the output converges to a constant times the amplitude of the input sinusoid. The phase converges to a linear function of the frequency difference, $\Delta\omega$, if the number of points in the summation, N, is large compared to 1.

USAGE

The biggest problem with using the filter is that the assumptions upon which it is based are rarely true. That is, all music instruments have harmonics that change with time, and many have frequencies that are not exact multiples of the fundamental frequency. Since the principal source of error due to these deviations from the ideal comes from "leakage" from adjacent harmonics, the output may be improved somewhat by further filtering of these harmonics.

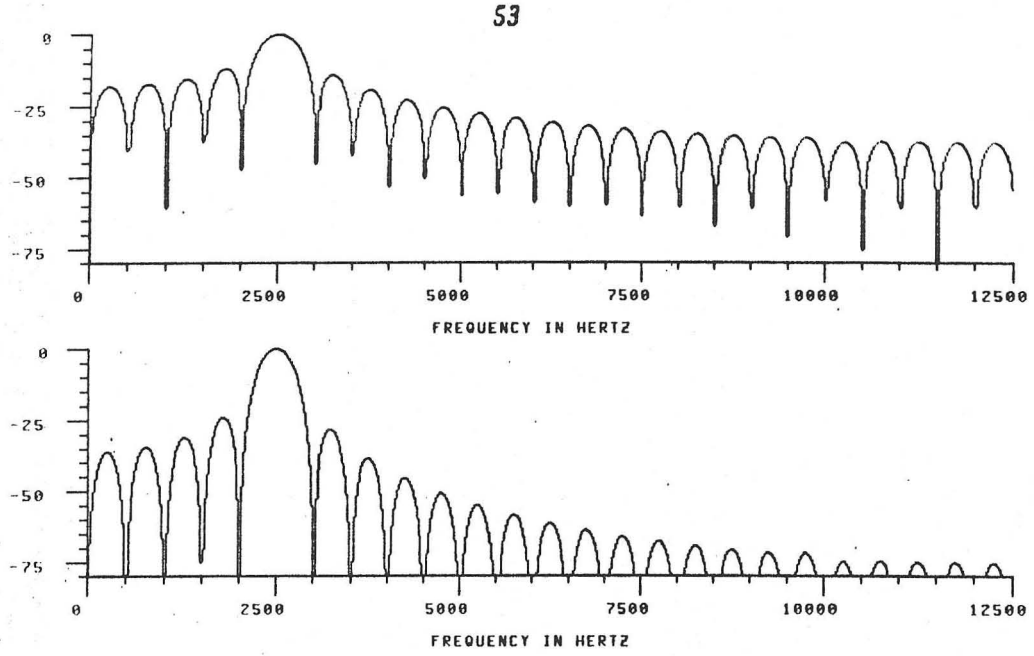


FIGURE 24. Equation (25) Evaluated for a wide range of frequencies. In this figure, we are analysing the fifth harmonic of a 500 Hertz tone. This is effectively the frequency response of the heterodyne filter for a particular tone. In the lower plot, the output has been smoothed by averaging over one period of the fundamental.

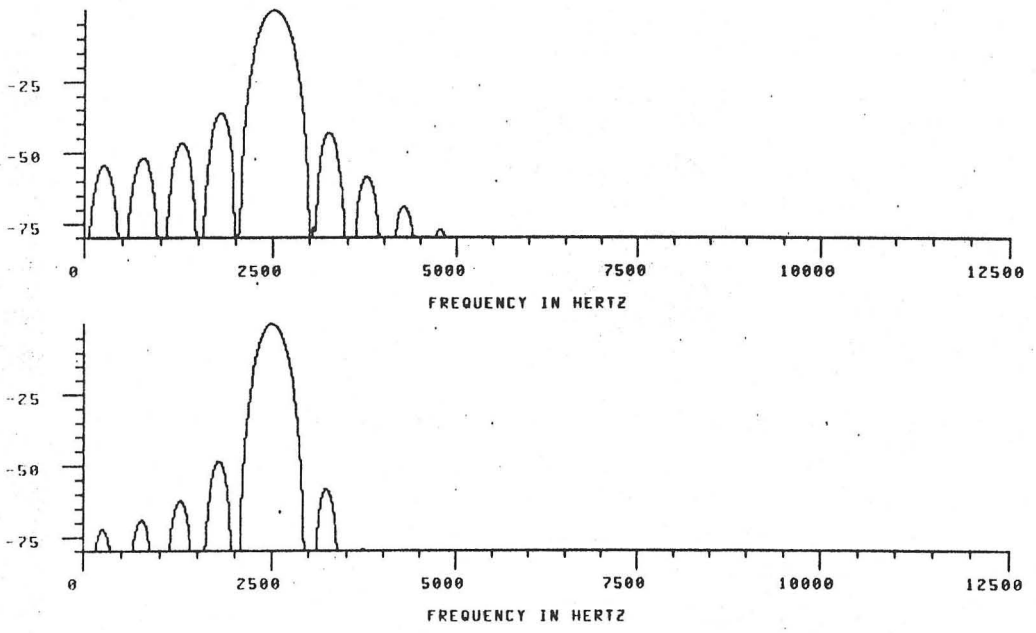


FIGURE 25 This is equation (25) evaluated, as above, for the fifth harmonic of a 500 Hertz tone and then smoothed by averaging over one period. The upper plot has been smoothed twice, and the lower plot has been smoothed three times.

Since the important part of the output of the heterodyne filter is around zero frequency, we can simply filter out the harmonics other than the one under analysis by replacing each point in the output by the average over one period of the fundamental frequency. This places an additional zero of transmission over each other harmonic. Figures 24b, 25a, and 25b show the results of applying such a filter once, twice, and three times. The sideband rejection becomes quite strong. We could use a classical filter, like the Butterworth or Chebychev low-pass design, but this would not put a zero of transmission at the other harmonics. We feel this feature is very important.

To get the slope of the phase function, we replace each point by the slope determined by a least-squares fit of a linear polynomial centered around that point. This provides further noise reduction by averaging as well as producing a band-limited approximation to the slope at each point.

Figure 26 shows a plot of the amplitudes of the harmonics of a music instrument tone. Time is the axis going from left to right (about .5 seconds total), and frequency is depth into the page. The first harmonic is in the rear. Figure 27 shows a spectrogram-like plot of this data as well as the detailed frequency deviations of each harmonic as functions of time. The analysis technique as described so far was used to analyse 16 music instrument tones for a study in perception of musical timbre [Grey 1975].

Tones were synthesized from these data. Putting the tones in this form allowed them to be normalized independently for pitch, duration, and loudness, as well as to be modified and blended. The synthetic tones were judged quite similar to the original tones. This is, of course, the final test of the analysis procedure. Appendix A shows the results of analysing several synthetic tones to determine how much perturbation the filter can tolerate before producing results that are grossly in error. It would appear that as much as a 2 percent deviation in frequency with rise times as short as 5 periods can be tolerated with reasonable results.

It is of interest to list the ways that this technique has been misused in the past with the hope that future users will avoid these problems.

As was described in the historical review, Luce used a method that was very similar to this, but limited by the extreme cost of computer time in those days. He selected single periods of the waveform and interpolated them to get exactly 24 points per period. He then did the summations to produce amplitudes and phases for 12 harmonics. Note that this method only gives one 24 numbers per period, whereas the heterodyne filter gives one 2NM numbers, where N is the number of points in a period and M is the number of harmonics under analysis. The advantage of this extra computation is that a particular difficulty of Luce's is avoided. The following quote is from Luce's thesis:

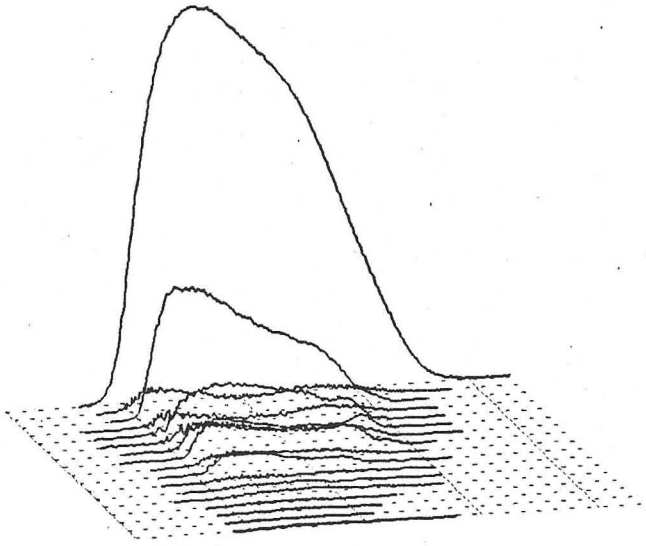


FIGURE 26. Perspective plot of analysis data from heterodyne filter for a clarinet tone, shown as an Amplitude x Frequency x Time perspective plot. The detailed frequency variation of each harmonic is not shown here. (X = time; Y = amplitude; Z = frequency, with the fundamental harmonic plotted in the background).

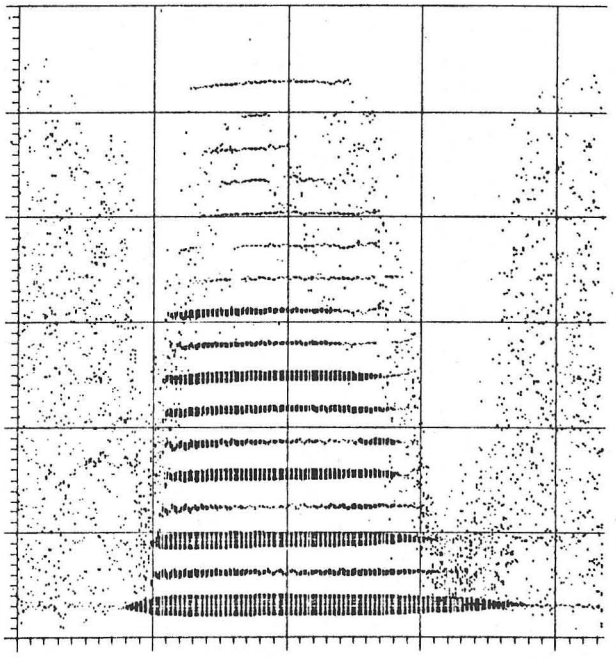


FIGURE 27. Analysis data from heterodyne filter for the same clarinet tone as in figure 26, shown above in the form of a spectrographic plot (X = time, with 1/10 second lines; Y = frequency, with KHz lines; Width of bars = relative dB to -40).

> "Another very serious difficulty arises for waveforms containing very narrow pulses well-separated from each other if only 24 ordinates per cycle of the fundamental of the note analyzed are used. Two neighboring data points are used in each interpolation. It is possible that none of these 48 data points, corresponding to the 24 points in time selected for interpolation during the cycle, contain the narrow pulse. Because of this phenomenon, a small error in the measurement of the fundamental frequency of the note may result in the pulse being missed in some cycles entirely and being selected in others. Large fluctuations (from cycle to cycle) in the calculated spectral components result."

By taking all the points in a period, we avoid this problem. We cannot, however, avoid a small (order of 1/N) fluctuation due to the fact that the true period is not an integral multiple of the sampling interval. Since this fluctuation is periodic with the same period as the note, the further filtering operations eliminate it entirely.

Pulse-like waveforms are quite common in music. All brass instruments have pulse-like waveforms. The human voice is often quite pulse-like. Pulse-like waveforms cannot be ignored in musical contexts.

Beauchamp and Freedman both thought of the summations in equations (21) and (22) as discrete analogs of the Fourier integrals. This is dangerous because it leads one to sum over one period, but to use an analysis frequency (ω_0) which does not correspond to a period equal to an integral multiple of the sampling interval. This produces imperfect pole-zero cancellation and all the resulting distortion. They too obtained only "a few" points per period, letting themselves in for the same kind of errors Luce's method obtains.

Beauchamp later used the FFT algorithm [personal communication 1974] with a Hamming window. The Hamming window is equivalent to a convolution in the frequency domain. It is equivalent to replacing each frequency-domain point ($a_{n\alpha}$, $b_{n\alpha}$) with the sum of itself and a portion of its neighbors [Bertram 1970; Blackman and Tukey 1959]. This means that "leakage" between adjacent harmonics, that very problem we have tried so hard to filter out, is directly encouraged by the application of a window function. Figures 28a and 28b show the frequency response of a filter designed this way. The zeros of transmission at the neighboring harmonics have been removed. This method cannot possibly produce accurate results.

This technique can be salvaged by doing the analysis at one-half the frequency (twice the period). This will produce an output that has only even harmonics, indicating a tone an octave high. This way, when we analyze for a certain harmonic, the adjacent "harmonics" will, of course, be zero, because the odd harmonics will be zero. This way, anything the technique produces on the odd harmonics can be ignored as artifacts of the analysis.

Keeler [1972] used Lagrangian interpolation to produce a much higher effective sampling rate and then computed an approximation to the Fourier integral by use of Simpson's rule. Even if we ignore the fact that the Lagrangian interpolation does not have good band-limiting

properties [Schafer and Rabiner 1973], there is a severe problem with the use of Simpson's rule rather than direct summation when considered from a signal-processing point of view.

With Simpson's composite rule, the successive samples are weighted by the following coefficients: 2, 4, 6, 4, 6, 4, . . ., 4, 6, 4, 2. The weighted samples are then summed. The problem is that this is equivalent to the sum of three separate weights:

```
first:   2, 2, 2, 2, 2, 2, . . ., 2, 2
second:  0, 2, 2, 2, 2, 2, . . ., 2, 0
third:   0, 0, 2, 0, 2, 0, . . ., 0, 0
```

We see that the first sequence is pure summation. The second sequence is a summation, but over N−2 points; a different fundamental frequency. The third sequence has every other sample zero, which is characteristic of a sampling rate a factor of 2 slower. This means that massive aliasing occurs, as well as annihilating the zeros of transmission. Figure 29a shows the frequency response of such a filter. We can see the aliased band up in the high frequency range, as well as the fact that the response no longer goes exactly to zero at every other harmonic. Probably the only reason that Keeler got as good results as he did is because he was analysing large organ pipes, which presumably had few high harmonics, and thus little aliasing. Figure 29b shows what happens if just a straight triangle rule is used. The plot does not show it, but the minima in the frequency response are not actually zeros of transmission. The use of the triangle rule has made the response non-zero at each of these points. This is because it is equivalent to the sum of two weightings, one of length N and one of length N−2.

Thus we see that there are a number of ways of doing this process incorrectly. It is hoped that this exposition will help others to find even better methods.

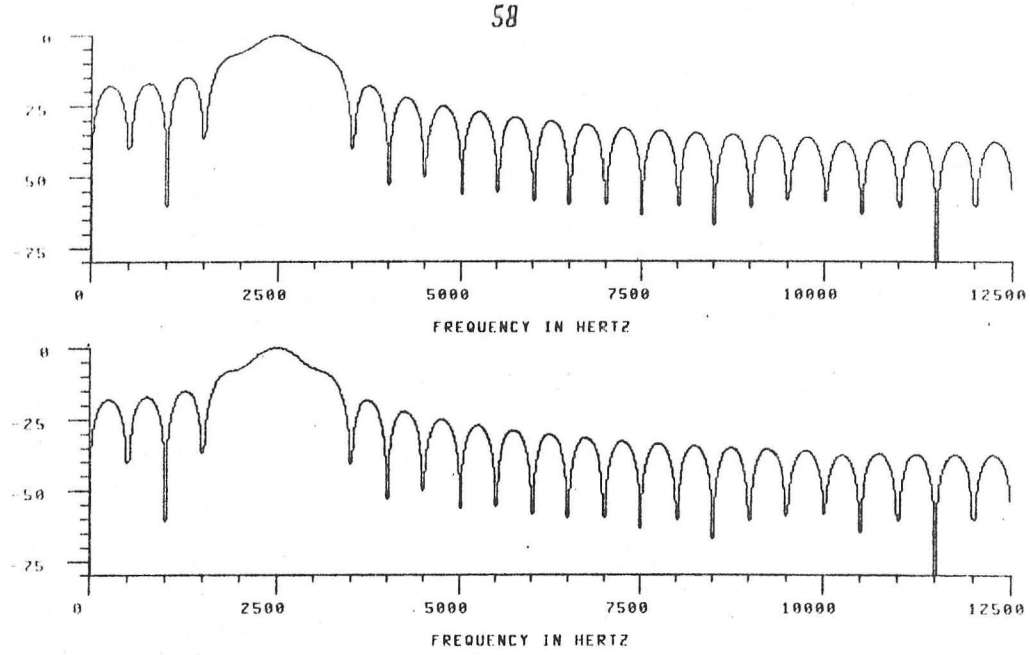


FIGURE 28. This is the magnitude frequency response of the heterodyne filter when a "Hanning" window function is applied. Since windowing in the time domain is equivalent to convolution in the frequency domain, the spectral zeros at the fourth and sixth harmonics go away. This cannot give accurate results. The lower plot uses the "Hamming" window.

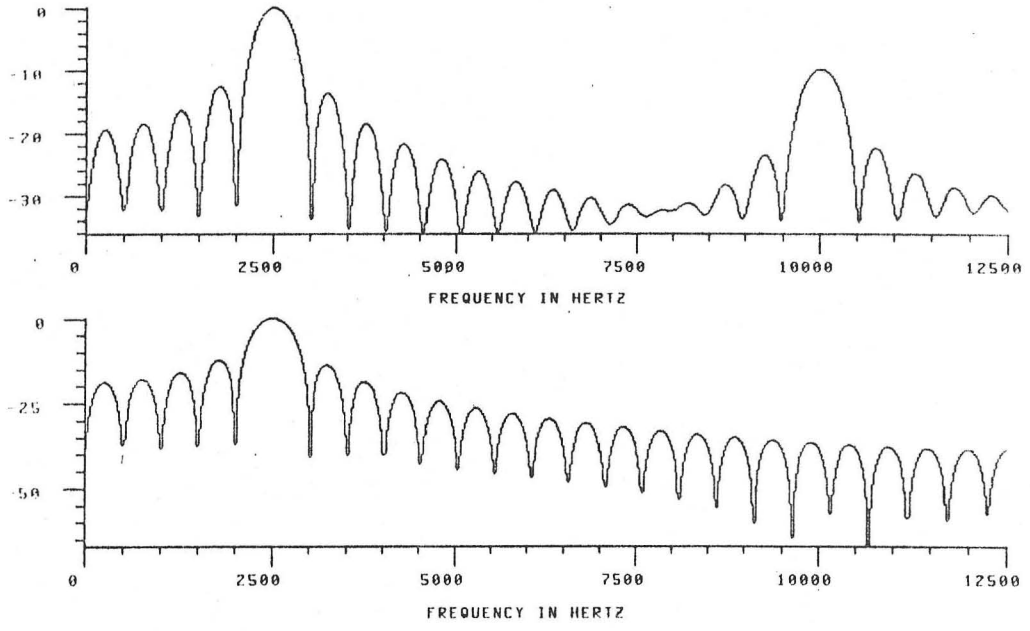


FIGURE 29. This is the result of approximating the Fourier integral by Simpson's composit rule. An effective halving of the sampling rate and corresponding aliasing occurrs. The lower plot approximates the integral by the triangle rule with somewhat better success.

BANDPASS FILTERING

INTRODUCTION

The bandpass filter is one of the oldest techniques for separating out a single harmonic. Backhaus [1927, 1932] used a bandpass filter for studying individual harmonics of music instrument tones, notably the violin. The bank-of-filters method of speech analysis has been widely used. There is much evidence that the basilar membrane in the ear is like a bank of bandpass filters.

We will not attempt to repeat the wealth of literature that exists on linear systems and linear filters, but let us just review some basic principles of filtering in general.

The output of a filter consists of its *particular* response and its *homogeneous*, or *transient* response. The particular response is directly related to the input signal. In fact, the spectrum of the particular response is just the product of the spectrum of the input signal and the frequency response of the filter. The transient response is, however, somewhat more complicated.

Any linear filter has what are called *natural frequencies*. These can be resonances or anti-resonances. The transient response of a filter is made up of sinusoids of these frequencies.

There is a relation between the frequency selectivity of a filter and how fast it can respond to changes in the input signal. A very narrow-band filter has a very long transient response and changes very slowly. This is illustrated in figures 30 and 31. In the first figure, we see the response of a very narrow band filter to a suddenly-applied pure sinusoid. The second figure shows the response of a wide-band filter to a suddenly-applied sinusoid. With this in mind, let us see how the bandpass filter can be used in practice.

USAGE

If we suspect that a harmonic exists at a certain frequency, we can use a bandpass filter to select it from a complex signal, with some ensuing loss of resolution in time. In fact, unlike the heterodyne filter, any sinusoid of nearly-constant frequency can be selected. It does not have to be harmonically related to any other sinusoids in the signal. Figure 32 shows, in the top plot, the response of a 4th order bandpass filter (Butterworth, 30 Hz between the 3dB points) to a complex signal. The center frequency of the filter is set to exactly the frequency of one of the harmonics of the signal. Notice the smooth amplitude envelope of the harmonic. The upper plot in figure 33 shows the output of a filter with the same input as the previous figure but its center frequency does not correspond to any partial in the input signal. The response consists almost entirely of transient response. The particular response is highly suppressed, as it should be.

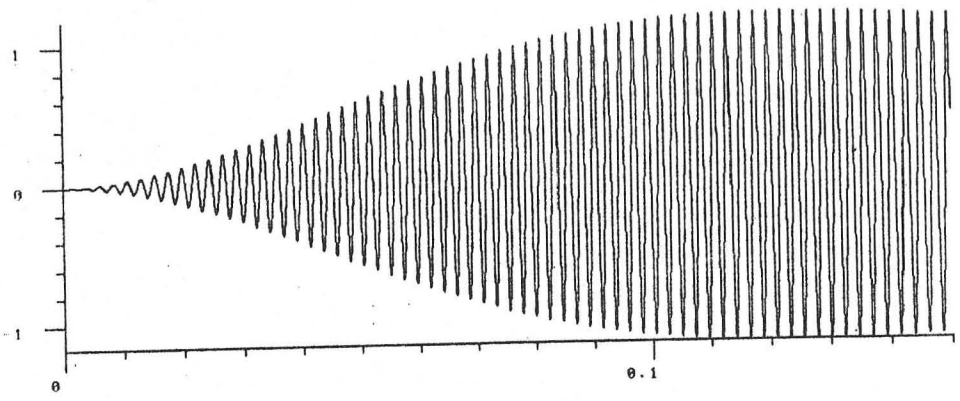


FIGURE 30. The response of a narrow bandpass filter to a pure sinusoid applied suddenly.

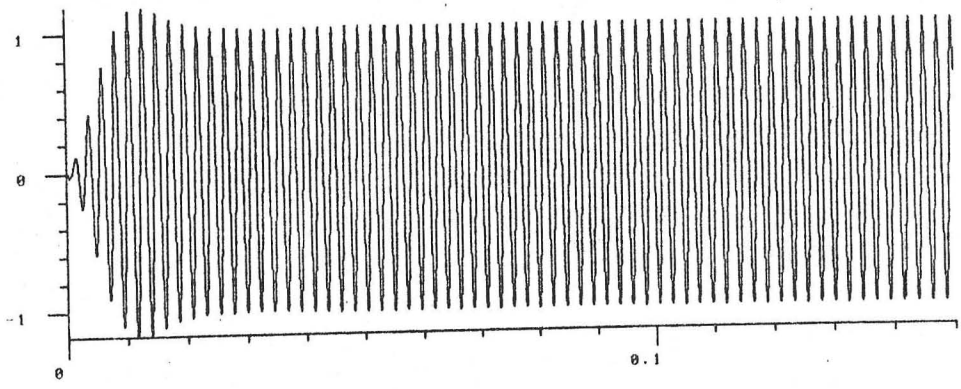


FIGURE 31. The response of a broad bandpass filter to a pure sinusoid applied suddenly.

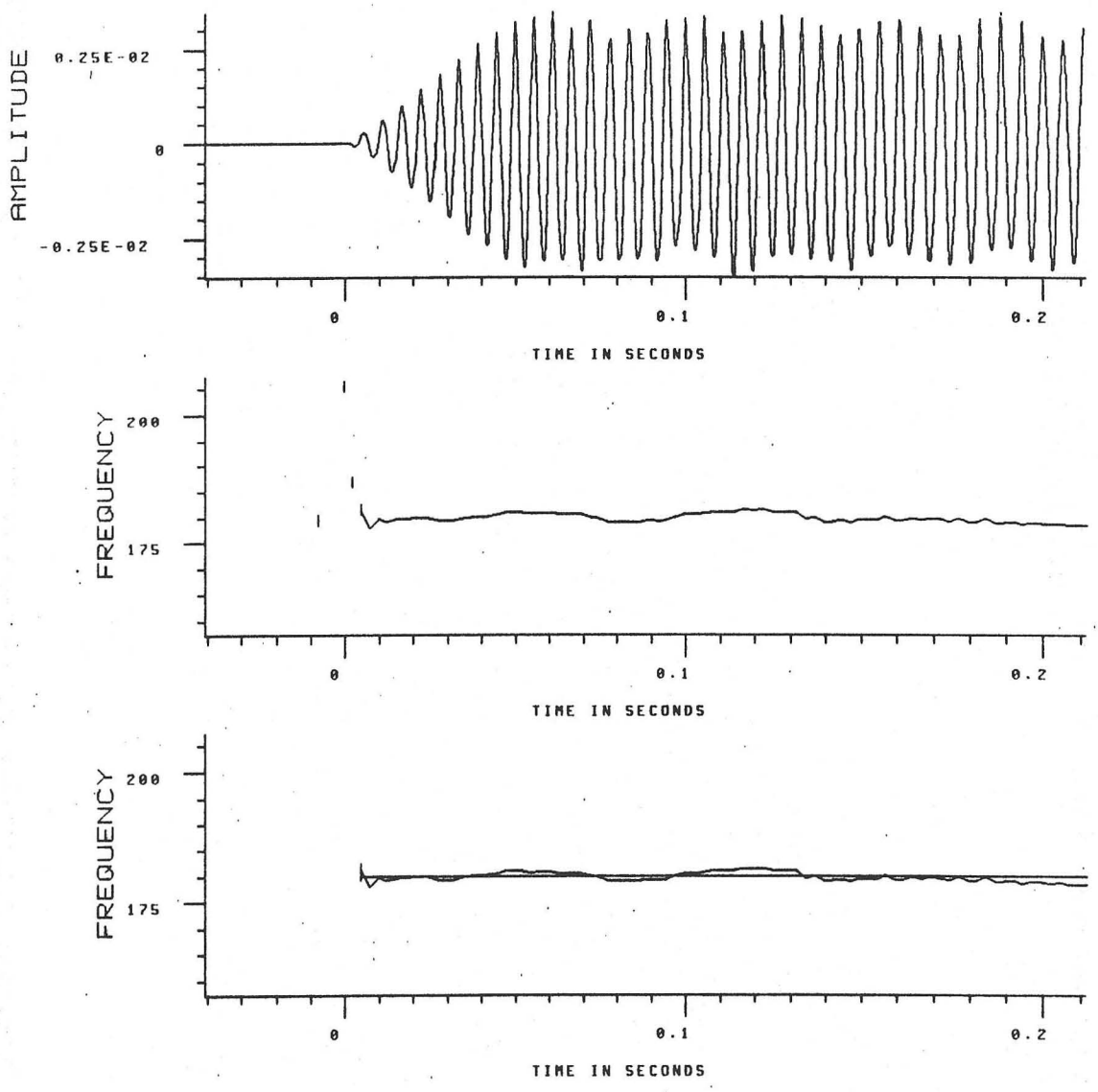


FIGURE 32. These three plots show steps in the processing of the fundamental harmonic of a piano tone in a piano duet. The upper plot shows the response to a bandpass filter the center frequency of which coincided closely with the frequency of the harmonic. The center plot shows the results of applying the optimum-comb to the waveform in the upper plot. The minima in adjacent time slices have been linked by a nearest-neighbor rule to form lists representing the frequency of the signal as a function of time. A vertical stroke has been placed at the beginning of each list. The lower trace shows the results of eliminating obviously spurious frequency lists. The dominant list has a horizontal line drawn through it representing the average frequency of the harmonic. The vertical stroke at the beginning of this line is two standard deviations high.

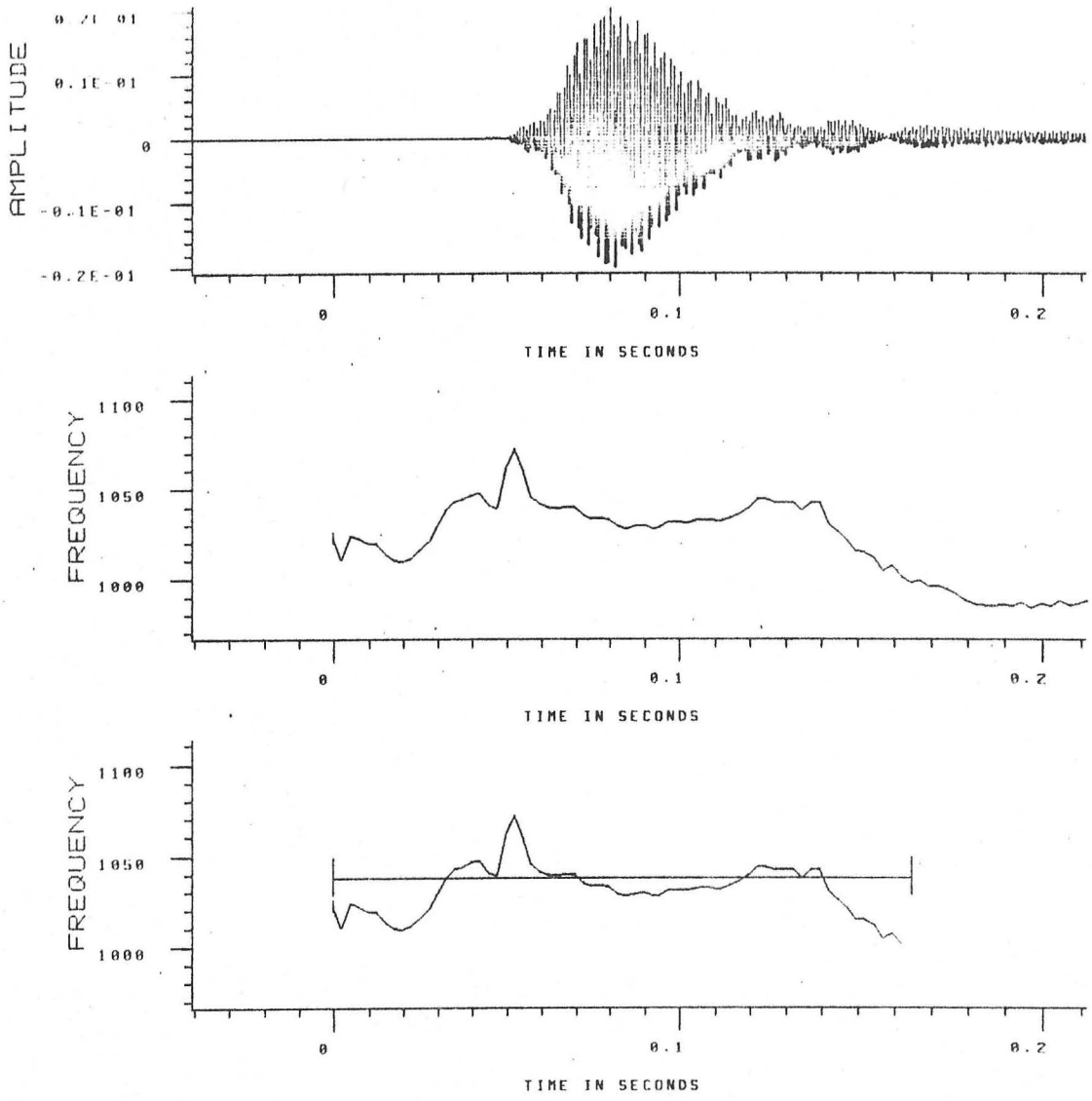


FIGURE 33. This figure, like the previous one, shows the processing of a single harmonic extracted from a polyphonic piece by a narrow bandpass filter. The upper plot shows the output waveform of the filter. The center plot shows the results of the application of the optimum-comb to detect any periodicity which may be present in the filter output waveform. The minima of the optimum-comb have been linked together to form lists. In the lower plot, obviously spurious traces have been eliminated. The remaining list has a horizontal bar through it denoting the average frequency in the list. There is, in fact, no sinusoid present at this frequency. This is a transient response and is entirely an artifact of the bandpass filter. This trace will hopefully be eliminated later due to its large frequency variation.

We may apply a pitch detector to the output of the bandpass filter to get the frequency of the harmonic as a function of time. This is also a good way to tell if there is really something there or not, because the output of the pitch detector will be gibberish if there is not a near-sinusoid present. The center plot in figures 32 and 33 shows the output of a pitch detector (the optimum comb) applied to that output of the bandpass filter shown in the upper plot. As we see, the frequency varies smoothly throughout the duration of the plot. If no harmonic is present, we do not get a consistent reading of pitch throughout the duration of the signal, thus no trace like the one shown is produced.

If the center frequency of the filter is very low, it is possible that the pitch detector can track sub-harmonics of the lowest harmonic in the sound at that point. Some of this low harmonic will sneak through the filter and fool the pitch detector. As was shown before, the autocorrelation-type pitch detectors respond just as well to integral multiples of the fundamental period as to the fundamental period itself. Figure 34 shows multiple traces of subharmonics of a harmonic produced by the optimum-comb technique. To eliminate the spurious traces (all of the traces in this figure are spurious), we may make some other crude measurement of the pitch which does not have this problem and compare the results. One simple technique is just to count the zero-crossings in the filter output. This provides a crude estimate of the pitch of the signal and is enough to eliminate the spurious traces.

To use the filter, we must know how to set its center frequency. One convenient method is to use a pitch detector (autocorrelation and comb filtering have been previously described) to get an estimate of the harmony of the signal. Since music uses ambiguous chords, we may expect several significant pitches to be indicated. We may then apply bandpass filters to all multiples of these pitches, up to some maximum. This will get approximations to the harmonics with limited resolution in time. We may then apply a pitch detector (again, autocorrelation or comb filtering will do) to get the frequency of the harmonic as a function of time, and we may average the energy of the signal to estimate the amplitude of the harmonic as a function of time. The bottom plot in figure 32 shows the final frequency contour of a harmonic of a complex signal. The straight line through the plot indicates the average frequency of the harmonic. The vertical bar at the beginning of the horizontal line is two standard deviations high. Figure 35 shows what happens if the center frequency of the filter is not exactly upon the frequency of the harmonic. This trace was not accepted, as is shown by its absence from the lower plot. The frequency deviation throughout the trace was unacceptably great.

In practice, the use of a pitch detector to determine which bandpass filters to apply only reduces the number of applications of the filter by a factor of about 3 from a dense covering. For example, a 30 Hz bandwidth was used in the analysis program. A dense covering from 100 Hz to 2000 Hz would be about 200 applications. In fact, only about 75 applications were needed. This is still a lot. It is enough so that this method of analysis can hardly be called practical at this point in time. Perhaps with the advent of high-speed special-purpose signal processing hardware, the method may become more than a demonstration. It should be noted that just as much time was spent doing the pitch detection on the filtered waveform as was spent doing the filtering itself.

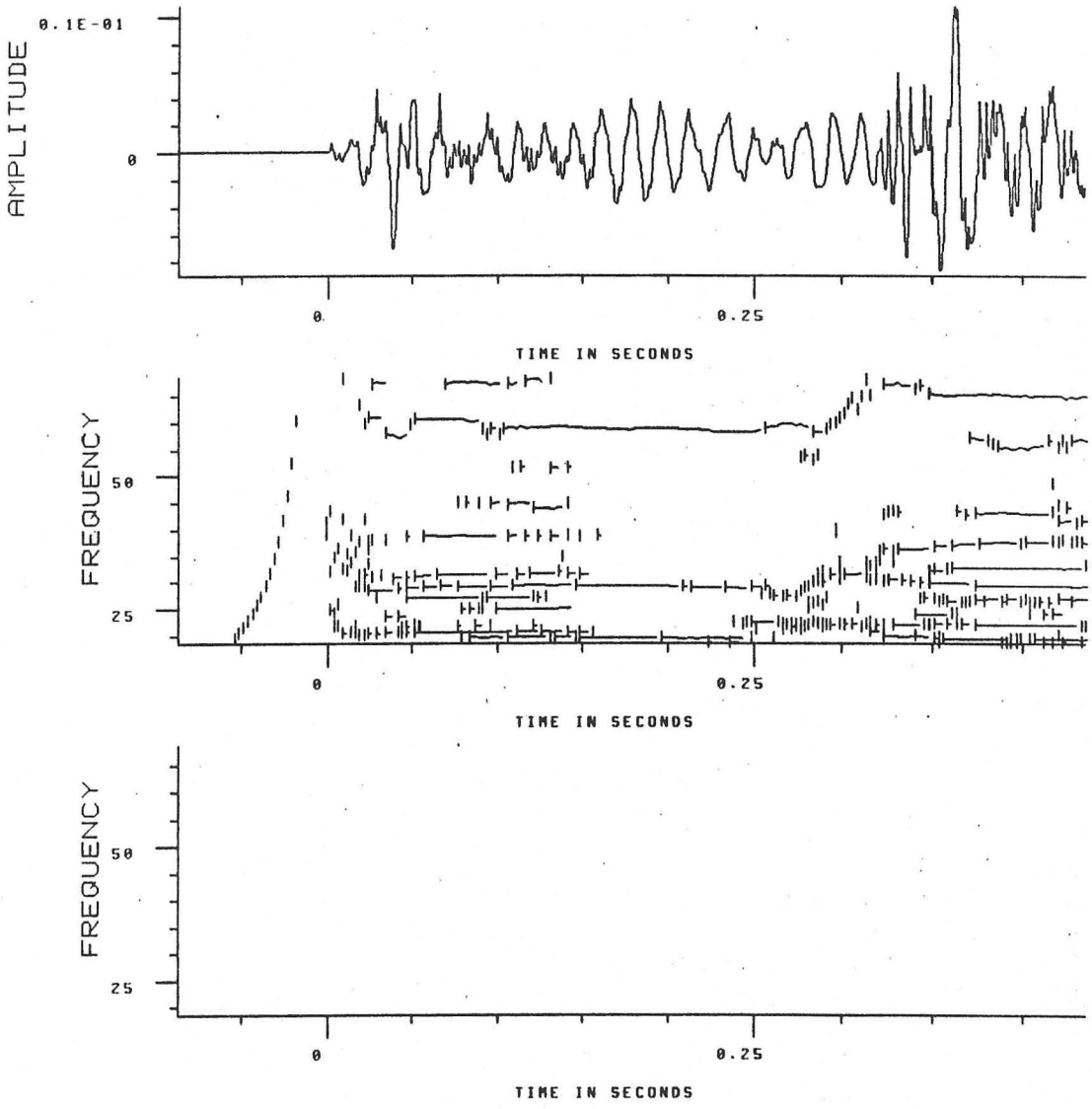


FIGURE 34. This figure is similar in format to the previous two. This shows the results of applying a bandpass filter at a very low frequency. The filter does transmit the lowest sinusoid in the signal greatly attenuated. The optimum-comb cannot by itself distinguish subharmonics of the filter output, so it finds many minima. These are linked into lists and shown in the center plot. A vertical stroke is placed at the beginning of each list. To eliminate subharmonics, we count the zero crossings in the filter output. This gives a rough pitch estimate that is sufficient to eliminate all the spurious subharmonic traces, as is shown in the bottom plot.

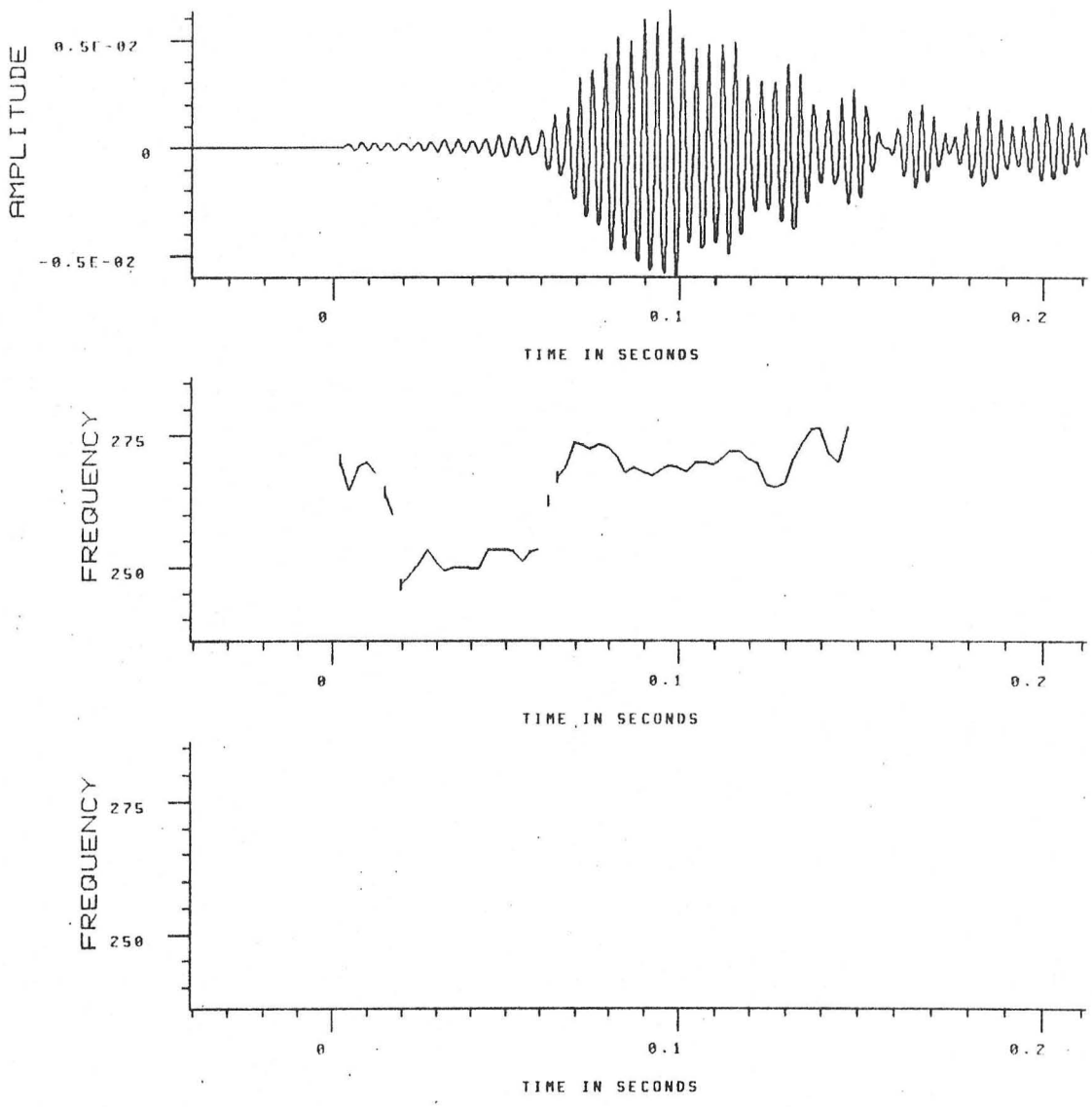


FIGURE 35. Here we see the results of applying a bandpass filter the center frequency of which does not correspond to any partial in the piece at that time. The filter, of course, passes in attenuated form the composit waveforms of the neighboring partials. The optimum-comb found some minima to track, as is shown in the center plot. Since the list of frequencies found by the optimum-comb is highly variable, it can be eliminated on this basis alone, as is shown in the lower plot.

POPULAR TECHNIQUES NOT FOUND USEFUL

INTRODUCTION

In this section, we will expose some of the weaknesses in other popular signal processing techniques that make them not useful for the musical scribe. We present these negative results for several reasons, perhaps the most important being the fact that the science and art of digital signal processing is new enough that a great deal of experience with its techniques has not had time to accumulate. Each of the techniques to be discussed has been found to be very useful in general. The linear predictor forms the core of most speech analysis systems in use today. The FFT is the "workhorse of the industry". The cepstrum is useful in speech as well as picture processing, sonar, radar, and many others.

THE CEPSTRUM

INTRODUCTION

The cepstrum is defined as the inverse DFT of the log of the magnitude of the DFT of an input signal. This may sound a bit perverse, but if we recall that the autocorrelation of two time-limited signals can be computed by the inverse DFT of the magnitude of the DFT of an input signal, we can see that the processes are related. The cepstrum of a signal is *a signal (a function of time) whose DFT is the log-magnitude of the DFT of the input signal.* The cepstrum is a time sequence, just like the signal itself, and also like the autocorrelation function.

The cepstrum is useful for dealing with signals that have been multiplied or convolved with other signals. For instance, we may think of the speech production mechanism as an excitation (the glottis) followed by a filtering operation (the vocal tract). In picture processing, the signal can be represented as the excitation (the light source) multiplied by the reflectance function of the illuminated object. In each of these cases, the log-magnitude DFT is related to the *sum* of the transforms of the individual signals. If these signals, by themselves, occupy different parts of the spectrum, then they can be separated by simply partitioning the cepstrum. In this manner, we may use the long-time end of the cepstrum to detect the pitch of a speech waveform [Noll 1967], or the short-time end of the cepstrum to compute an approximation to the impulse response of the vocal tract [Oppenheim 1968, 1969; Miller 1974]. In speech, the signals separate nicely.

One place where the cepstrum may be of great use in music is in analysis for the purpose of synthesis. Since we can separate the functions of periodicity generation from spectral shaping with the cepstrum, we may use it to generate the impulse response of a filter which can duplicate, as a function of time, the spectral shape of the waveform of a music instrument. Since a number of instruments are almost perfectly periodic (brasses, most woodwinds except during the attack), it may be possible to synthesize many tones using these impulse responses. There are, however, a large number of instruments which are not perfectly periodic (all stringed instruments) and are thus not suitable for simulation in this manner, unless some technique for deriving and modeling the excitation function is found. (We can compute the excitation function simply from the long-time part of the cepstrum, but unless we can model it more simply, it is not amenable to modification and is thus not useful for musical purposes).

DISCUSSION

The problem with using the cepstrum to compute, say, the pitch of music instruments is that in polyphonic music, we are dealing with the *sum* of a number of waveforms. When we take the log of the magnitude of the DFT of the input signal, we get a very complex result where the signals do not partition nicely. The information for each voice is spread all over the cepstrum in complex ways. For instance, figure 36 shows the cepstrum of a single violin tone. Notice the single peak corresponding to the period of the input signal.

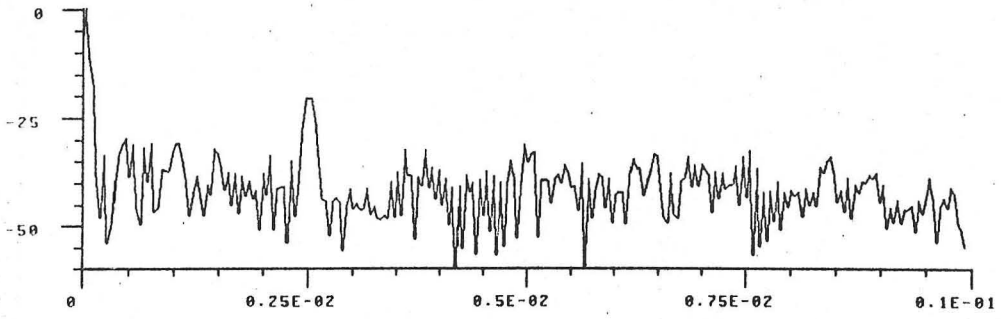


FIGURE 36. This is the cepstrum of a segment of the waveform of a trumpet solo. The waveform was taken from the first note of Ravel's orchestration of Tableaux D'une Exposition. The note is a G4, or about 396 Hz. As we see, a single peak is evident at about 25 milliseconds, which represents the period of the detected signal. The cepstrum is quite insensitive to reverberation, as the trumpet was recorded in a large concert hall with extensive reverberation.

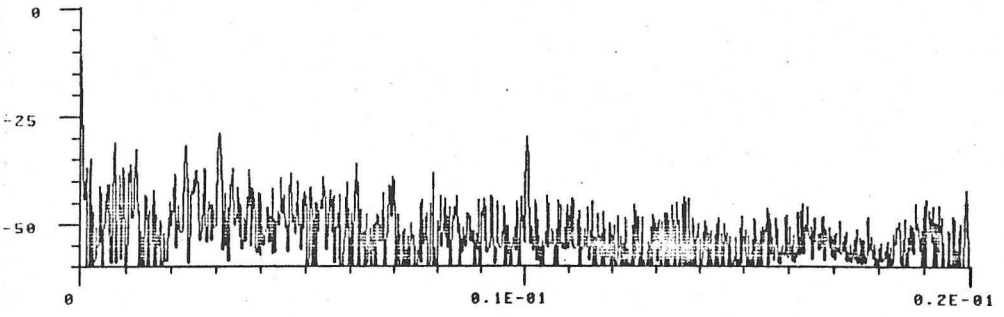


FIGURE 37. This is the cepstrum of a segment of the waveform of a brass choir. The waveform was taken from the first brass chord of Ravel's orchestration of Mussorgsky's Tableaux D'une Exposition. The cepstrum does not seem to produce a distinct peak corresponding to any periodicity in the input signal in this polyphonic case.

Figure 37 shows the cepstrum of two violins being played at different frequencies. The peaks no longer correspond to frequencies in the original signal. There is no clear way to extract from the cepstrum the information about the pitches of the two notes being played.

THE DFT

INTRODUCTION

The Fourier transform in all its many forms is possibly the oldest and most widely useful signal processing technique of all. Special processors to compute the DFT by the Fast Fourier Transform algorithm [Cochran *et al* 1967; Gentleman and Sande 1966; Gold and Rader 1969; Rabiner and Gold 1975; Oppenheim and Schafer 1975; Singleton 1967, 1968, 1969] are available from numerous sources. When we began this project, the DFT was the first technique called upon to help accomplish the task. It was later abandoned for reasons that will be explained below. It may, in fact, be possible to accomplish the task at hand with the DFT, but certain problems would have to be solved which did not seem to have simple solutions.

DISCUSSION

Let us begin by examining the DFT of a pure sinusoid with an exponential amplitude. The (complex) signal that we shall transform is as follows:

$$(28)\quad S_n = e^{n(\sigma+j\omega)T}$$

Where S_n is the value of the sinusoid (the input signal) at time nT, where T is the time between consecutive samples

σ is the decay rate. $1/\sigma$ is the *time constant* of the signal, *i.e.*, the time it takes the signal to decay to $1/e$ of its value at time=$\emptyset$.

ω is the radian frequency of the sinusoid.

j is the square-root of -1.

The transform can be computed as follows:

$$(29)\quad A_k = \sum_{n=\emptyset}^{N-1} S_n e^{-2\pi nkj/N} = \sum_{n=\emptyset}^{N-1} e^{n\{(\sigma+j\omega)T-2\pi jk/N\}}$$

A_k is the k^{th} value of the discrete Fourier transform. It represents the frequency $k/(NT)$.

N is the number of points in the transform.

Since this is just the sum of a finite exponential series, we can compute this summation in closed form:

$$(30)\quad A_k = \frac{e^{\{N(\sigma+j\omega)T-2\pi jk\}}-1}{e^{\{(\sigma+j\omega)T-2\pi jk/N\}}-1}$$

After some manipulation, we find that the squared magnitude of this expression is then the following:

$$(31)\quad |A_k|^2 = e^{(N-1)\sigma T}\ \frac{\sinh^2(N\sigma T)+\sin^2(N\omega T-2\pi k)}{\sinh^2(\sigma T)+\sin^2(\omega T-2\pi k/N)}$$

It is easy to show that this expression is maximized when the following is true:

$$(32)\quad k = \frac{N\omega T}{2\pi}$$

This maximum is unique in the range $0 \le \omega T \le \pi/2$. We can see from the expression above that the peak widens as N gets smaller and as σ gets larger. Figure 38b a shows equation (32) evaluated for N=128, figure 38d for N=2048, and figure 38f for N=16384. We see that as N is increased, the peak becomes sharper and sharper.

Figures 38a, 38c, and 38e show the actual DFT of a pure sinusoid at 314.159265 Hz evaluated by the fast Fourier transform algorithm for 128 points, 2048 points, and 16384 points. The results differ from the calculated values because of roundoff error. In the longer transforms, the error manifests itself as a spreading of the peak. It is roughly analogous to a multiplicative noise (rather than an additive noise).

Likewise, figures 39a, 39b, 39c, and 39d show the spreading of the peak as σ increases. The reciprocal of σ is the attack time in seconds, so σ increasing means faster and faster attack. A σ of 100 implies a 10 millisecond attack, which is quite common in music waveforms.

These cases were idealized. In general, the attack is not a pure exponential. Figure 40 shows the DFT of a segment of a 2-voice piano piece. The time window is centered over the boundary between two notes. The lower voice persists throughout the window at a constant C4 (261.6 Hz). The upper voice is changing between an E4 (329.6 Hz) and an F4 (349.2 Hz). It is clear that the region around the E4 and the F4 is quite muddled with many peaks in evidence. This DFT used 4096 points and occupied about 200 milliseconds width in time.

There is another problem with the use of the DFT for sounds that were recorded in highly reverberant rooms. In this case, the effect of the room can be modeled by a linear time-invariant filter. The music is then convolved with the impulse response of the room. This is equivalent to multiplying the transform of the music by the frequency response of the room (or adding the logarithms of the transforms). Since it is well known that concert halls have frequency responses with many narrow peaks and valleys of depth up to 20 and 30 dB [Schroeder 1962, 1962, 1970], these peaks and valleys can produce spurious peaks in the DFT of music recorded in such a room.

Figure 41 shows the DFT of a 200 millisecond segment near the center of the first block chord. This chord is a G-minor chord. It has notes at G2 (98 Hz), G3 (196 Hz), Bb3 (233.1 Hz), D4 (293.7 Hz), and many more. We can notice many spurious peaks. In the region of the Bb3 (233.1 Hz), there is an extra peak that is only 5 dB lower than the main peak. The same is true of the G3 (196 Hz).

For these reasons, we decided not to use the DFT in this investigation. Later on, we show cases where we used the DFT as the front end of a hypothetical music analysis system and compare the results with our preferred implementation.

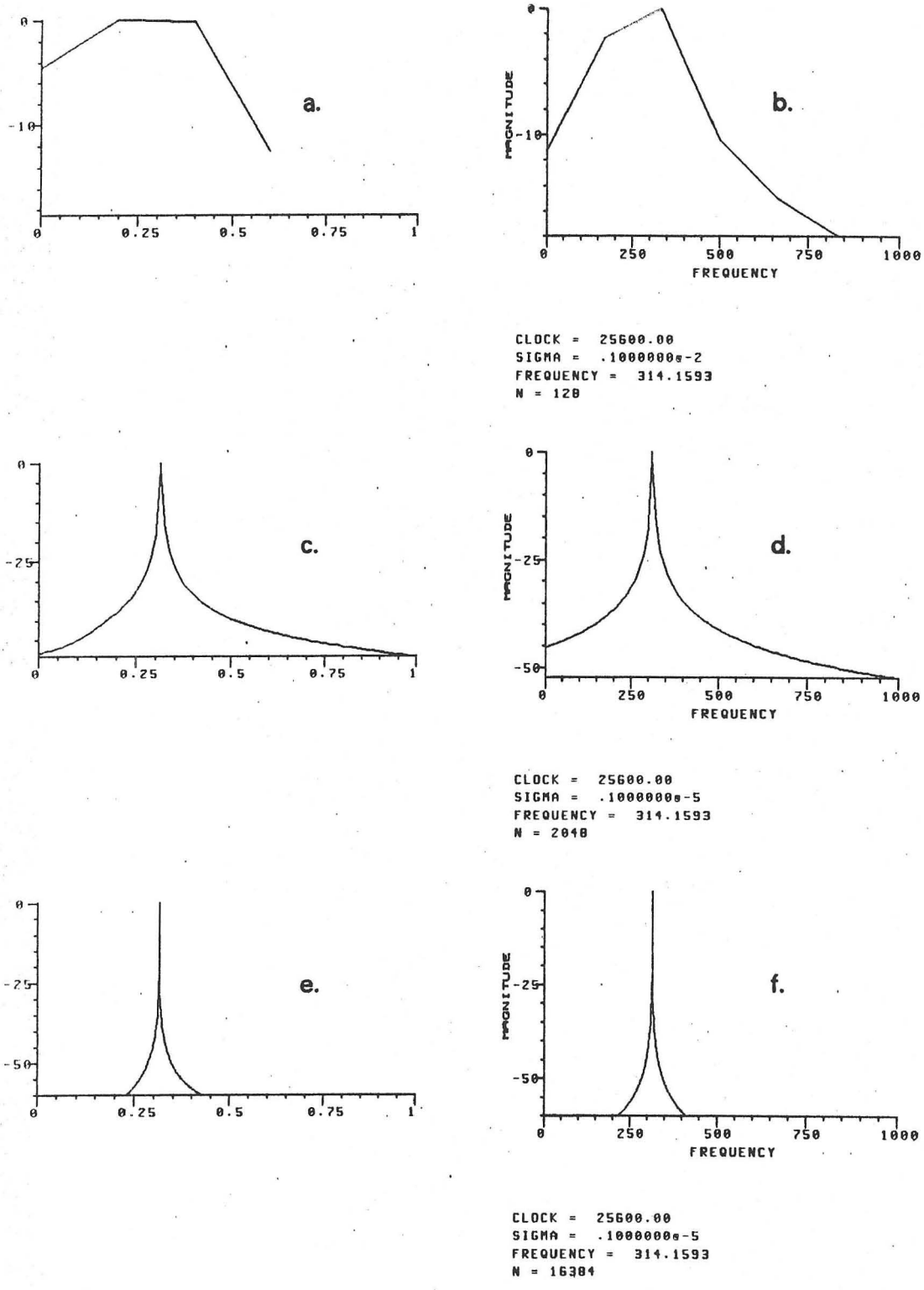


FIGURE 38. Comparison of predicted (right-hand column, figures 38b, 38d, and 38f) and actual (left-hand column, figures 38a, 38c, and 38e) DFT of pure sinusoid with DFT of increasing amounts of data. The top figures, 38a and 38b, used a 128 point DFT. The center pair, 38c and 38d, used 2048 points. The bottom pair, 38e and 38f, were done with 16384 points. The discrepancies are due to roundoff error which tends to increase the apparent background noise.

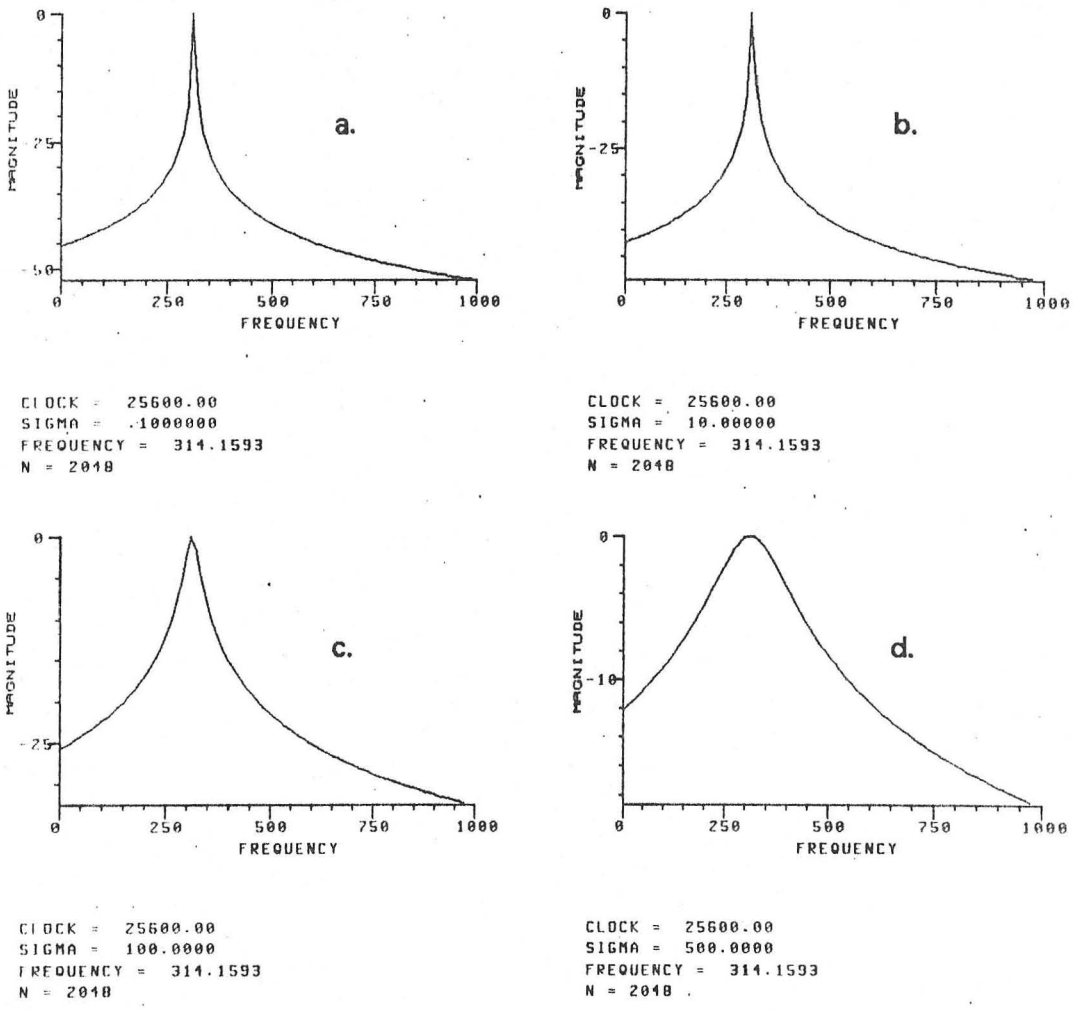


FIGURE 39. Comparison of predicted DFTs of exponentially-damped sinusoids for different values of damping factor, σ. Figure 39a has a damping factor σ=1, which represents a 1 second decay time. Figure 39b has σ=10, or 100 milliseconds decay time. Figure 39c has σ=100 for 10 milliseconds decay. Figure 39d has σ=500 for 2 milliseconds decay. We can see from this that the more transient the waveform is, the more the peak in the DFT is spread.

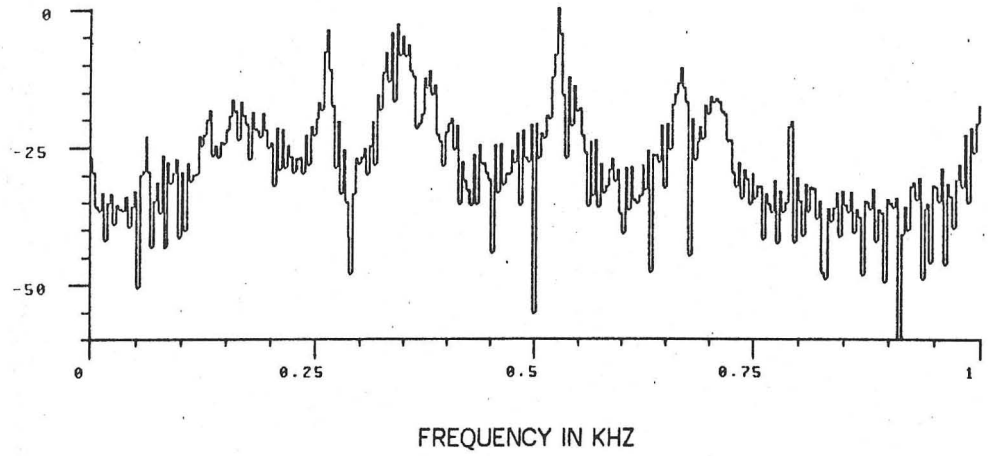


FIGURE 40. Discrete Fourier transform of a 4096 point (200 millisecond) segment of a piano duet. The time window is centered over the boundary between two notes. The lower voice persists throughout the window at a constant C4 (261.6 Hz). The upper voice is changing between an E4 (329.6 Hz) and an F4 (349.2 Hz). The region around the E4 and F4 is quite muddled with many peaks in evidence.

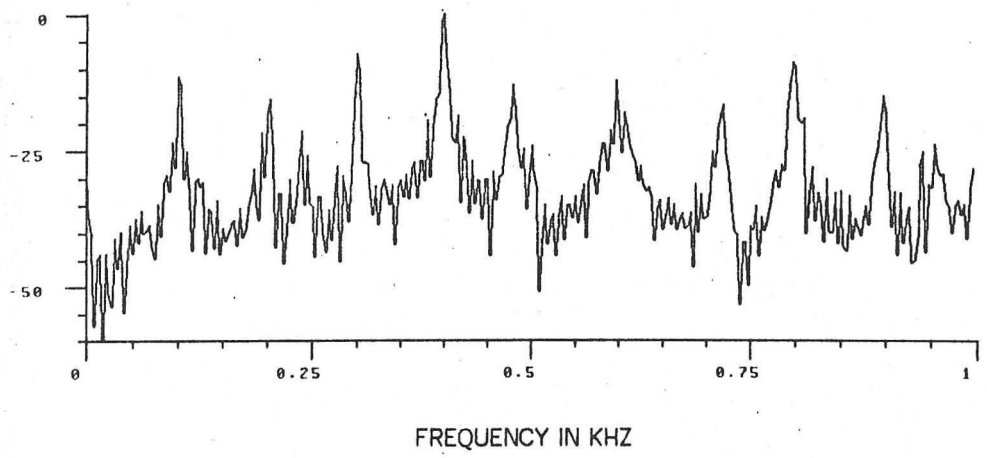


FIGURE 41. Discrete Fourier transform of a 4096 point (200 millisecond) segment selected from the center of the first G-minor brass chord in Tableaux D'une Exposition. Some of the principle notes present in the chord are G2 (98 Hz), G3 (196 Hz), Bb3 (233.1 Hz), and D4 (293.7 Hz). This recording was made in a highly reverberant concert hall. Since this is equivalent to multiplying the transform of the music with the frequency response of the concert hall, we see many superfluous peaks representing the natural modes of the hall. Near the Bb3 (233.1 Hz) there is an extra peak that is only 5 dB lower than the main peak. This causes considerable confusion in trying to use the discrete Fourier transform for polyphonic music analysis in reverberant environments.

THE LINEAR PREDICTOR

INTRODUCTION

The linear predictor [Atal and Schroeder 1968; Atal and Hanauer 1971; Boll 1973; Griffiths 1975; Itakura and Saito 1968, 1970, 1971; Levinson 1947; Wiener 1947; Makhoul and Wolf 1972; Makhoul 1975; Markel 1972] is a technique for computing an all-pole filter the frequency response of which best approximates the spectrum of the input signal. It has become very popular recently in the speech community because one can approximate the spectrum of a speech signal and then determine the formant regions by examining the frequency response of this filter. It provides much-needed smoothing of the spectrum, giving quite often clear, unambiguous peaks at the formant frequencies. This technique belongs to the world of "system estimation", in that the filter thus created models the filtering activity of the vocal tract. The linear predictor *estimates the system* consisting of the resonant regions of the vocal tract.

DERIVATION

A simple way to derive one form of the linear predictor was given by Markel [1972]. First, we define a linear finite impulse response filter of the following form:

$$(33)\quad A(z) = 1 + \sum_{i=1}^{M} a_i z^{-i}$$

Where $A(z)$ is the Z-transform of the filter transfer function.
Z is the *unit time-advance operator*
a_i are the coefficients of the difference equation that defines the filter, shown below equation (35).

If X_i is the input sequence and Y_i is the output sequence of the filter, we may obtain the energy in the output of the filter by merely suming the squares of the output of the filter.

$$(34)\quad \text{Energy} = \sum_{n=0}^{L} Y_n^2$$

Where Y_i is the output of the filter at time iT.

and also:

$$(35)\quad Y_n = X_n + \sum_{i=1}^{M} a_i X_{n-i}$$

After substituting (35) into (34), differentiating with respect to a_i, setting the energy to zero, and collecting terms, we get the normal equations for the filter coefficients:

$$(36)\quad \sum_{i=1}^{M} a_i \sum_{n=0}^{L} X_{n-i}X_{n-k} = -\sum_{n=0}^{L} X_{n-k}X_n$$

for k=1, 2, . . . M

This is a system of linear equations in the variables, the a_i. It can be solved in a number of efficient ways [Levinson 1947; Markel 1972]. It produces a filter that best reduces the input sequence to zero. Such a filter has a frequency response that is the inverse of the spectrum of the input signal. We can invert the filter simply by making it an all-pole filter, using the coefficients, a_i, on the delayed output signal rather than the delayed input signal. This filter has a frequency response that approximates the spectrum of the input signal. This is a discrete realization of the Wiener-Hopf integral [Levinson 1947; Wiener 1947; Lee 1960], and uses the RMS error criterion for optimality. This technique also belongs to a larger topic of "system estimation" [Tribolet 1974; Sage and Melsa 1971], where one attempts to infer a linear system from its impulse response. A superb review of linear prediction may be found in Makhoul [1975].

USAGE

This is commonly used in vocoder and speech analysis systems. For vocoder use, the input speech is processed for pitch, voiced-unvoiced decision, and filter coefficients a_i. These parameters are transmitted to the receiving station. The speech is then resynthesized using a pulse train at the computed pitch for voiced excitation, and white noise for the unvoiced excitation. The filter then simulates the spectral shaping imposed by the vocal tract.

This technique can also be used to aid pitch detection. The input signal is filtered by the inverse filter. This evens out the spectrum, removing the effects of the formants. The resulting waveform is much more pulse-like. This output can then be autocorrelated to produce peaks which are much more sharp than those produced by autocorrelating the unfiltered waveform.

This technique of "spectral flattening" or "prewhitening" does not apply to polyphony. Unless the filter is of extreme order, making it expensive to compute, the interleaved harmonics of the notes will not be adjusted equally. The autocorrelation then shows one sharp peak corresponding to the dominant tone and a multiplicity of other peaks, corresponding to the other tone.

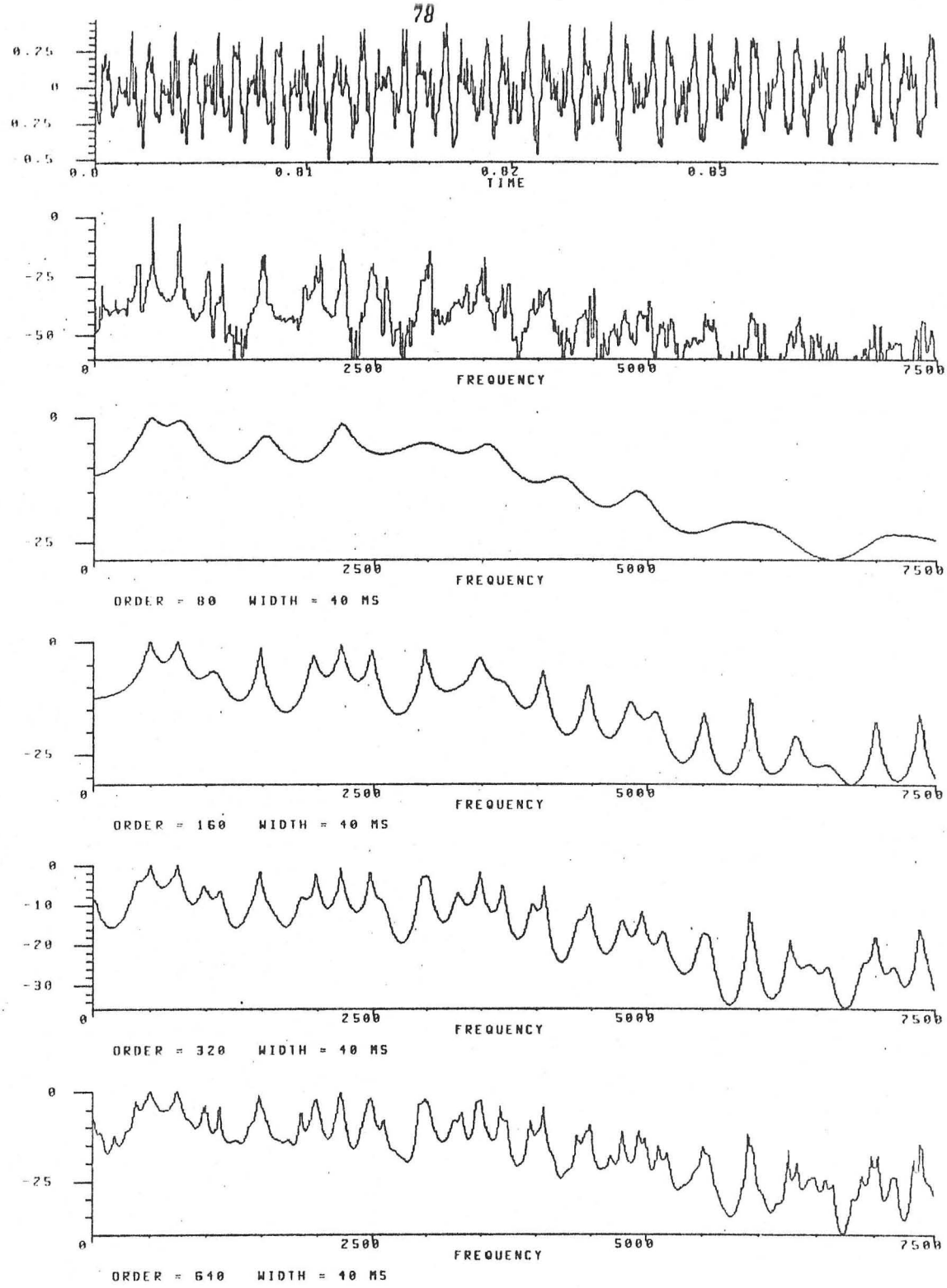


FIGURE 42. Frequency responses of filters computed by the linear predictor for different filter orders. The top plot is the sound waveform itself. The second plot is the discrete Fourier transform of that sound waveform. There are two violins playing here. The sound segment is 40 milliseconds long. The next plots are the magnitude frequency responses of linear predictors of orders 80, 160, 320, and 640 respectively. As the order approaches the number of points in the sound sample, the frequency response of the filter approaches the magnitude of the discrete Fourier transform of the sound sample.

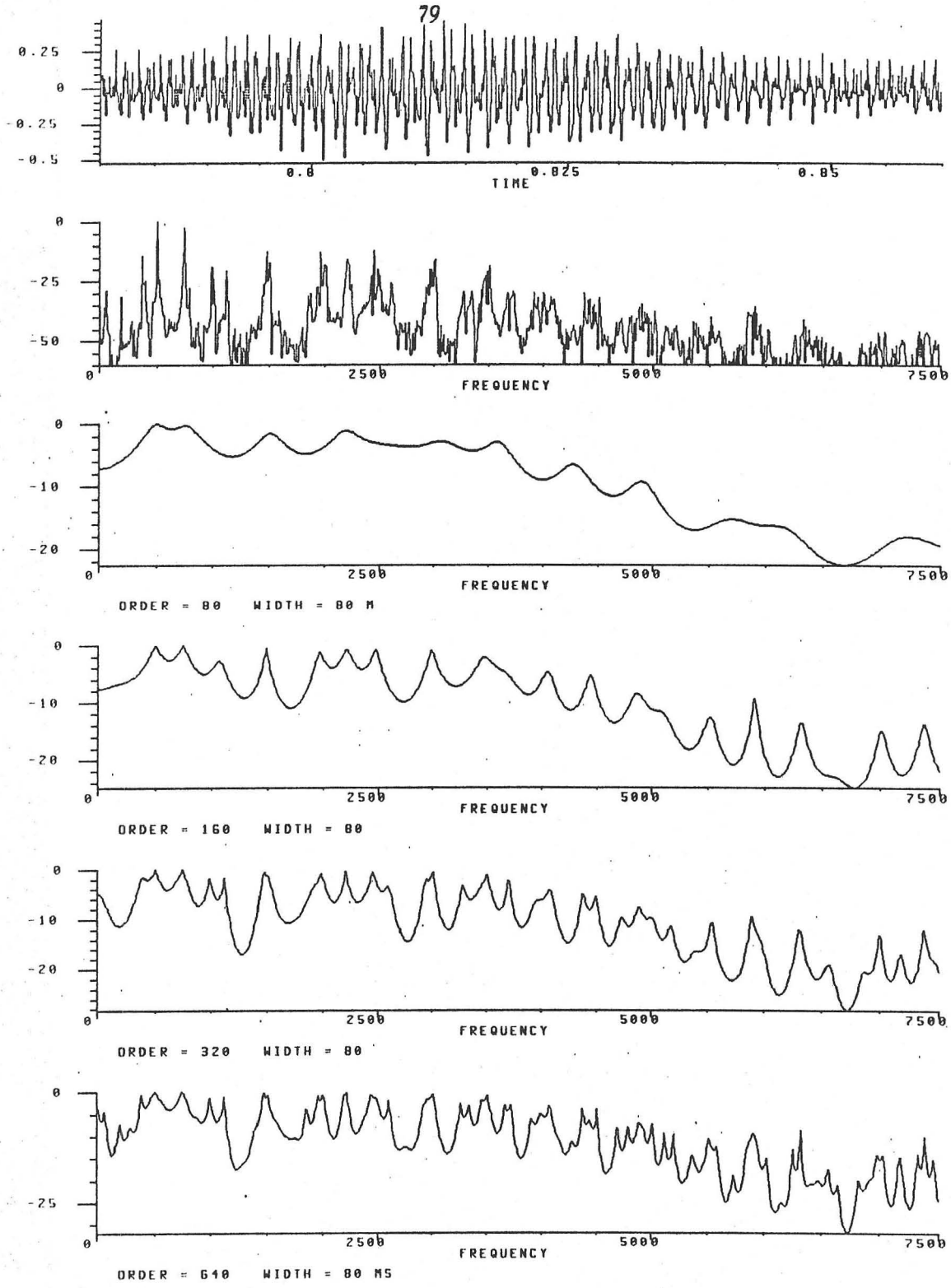


FIGURE 43. Frequency responses of filters computed by the linear predictor for different filter orders. As in the previous figure, the top plot is the sound waveform itself. The second plot is the discrete Fourier transform of that sound waveform. The sound segment is 80 milliseconds long. The next plots are the magnitude frequency responses of linear predictors of orders 80, 160, 320, and 640 respectively. Increasing the size of the sound sample from 40 to 80 milliseconds has the effect of sharpening the peaks in the transform. It also lessens the chances that the signal will be homogeneous throughout the interval.

Another possible usage would be to compute a filter of high enough order that it simulated the harmonics themselves as high-Q resonances. Figures 42 and 43 show frequency responses of filters of various order computed by the autocorrelation method [Markel, Itakura]. As we see, the frequency response approaches the spectrum as the order is increased. This points up again that the linear prediction algorithm is a spectral matching process [Makhoul 1972]. Since the DFT itself has not proved useful in this task, there is no reason to believe that an approximation to the DFT would be any more useful.

Griffiths [1975] used this method for determining the frequencies of a number of sinusoids which were added together. With a 12 pole filter and a 25 dB signal-to-noise ratio, he obtained estimates for the frequencies of up to three sinusoids added together. The error was as much as 12 percent, and sometimes peaks were not even located. In our case, we must detect up to 40 sinusoids and determine the pitches to better than 3 percent in all cases.

INTERCONNECTION

OVERVIEW

The music analysis system as it was implemented for the purposes of this thesis combines the previously discussed low-level routines into a complete system. This is done in the following steps:

An estimate of the frequencies present is obtained by running the optimum-comb pitch detector over the entire music sample at 10 millisecond intervals. We call these "windows" into the sound file. If a particular period appears in many consecutive windows, a list is made of its occurrences. A list is redundant if it is a harmonic of some other list. Redundant lists are eliminated. This produces a list of regions which have the same periodicities present. These are regions wherein the harmony does not change. These are arbitrarily grouped into larger regions so that more data may be dealt with at once. These macro-regions are then used as the guide for the bandpass filter.

The bandpass filter is set to all harmonics of all the periodicities that are present in a given macro-region up to a certain maximum frequency. For the examples shown later, a maximum frequency of 1.5 KHz was sufficient. Any more comprehensive system would have to use a much higher frequency range than this. The output of the bandpass filter is run through an optimum-comb pitch detector which is swept over the frequencies in the passband of the filter. The minima of the optimum-comb output are linked into lists which indicate the existence of a frequency at that pitch over the time that the minima are found. The amplitude envelope of the filter output indicates the amplitude function of the harmonic in question. It is these amplitude and frequency functions that are passed to the intermediate-level routines for scoring and grouping into notes. Before we leave this level, many checks are done to throw out traces that are obviously spurious.

We will first discuss the theoretical basis and the constraints on the music that allow us to analyse it in this manner. We will then discuss the details of the algorithms.

THEORETICAL BASIS

To allow this dissertation to be completed in a finite amount of time, certain restrictions have been placed on the music that will be allowed. These restrictions, combined with the properties of music instruments, make the problem manageable. These properties and restrictions are discussed below.

ALL TONES ARE NEARLY PERIODIC

This restricts the class of instruments to woodwinds, brass, strings, and some percussive instruments (piano, marimba, etc). This assumption allows us to infer a note from its harmonics. It insures that notes will have harmonics. It does not tell us what the harmonic structure will be, or how the harmonic structure changes with time. It can still be that the note will not have a first harmonic (a sinusoid at the fundamental frequency). The note can also consist of a single sinusoid. Later, in the intermediate-level processing, further restrictions will be placed on the tones. For the low-level, this is sufficient.

ALL FREQUENCIES ARE NEARLY PIECEWISE-CONSTANT

This restriction eliminates strong vibrato, glissandi, and other cases of non-constant pitch. This allows us to filter out a single harmonic by using a filter of a constant frequency. We are assured that the tone will not jump out of the range of one filter and into the range of another. Vibrato can be tolerated up to a point, but some intermediate-level routines attempt to model the sound as having constant frequencies, and would thus make errors if strong vibrato was present.

THE FUNDAMENTAL OF ONE NOTE WILL NOT OVERLAY A HARMONIC OF ANOTHER NOTE

This is very important. If the fundamental frequency of a note is the same as the frequency of a harmonic of another note that is sounding at the same time, it appears to be very difficult to distinguish this case from the case of a single note with a complex harmonic structure. It is not clear how (or that) we distinguish these cases. It is possible that we hear differences in the times that the instruments begin, or that we can distinguish because the instruments are invariably at slightly different pitches. It is clear that a more advanced transcription system should be able to separate the notes in these cases. It is certainly the case that separate vibratos on the tones makes them aurally separate much more convincingly. The subject of when a group of harmonics fuses into a single percept has not been researched fully in the past. Rather than attempting to solve the problem here, we will finesse it by requiring that the input music not exhibit that property. Or likewise, if it exhibits the property, we will not expect the higher note to appear in the output manuscript. This gives us the property that a set of harmonics uniquely imply their fundamental. All we must deal with is noise and processing error which may cause some harmonic to be missed. We do not have to try to expand a single set of harmonics into more than one note.

THE PIECE CONTAINS NO MORE THAN TWO VOICES

This restriction allows us to compute the musical harmony from the periodicity of the waveform without having to worry about whether some voice is lost because it is masked by several other voices. When using the diatonic scale, any two notes imply a harmony, thus a two-voice piece will always imply at least one root frequency, and generally will imply several.

OTHER CONSIDERATIONS

We also expect the tones to be smooth. The amplitude and frequency functions of the harmonics of music instrument tones vary slowly with time, except during the attack and decay portions of the note. Since these portions are relatively short, compared to the total length of a note, we need not consider them. This assures us that the amplitude and frequency contours will be continuous and will not vary greatly. This is important, because then we can use this smoothness criterion to eliminate noisy traces. This eliminates certain instruments, like drums and cymbals, which not only do not have harmonics, but they do not have smoothly varying partial tones. This also eliminates heavy reverberation. Recording in a highly reverberant room causes phase and amplitude jitter in each harmonic. Each time a reflection reaches the microphone, the attack of the note with all its inharmonicity occurs again. Figure 44 shows the amplitude and frequency trace of a harmonic from a piece that was recorded in a highly reverberant concert hall. The jitter due to the reflections is quite apparent here in both the amplitude and frequency plot.

With the above restrictions, we have some hope of accomplishing the task. Let us look now at how the routines can coax out the secrets of the input waveform.

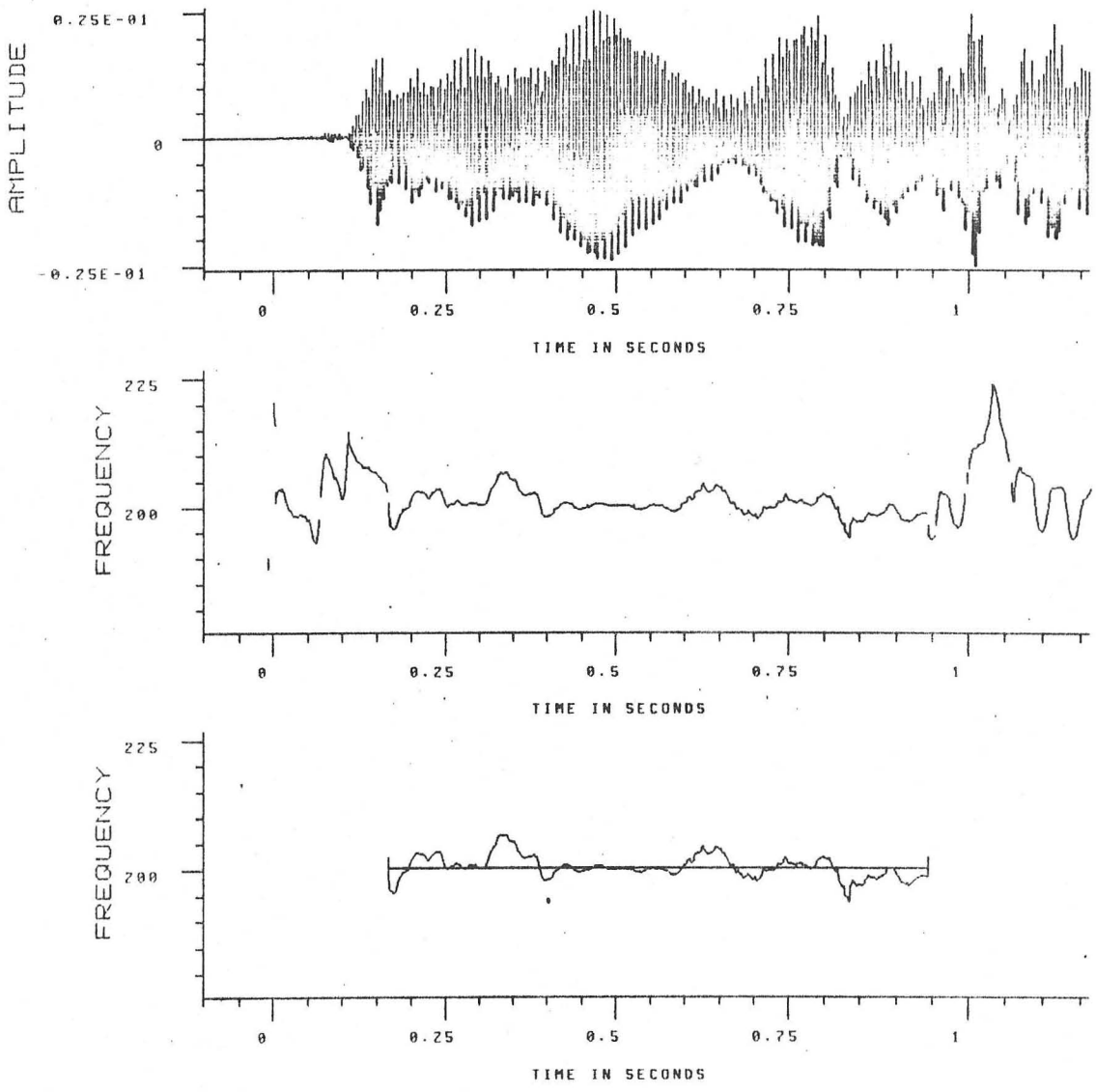


FIGURE 44. The upper plot shows the waveform of the output of a bandpass filter centered at G3 (196 Hz) on the first brass chord in Tableaux D'une Exposition. The center plot shows the pitch as a function of time as tracked by the optimum-comb. The jitter both on the amplitude of the signal and on the frequency is due both to the extremely reverberant environment of the concert hall and the choral effect of having many musicians playing the same note (or notes at octaves). The notes and their harmonics beat highly due to the inevitable mistunings among the musicians. Despite this variability, the frequency function is accepted as is shown in the lower plot.

PRIMARY SEGMENTATION

We seek first to partition the piece on the basis of its musical harmony. This gives us a guide as to where to look for harmonics. As mentioned before, this can be done using the optimum-comb as a periodicity detector.

Figure 45 shows the waveform of two violins playing simultaneously. One is playing B4 (494 Hz) and the other is playing F#4 (370 Hz). It is difficult to detect any periodicity in the waveform by direct observation. Figure 46 shows the output of the optimum-comb for the above mentioned waveform. We can see strong periodicity at about 4, 8, and 12 milliseconds. These correspond to about 250 Hz, 125 Hz, and 62.5 Hz. The F#4 is roughly the 3rd harmonic of the 8 millisecond period and the B4 is roughly the 4th harmonic of the 8 millisecond period. This shows that the periods detected by the optimum-comb are sufficient to assure that we can find the frequencies of all the harmonics present by taking multiples of the frequencies represented by those periods. The problem is that there are more periodicities found by the optimum comb than are actually needed for this task. Since there does not seem to be any good *a priori* way of eliminating the unnecessary ones, we must settle for doing more work than we have to. We can, however, notice that one period is a harmonic of other periods and is thus redundant. For instance, in the set 4, 8, and 12 milliseconds, 4 milliseconds is redundant and need not be included.

ON THE OPTIMUM-COMB

The first pass through the piece is a straightforward application of the optimum-comb periodicity detector. There is little of interest here except that there is a way to reduce the computation time. If the time step between applications is less than the summation interval, then the summation can be broken up into intervals whose length is just the time between applications. The total summation may be obtained by summing a number of these intervals, thus reducing the computation to a fixed amount, regardless of the total summation width.

To enhance the accuracy of locating the minimum, the four points around the minimum are used to generate a Lagrange polynomial which is then differentiated and the location of the minimum extracted. This allows us to get somewhat finer resolution than an integral number of samples would allow.

Consecutive minima which are very close in period are linked together into lists. Figure 12 shows these lists as determined for the first brass chorale in *Tableaux d'une Exposition*. The only special consideration here is that loss of a minimum at a single point is tolerated. A list remains continuous even though an application of the optimum-comb does not have a minimum at that period, but has one in the neighboring applications.

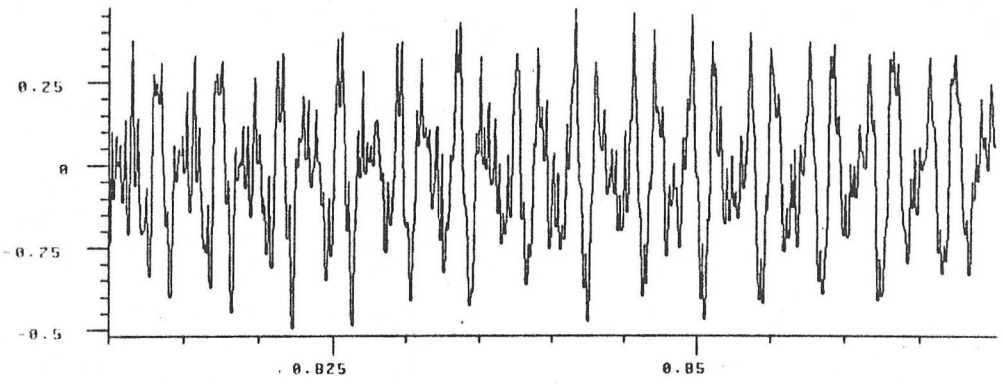


FIGURE 45. This is the waveform of a violin duet. One violin is playing a B4 (494 Hz) and the other is playing an F#4 (370 Hz). There is no periodicity evident to the unaided eye in the waveform.

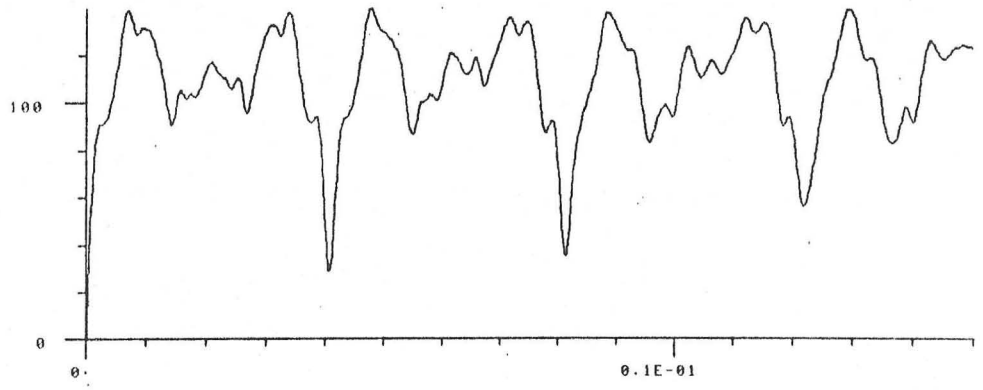


FIGURE 46. When the optimum-comb is applied to this waveform, it produces the above plot. We can clearly see the minima at about 4, 8, and 12 milliseconds. These correspond to 250 Hz, 125 Hz, and 62.5 Hz. The F#4 is roughly the 3rd harmonic of the 8 millisecond period and the B4 is roughly the 4th harmonic of the 8 millisecond period. The frequencies detected by the optimum-comb are generally sufficient to assure that all the harmonics of all the notes in the piece at that time are at frequencies which are multiples of those found by the optimum-comb. This is very important for planning at which frequencies the bandpass filters should be placed.

ON THE ESTIMATION OF ROOTS

These lists are then examined to generate regions. Each region is characterized by a number of "roots". A root is a frequency such that a number of the harmonics present in the region are integral multiples of the root frequency. Some number of roots will account for all the harmonics in a region. For N-voice pieces, only N roots at most are required. We cannot, however, tell on an *a priori* basis which roots form a complete set. We must settle for some duplication.

The first estimate of the regions is determined just by the beginning and ending times of the lists of minima. For each region, the minimum number of frequencies is determined which can produce all of the frequencies in the region. In other words, redundant harmonics are eliminated as candidates for the roots in a region. Adjacent regions are then merged if they contain the same roots.

The following is a table that presents the results so far for the first second of a two-violin piece. The first column gives the beginning time of the region, the second column gives the frequencies of the roots found in that region. The third column gives the frequencies of the notes that were sounding during that region, and the last column comments on the roots.

TIME (MS.)	ROOTS (HZ.)	NOTES (HZ.)	COMMENTS
0	1835	165, 196	11th harmonic of 165 Hz
10	189		Poor approximation to 196 Hz
20	32.1		5th subharmonic of 165, 6th subharmonic of 196
180	162.4		Poor approximation to 165 Hz
230	179.8	165, 185	Poor approximation to 185 Hz
240	20.1		8th subharmonic of 165, 9th subharmonic of 185
350	179.8		
390	186.8		Approximation to 185 Hz
400	186.8	262, 208	Leftover from last note
	272		Poor approximation to 262 Hz
410	186		
	200		Poor approximation to 208 Hz
	272		
430	200		
	272		
440	51		4th subharmonic of 208, 5th subharmonic of 262
450	42.6		5th subharmonic of 208
	51		
520	42.6		
730	48	262, 220	difference tone between 262 and 220
	63.7		4th subharmonic of 262
	85.7		3rd subharmonic of 262
850	24	262, 196	8th subharmonic of 196, 11th subharmonic of 262
	63.7		
	85.7		

From this table, it should be clear that the roots determined by this process are not entirely reliable. The problem is that there is no way to judge the quality of a minimum produced by the optimum-comb method. The exact depth of the minimum is highly variable from application to application, depending on the exact amplitudes of the notes involved. The period estimates do not vary appreciably from application to application. Since we cannot tell whether a particular periodicity estimate is better than any other, there is no way to eliminate the less useful root estimates. To make sure that no tones are lost, root estimates for adjacent regions must be merged before planning the filter frequencies.

BANDPASS FILTERING

ON LOCATING CENTER FREQUENCIES

First, we must determine at what frequencies to apply the filters. This comes from examining the estimates of the roots of each region of the piece. The only measure of quality of the root estimates is the length of a region. A long region means that these roots were present for a long time. This is evidence that they are not transient phenomena. Based on this observation, we form macro-regions by starting with the widest regions and grow outward by absorbing adjacent regions until the entire piece is covered. Because of memory limitations, we cannot handle more than .5 seconds of sound at a time in the filter routines, thus we cease growing a region when it approaches .5 seconds in length.

To some extent, the procedure described above is an *ad hoc* one. This is because there does not seem to be, at this time, anything better to be done. Since the purpose of locating the roots of the regions is to reduce the number of filtering operations over what would be required for a dense covering, it is not damaging that we include spurious roots. This just means that we will not realize the minimum number of filtering operations. In every case examined so far, some savings have been realized, so the procedure seems worthwhile. The average savings seems to be roughly a factor of three over the dense covering.

Once the macro-regions are defined and the roots determined, a list is made of all the harmonics of each root up to some maximum frequency. This maximum could have been set as high as the Nyquist rate, but was arbitrarily set to include up to the 5th harmonic of the highest note in the piece under analysis. This maximum frequency setting does not affect the analysis, providing it is set high enough, so that setting it any higher simply wastes time without adding to the quality of the analysis.

This list of candidate center frequencies is examined for redundant entries. An entry is redundant if it is within the passband of a filter set at an adjacent frequency. This reduced list is then taken as the final list of center frequencies.

ON FILTER PARAMETERS

A bandpass filter is defined by many parameters. For communication value, we use traditional filter types: Chebychev, Butterworth, etc [Guilliman 1957; Karni 1966], transformed to the discrete domain by use of the bilinear transform [Gold and Rader 1969]. The resulting filters have infinite length impulse responses. The filter coefficients are determined by a program which takes the filter specifications and computes the coefficients (see Appendix B). In selecting a filter type and parameters, the considerations are as follows:

1 - What is the band width? A bandpass filter attenuates frequencies outside of its passband. We determine the band width by choosing two frequencies which represent the endpoints of the passband.

2 - What is the attenuation outside of the passband? This determines the *order* of the filter. The order of a filter is an integer. It determines how many natural frequencies the filter has. Outside of the passband, the frequency response (before transformation to the discrete domain!) drops off roughly 20 dB per decade (factor of 10 in frequency) for each order. Since a bandpass filter has two skirts (places where the response drops off sharply), the effect is halved. That is, increasing the order by 2 causes an increase of the attenuation rate of 20 dB per decade on *both* sides of the passband.

3 - How close to constant is the response in the passband? This determines how accurate the harmonic amplitudes will be as they emerge from the filter.

The relations among these parameters are complex. Generally, it works like this: the transient response is directly related to the band width. It is secondarily related to the attenuation rate. The more narrow the band, and the faster the falloff, the longer the transient response. There is a tradeoff between constancy in the passband and the attenuation rate. In the Chebychev filters especially, there is a direct relation. The more ripple (distortion) you allow in the passband, the greater the attenuation rate.

Making a choice of exactly the parameters to use is an exercise in whim, since there is generally no "optimum" setting. When thinking about musical sound, we might conclude that since harmonics are linearly spaced in frequency, a linear frequency scale is what is called for, that we should maintain a constant bandwidth throughout the frequency range, and that center frequencies should be placed at uniform intervals. Linear distance, however, on a piano keyboard reaches frequencies that increase exponentially. This might lead one to think that the bandwidth could be wider for higher frequencies because the spacing of musical notes gets wider with frequency. The ear is physically set up on a scale that is somewhat between linear and exponential, and since we are mimicing the ear's performance, we perhaps should take advantage of the experimentation that nature has done for us. Figure 47 shows the relation between distance along the basilar membrane (corresponds to filter bandwidth) with frequency. It is clear that this relation is not simply logarithmic or simply linear. The vertical axis on the plot represents what is called "tonalness" (a poor translation from the German) and is measured in "Barks", after the great researcher Barkhaus. Tonalness represents critical bandwidths in the ear. If we think of the ear as a band of bandpass filters, a critical band is analogous to the bandwidth of the filter. For instance, two sinusoids will sound rough if their frequency separation is smaller than a critical bandwidth, and will sound smooth for frequency separations wider than a critical bandwidth. A difference of 1.0 on the tonalness scale represents one critical bandwidth. This corresponds to equal lengths along the basilar membrane.

However, the program currently uses a linear frequency scale. The bandwidth is set to a constant 20 Hz throughout the range, which extends from about 80 Hz to about 5000 Hz. It would be very interesting to use the biological model and see if good results were obtained and time was saved. This experiment is deferred for the time being.

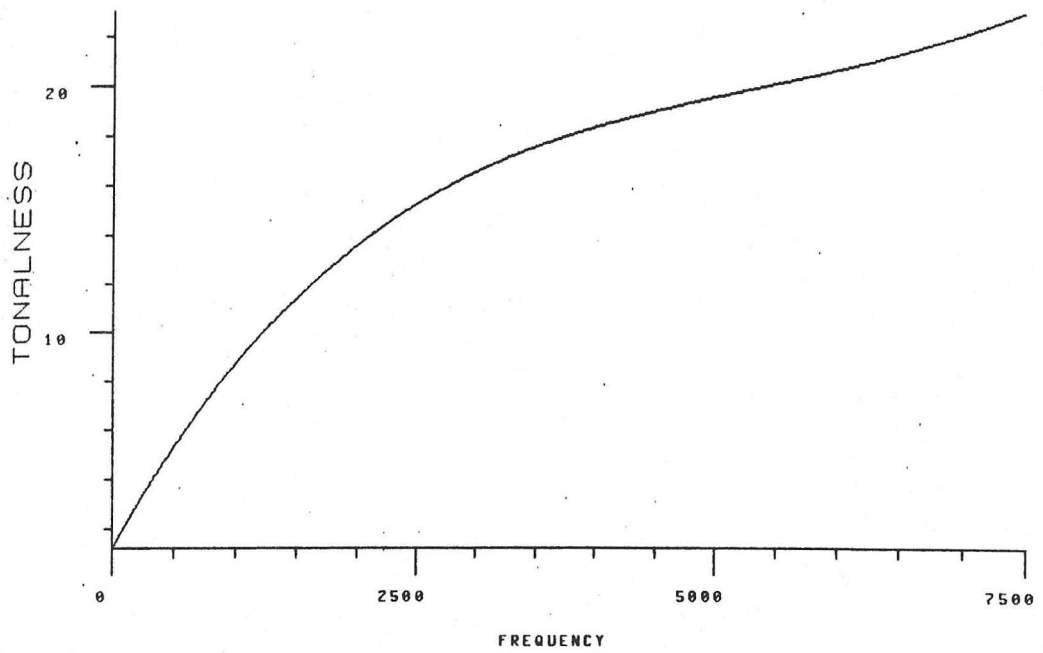


FIGURE 47. This is a plot of length along the basilar membrane versus frequency (after Zwicker). The vertical axis label is called "tonalness" and is measured in "barks" (after Barkhaus). One bark corresponds to one critical bandwidth. Thus this curve gives us the frequency resolution of the ear. Note that a critical bandwidth is not the inherent bandwidth of the hairs along the basilar membrane, but is a much more narrow bandwidth which is hypothesized to be a consequence of the neural interconnections of the hairs. The point is that the curve is neither exponential (like the piano keyboard) nor linear (like harmonics) but is something in between. The greatest slope is below 500 Hz and represents the greatest resolution. Most of the lower partials of musical sound can be independently discriminated. Generally, it is thought that "dissonance" occurs when more than one partial falls within a single critical bandwidth. This plot is suggested as a possible guide for placement of bandpass filter frequencies for a dense covering of the frequency spectrum.

The program uses a 4th order Chebychev filter with a 3dB passband ripple. If it were being done again, we feel that less ripple is in order. The ripple caused certain harmonic amplitudes to be estimated incorrectly. Figure 48 shows the impulse response and the frequency response for this kind of filter when centered around 100 Hz.

The impulse response associated with a 20 Hz bandwidth is quite long, as can be seen from the figure. With some of the higher harmonics, where the activity is quite weak, considerable transient response was excited. The use of wider bands, as suggested by the physical model and the exponential models mentioned above, would help alleviate this problem.

ON PROCESSING FILTER OUTPUT

The output of each filter is sent to an optimum-comb pitch detector. The detector searches for frequencies within the passband of the filter. It is applied every 2.5 milliseconds throughout the macro-region. The output of the pitch detector at each application is a list of the frequencies where minima in the comb output were found. Again, polynomial interpolation is used to locate the minima more accurately. This is essential. At 5 Khz, for instance, at 50 KHz sampling rate, the period is only 10 samples long. A shift of one-half sample is equivalent to a frequency change of about 250 Hz. Interpolation, then, is essential for the higher harmonics. Each such frequency is compared with the previous application. Frequencies whose periods are within 2 samples are considered for linking. Each frequency is linked to its best match from the previous application. These links produce lists of minima.

After all the lists have been formed in this macro-region, a "weakest boundary first" merging algorithm [Yakimovsky 1973] is used to link adjacent lists whose average periods are very close. This merging algorithm is used because each time two lists are merged, the resulting list has in general a different average period, so that it must be compared again with its neighbors. Each time two lists are merged, the boundary between them is deleted and the "scores" (magnitude difference between the average periods of the lists) of the two remaining boundaries are recomputed based on the new composite average period for the list. We cannot just merge lists which have scores better than some threshold without recomputing the averages. This could allow glissandi, which would have small *local* changes in frequency but large *global* changes.

This procedure is sensible because we know that the frequencies present in the music change slowly and smoothly, so we can be sure that minima whose frequencies are very close are quite likely to belong to the same harmonic. Since we know that the frequencies of notes, and thus their harmonics, are *nearly piecewise-constant*, we can eliminate glissandi, and certain noise traces which appear to have swiftly-changing frequencies.

With the lists that remain, some simple tests to eliminate noise traces are done. A list whose total deviation (maximum frequency in the list minus the minimum frequency) is too large is eliminated. Lists whose frequencies change too rapidly (has too great a slope) are eliminated.

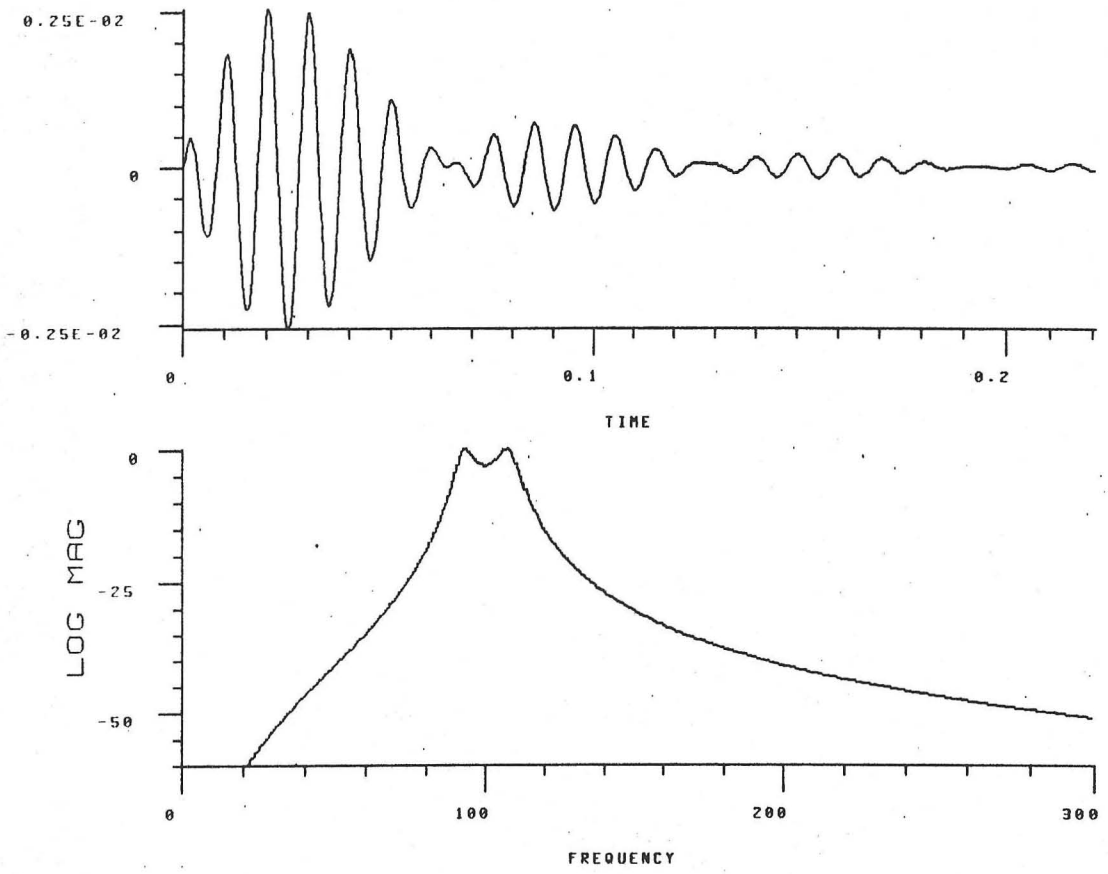


FIGURE 48. Impulse and frequency responses of the bandpass filters that were used for the harmonic extraction. The bandwidth is about 20 Hz. This filter is centered on 100 Hz. The filter was made by first designing a 2nd order Chebychev low-pass filter with 3 dB of ripple, transforming it to a 4th order bandpass filter (all in the continuous domain), then transforming to the discrete domain via the bilinear transform. Of course, the 3 dB points had to be mapped first to assure the correct cutoffs after transformation. The advantage of designing the filter in this manner is that it is a closed form solution (no iteration) and thus can be programmed very efficiently. It takes only a few milliseconds on the computer to set up the coefficients for a filter of arbitrary ripple and cutoff frequencies. If we were to attempt the task again, a filter with less passband ripple would be preferred. The passband attenuation sometimes reduced the amplitude of a good harmonic to the point that it could not be distinguished from a noise trace.

As was mentioned before, the optimum-comb pitch detector (and, in fact, all autocorrelation-type pitch detectors) responds as well to subharmonics of a frequency as to the frequency itself. We must have a way to eliminate these subharmonics. This is done by applying a crude pitch detector which does not have this problem and comparing the results. The pitch detector used is just the length of the list in time divided by half the number of zero crossings in that interval. This gives an order-of-magnitude pitch estimate which is then used to eliminate lists corresponding to subharmonics.

INTERMEDIATE-LEVEL TECHNIQUES

INTRODUCTION

At this point of the analysis, we are presented with a list of sinusoids that are present in the original sound. We have their amplitudes and frequencies as functions of time. The purpose of the intermediate-level programs is to infer from these data what notes are present, their frequencies and their extent in time.

At this level, we must also eliminate information that is not strictly erroneous, but nonetheless is not desired. One example of this is found in string instruments. When a musician plays a string instrument, like violin or guitar, the strings other than the ones being manipulated also sound. It would be extremely difficult for a musician to damp the other strings all the time. It is not common practice to do so on stringed instruments except in some schools of classical guitar. The resonances of the other strings are usually 15 dB or more softer than the principal sounds, so they are generally not heard unless one listens very carefully. Our program, however, picks these extraneous tones out quite nicely. They appear in all the output. Rather than report exactly what is present, we wish to mimic human behavior and suppress these tones that do not have immediate musical meaning. Other extraneous sounds include box resonances (stringed instruments, for instance, have very strong box resonances), and strings that continue to vibrate past the intended ending of the note (common with open strings).

In the following sections, we describe the processes as they roughly correspond to separate programs in the processing path. First is segmentation and scoring. The scoring is the key to this entire section. Without rating the output of the low-level processes as to quality and suitability, no cogent decisions as to what notes are present could be made. With these ratings, the notes can be inferred by accumulating groups of high-quality harmonics without combinatoric searches. After the notes are derived, we proceed to separating the notes into the upper and lower voices. This is done using the assumption that the piece has no more than two voices at any given time. Finally, the output is prepared for the manuscripting program. This involves some cleverness to assure good readability.

HARMONIC PROCESSING

SEGMENTATION AND SCORING

INTRODUCTION

From bandpass filtering and pitch detection, we get rough traces of the amplitude and frequency contours for each harmonic present in the piece. The problems are many. First, any given trace may not include the full duration of a note. This is because of space limitations in the filtering program. The signal must be broken up at arbitrary places and processed in pieces. These pieces must be glued back together later. Second, any given trace may include more than one note, one after another. This is because the transient response of the filter may continue to ring after a harmonic disappears. It can be excited by activity elsewhere in the spectrum. This can continue indefinitely, or another harmonic of similar frequency may be picked up. Third, poor traces are caused not only by weak signals, such as extraneous resonances or high harmonics, but can also be caused by having the center frequency of the filter be offset from the actual frequency of the harmonic. In fact, there are usually 3 traces for each harmonic: one right on the frequency, one above, and one below.

From this, we can see that the first thing that must be done is to break up the traces into units that we know contain no more than one harmonic of one note, if they contain anything meaningful at all. The next thing that must be done is to produce a score for the trace which reflects its "quality" in some way. We must decide what "quality" means in this context. Gluing together component pieces of a long note can be done later.

SEGMENTATION AND SCORING

The segmentation is actually the easiest part of the processing. Here, we simply determine the threshold on the amplitude function such that 90 percent of the energy in the harmonic is at amplitudes above this threshold. The amplitude function is then scanned for regions that exceed this threshold. Segments that are shorter than 35 milliseconds are assumed to be unimportant and are discarded. This is based on the fact that most meaningful musical notes are longer than 100 milliseconds. Occasional grace notes and trills will involve notes as short as 50 milliseconds. Our programs are set up (from this point on) to favor notes of duration 80 milliseconds or longer. This number is a compromise with the desire to include meaningful musical notes and the desire to eliminate noise traces. We must set the threshold on length long enough to eliminate as much spurious transient response of the bandpass filters as possible. We include harmonics at this point of durations 35 to 80 milliseconds because they may get merged into a longer note subsequently.

Before we proceed further, let me point out an ambiguity of terminology. When a piece of music is written down in traditional music notation, the resulting document is called a *score*. Alternately, when we rate an entity by assigning it a number which reflects its quality, this number is often called a *score*. We hope the context will distinguish these meanings clearly. In this section, we are interested in assigning a quality measure to the traces, so it is the second meaning that is relevant here.

The scoring of a harmonic is the most important process because it is the only clue as to the viability of a note that is assembled from a group of harmonics. As an example of how much data is assembled, a single 2-bar piece that was processed contained 27 notes, or about 150 meaningful harmonics (about 5 harmonics per note). The output of the bandpass filtering and pitch detection produced about 2000 amplitude-frequency traces. That means that over 90 percent of the traces produced by the filtering and pitch detection must be discarded. The traces come from multiple detections of single harmonics, and traces of transient responses and noise patterns in the high-frequency ranges. The score must reflect the likelihood that a given harmonic is real and not just a noise trace.

The criterion we have chosen is smoothness of the curves. We require the amplitude curve to correspond well to a low-order polynomial (6th order or so), and we require the frequency to be nearly constant. Since the sluggishness of the bandpass filter smooths out any fine detail in the harmonic, this is a reasonable consideration. Strong, valid harmonics tend to have clean, smooth traces and nice even frequencies. Vibrato can cause the frequency to be non-constant. Rather than deal with this aspect now, we have finessed the problem by not considering it. Any more comprehensive musical scribe should allow certain forms of frequency variation like vibrato, glissando, and expressive frequency changes.

We produce a composite score for the trace by taking into account the residual error of the

amplitude and frequency fits as well as the coefficients of the frequency fit. This not only gives a measure of the quality of the fit, but also a measure of the constancy of the frequency during the note. We also use the distance between the center frequency of the filter and the frequency of the harmonic. Since the traces are better as they approach the center frequency, because they are maximally distant from the high-Q resonances of the filter, this is a reasonable measure to help discriminate good traces from transient response. Each of these measures must be made commensurate with one another. For instance, the coefficient of the second degree term of the frequency polynomial is a squared quantity and its square root must be taken.

One of the bigger problems in normalization of the components of the score is equalization for duration. We want scores for long notes to be commensurate with scores for short notes. The terms in question here are the residual errors for the polynomial fits. If we view the fit as a regression process, then the residual error will be distributed as X^2. To show this and the assumptions it involves, let us show where this result comes from. This presentation is patterned after Freund [1962]. Since this is a standard derivation, we shall only present the results, not the intervening steps.

Given a sequence of abscissa, X_i, and their ordinates, Y_i, representing, in this case, equally spaced points in time and the value of the amplitude or frequency curve at that point in time, we can fit a polynomial to Y as a function of X and use its residual error as a measure of the quality of the fit. We assume, then, that the Y_i are *independent* random variables having the following conditional probability distribution:

$$(37)\quad \phi(Y_i|X_i) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{\frac{1}{2\sigma^2}\left(Y_i-\sum_{j=0}^{N} a_j X_i^j\right)^2}$$

Where X_i are the *independent* variable, $1 \le i \le M$

Y_i are the *dependent* variable, but are *independent* random variables distributed normally about an N^{th}-order polynomial.

σ is the standard deviation of said distribution.

a_j are the coefficients of said polynomial.

Here, the a_j are the same for each value of i. We obtain maximum likelihood estimates of the regression coefficients, a_j, and then compute the residual error as follows:

$$(38)\quad \rho^2 = \frac{1}{M}\sum_{i=1}^{M}\left(Y_i-\sum_{j=0}^{N} a_j X_i^j\right)^2$$

Where ρ is the root-mean-square residual error of the abscissa and the polynomial

ρ^2 will then be an estimator of σ^2 and is thus distributed as χ^2. The main assumption here is that the ordinates are distributed normally around a polynomial. This is, of course, not entirely true. There is nothing in the physics of music production that requires the harmonic amplitudes to be polynomials. We violate the assumption with the hope that the resulting computations will still be meaningful.

The use of the χ^2 property of the residual error is that traces of different lengths (different values of M, i.e., different numbers of degrees of freedom) can be compared by first normalizing by the χ^2 value for that number of degrees of freedom. In fact, we find that this does help produce more commensurate residual errors between long segments and short segments, but due to the fact that the assumptions fundamental in the process are violated, the correction does not seem to be enough. Long segments still have somewhat higher residual errors than short ones.

To be explicit, the score, representing the "badness" of the trace (that is, inverse quality) is computed as the sum of the following terms:

$\mathfrak{E}_1$ - The quotient of the residual error of the amplitude fit, as defined in equation (38), and the average amplitude of the harmonic. The residual error of the amplitude fit was normalized by the χ^2 value for the number of degrees of freedom (points) in the amplitude function that were used in making the polynomial fit.

$\mathfrak{E}_2$ - The quotient of the residual error of the frequency fit and the average frequency of the harmonic. The residual error is again normalized by the χ^2 value.

$\mathfrak{E}_3$ - The first-order coefficient of the frequency fit, divided by the average frequency of the harmonic.

$\mathfrak{E}_4$ - The square root of the second-order coefficient of the frequency fit, again divided by the average frequency of the harmonic.

$\mathfrak{E}_5$ - The magnitude of the difference of the average frequency and the center frequency of the filter.

The total score was then computed as the weighted sum of these terms:

$$(39)\quad \mathfrak{E} = k_1\mathfrak{E}_1 + k_2\mathfrak{E}_2 + k_3\mathfrak{E}_3 + k_4\mathfrak{E}_4 + k_5\mathfrak{E}_5$$

Where the k_i are the weightings of the various error terms

The first four terms, $\mathfrak{E}_1$ through $\mathfrak{E}_4$, were normalized by the average value (amplitude or frequency) of the harmonic. This gives a measure of the *relative* error rather than the *absolute* error. This allows us to compare strong harmonics with weak, high frequency harmonics with low frequency ones. Otherwise, the expected error range would vary with these parameters.

In $\mathfrak{E}_4$, the square root was taken because the second-order coefficient is a *squared* quantity. The root must be taken to make it commensurate with the other error measures, which are all *linear* quantities.

For reference, the values for the weights, k_i, were $k_1=100$, $k_2=3000$, $k_3=10$, $k_4=20$, $k_5=4$.

Figure 49 shows four examples of segmentation, polynomial fitting, and scoring on a single harmonic. The harmonic chosen is the second harmonic of a 262 Hz note (C4), which is at about 525 Hz. These traces are taken from the first notes of a two-part piano piece. There is also a 332 Hz note (E4) sounding at this time. The four traces in the figure are separate traces of the same harmonic. This shows how adjacent filters will pass the same harmonic with differing degrees of faithfulness. Over each figure is a list of parameters: CF represents the center frequency of the filter that produced the trace. In each figure, the upper plot represents the amplitude envelope of the filter output. The bottom plot represents the output of the pitch detector which was applied to the filter output. Across the amplitude plot is a horizontal line which represents the threshold such that 95 percent of the energy in the amplitude envelope is at values above that threshold. This is how the segmentation is done. The small arrows point out the limits of the region above threshold that is being processed. Sometimes a single trace will have several disjoint traces above the threshold. The next figure shows such an example. Both the amplitude and frequency functions were fit with polynomials. The polynomials are also plotted. They are the smooth lines through the plots. The amplitude polynomial is of order 6, and the frequency polynomial is of order 2.

Above each figure is listed the contributions to the total score from each of the five error functions. The label CONT1 on the figure refers to the weighted, normalized quantity $k_1\mathfrak{E}_1$. The label CONT2 refers to the weighted, normalized quantity $k_2\mathfrak{E}_2$, and so on. The total score, which is the sum of these contributions as expressed in equation (39), is labeled SCORE in the figures. The parameter AVFR is the average frequency in the region under analysis.

As we can see, the error score decreases monotonically as the center frequency of the filter approaches the actual frequency of the harmonic, even if we discard the contribution from $k_5\mathfrak{E}_5$, which represents exactly the distance from the frequency of the harmonic. The contribution from $k_5\mathfrak{E}_5$ is included to strengthen this bias toward centered filters. Remember that the frequencies of the filters was determined by the comb filter, so that they do not necessarily represent the frequency of the harmonic that passes through the filter. We include this last term to represent only the fact that the trace is better when the frequency of the harmonic is near the center frequency of the filter, and thus the overall score for the harmonic is more likely to be meaningful.

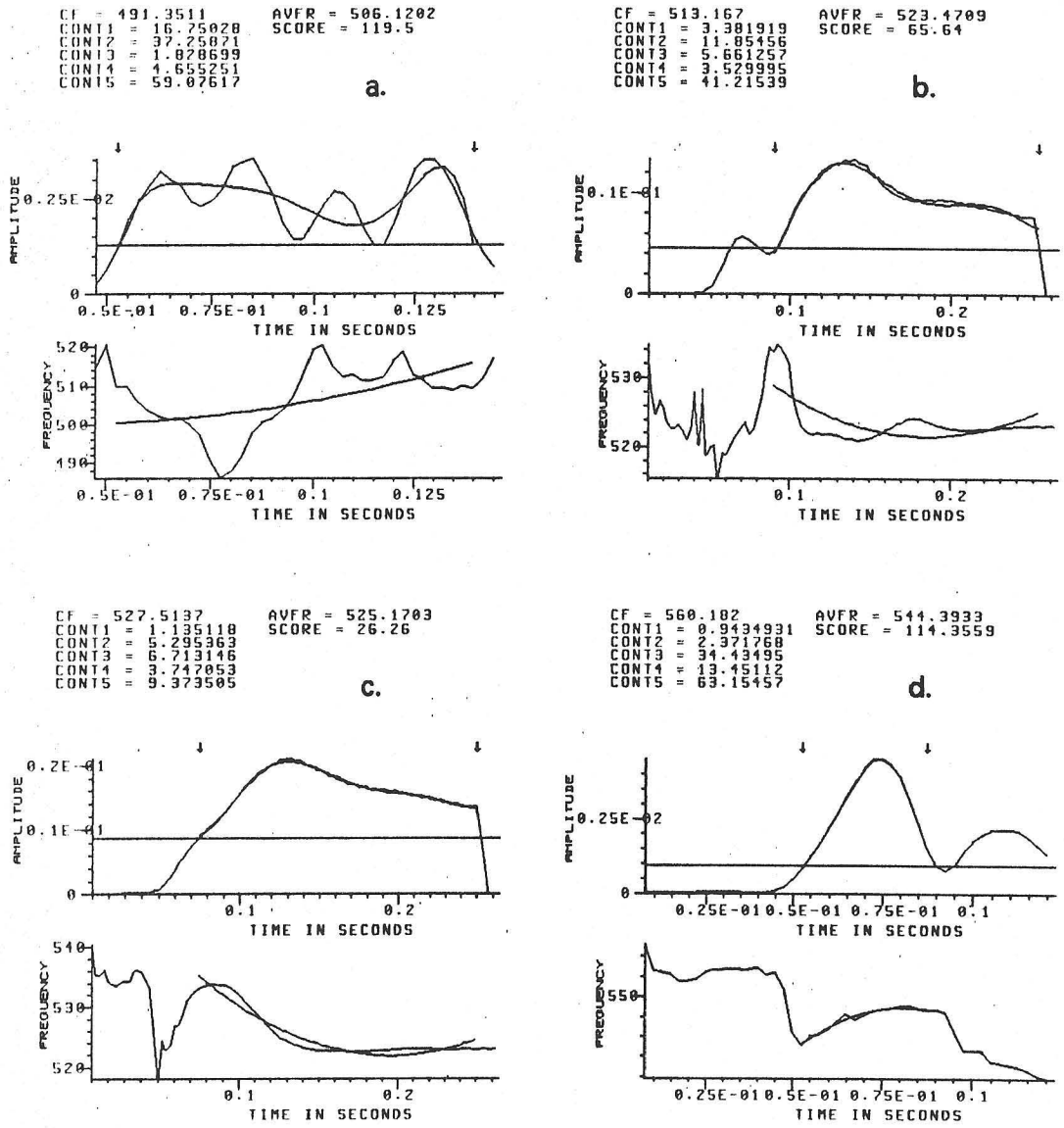


FIGURE 49: Plots from the segmentation and scoring algorithm. Each picture shows an amplitude and a frequency curve. The horizontal line across the amplitude plot denotes the threshold where 95% of the energy of the plot lies at amplitudes above this line. The small arrows denote the region being fit and scored. The smooth curves through the amplitude and frequency plots are the polynomial fits to these curves. In figure 49c, the polynomial fit for the frequency rises at the end of the plot. This is a boundary effect common in this kind of approximation that the slope of the approximation strays at the ends of the window. The numbers at the top represent the various scoring contributions, already weighted and normalized, as described in the text. CF represents the center frequency of the filter that produced these plots, AVFR represents the average frequency in the region being fit, and SCORE represents the sum of the contributions from all five error sources. These traces were taken from the analysis of a two-part piano piece. There was a 262 Hz note and a 332 Hz note being played at this time. We see four traces of the same harmonic: the second harmonic of the 262 Hz note, at about 525 Hz. It is clear that the score improves (gets smaller) as the center frequency of the filter (CF) approaches the actual frequency of the harmonic. This is a good demonstration of why a scoring system is necessary. Each harmonic produces many traces. The good ones must be separated from the spurious ones. The error criteria used here seems to accomplish this effectively.

In figure 50 we see four more plots, again of the same harmonic, which is the third harmonic of the 332 Hz note (E4) at about 987 Hz. Since the strengths of the harmonics generally decrease as the harmonic number increases, these upper harmonics become increasingly difficult to follow. Often, even when the filter is exactly centered on the harmonic a good trace with low error cannot be obtained. As a result, these upper harmonics cannot be used with great confidence to infer the existence of notes.

Figure 50a and 50b show how a single harmonic can get spuriously broken into two pieces. Here the harmonic was beating with the transient response of the filter and went below the segmentation threshold and was thus broken up.

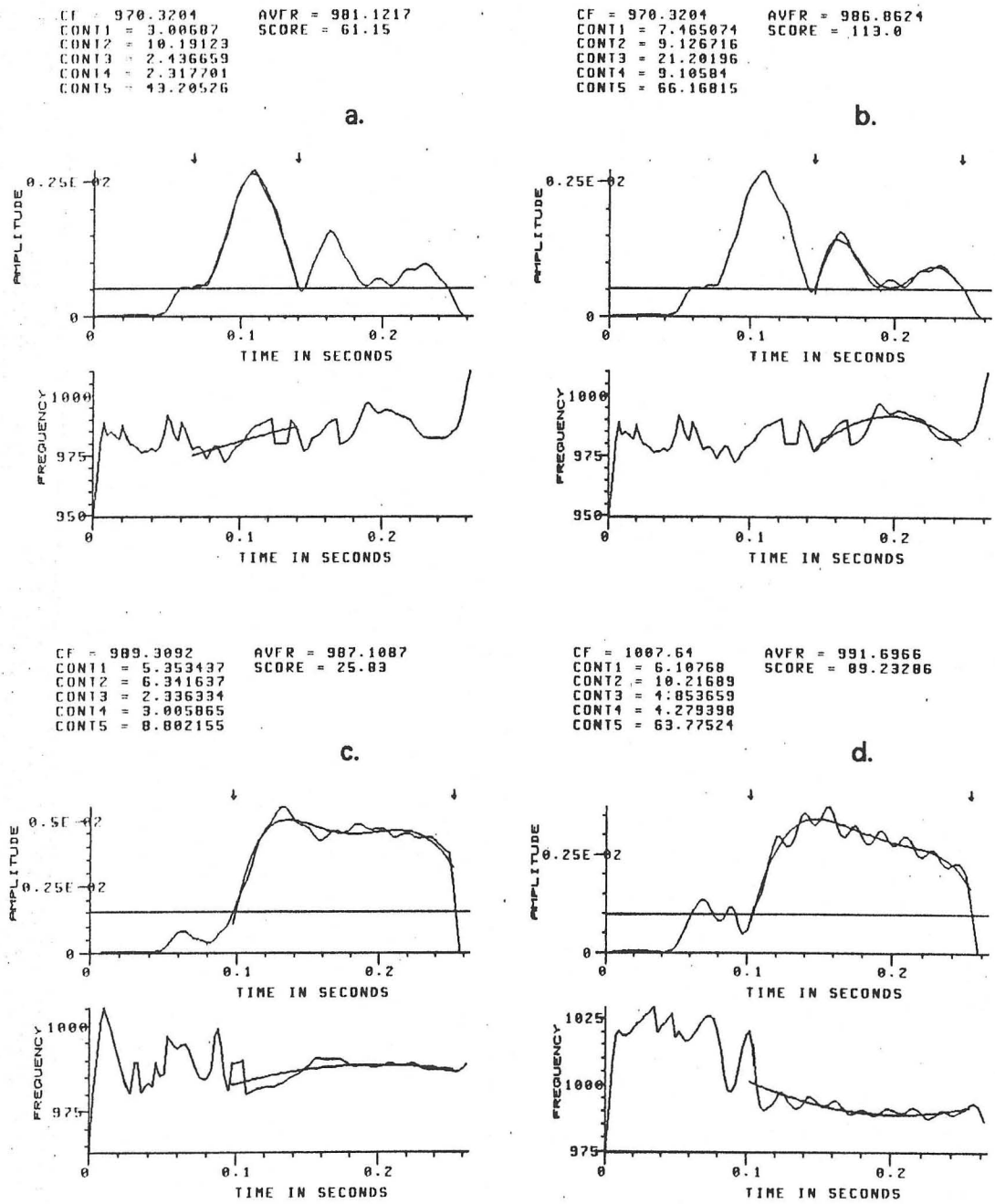


FIGURE 50: Plots from the segmentation and scoring algorithm. As with the previous figure, these traces were taken from the analysis of a two-part piano piece. There was a 262 Hz note and a 332 Hz note being played at this time. We see four traces of the same harmonic: the third harmonic of the 332 Hz note, at about 987 Hz. As we ascend in harmonic number, the traces get weak and noisy, such that there are many spurious traces, and high error scores on the good traces. For this reason, we cannot rely on the higher harmonics as evidence for notes except for certain instruments with especially strong high harmonics.

INFERRING THE NOTES

At this point in the analysis, we have a large set of possible harmonics. For each possible harmonic, we preserve only a few numbers: the average amplitude, the beginning time, the ending time, the average frequency, the error score, and the amplitude function polynomial. All information regarding the exact shape of the amplitude or frequency function has been discarded.

It seems to be a property of machine perception programs that they get more and more heuristic and less and less defensible on theoretical bases as they proceed to higher and higher levels of processing, away from the low-level, signal-processing techniques. This program is no exception. Each heuristic is based in the properties of musical sound, but sometimes the connection is especially tenuous.

Our first task is to merge duplicate traces. Since we get several traces for each harmonic, we can combine these into one composite harmonic. This reduces the data immediately by a factor of three or so. This initial merging is only done for traces that overlap significantly in time and whose pitches are within a few percent of one another. We call these *reduced* harmonics. The parameters of the reduced harmonic are taken from the parameters of the harmonic with the lowest error score. In the case of several harmonics with low scores, a weighted average is taken to form the new amplitude and frequency. The parameters are weighted by the reciprocals of the scores of the individual harmonics.

Next, a list is formed of these reduced harmonics in order of their average amplitude divided by their error score. This provides simultaneously a measure of the strength and the quality of the reduced harmonic. We then attempt to group together a number of harmonics that infer a note. One problem in so doing is avoiding a combinatoric search. Assuming that the lower-level procedures have produced faithful traces, we can just pick off the best reduced harmonic (in the sense of having the largest amplitude-error score quotient) and assume that this is the first, second or third harmonic of a note. This is a purely heuristic assumption but it is based on the observation that most musically interesting tones have strong lower harmonics. This does not account for many effects present in human hearing, like the existence of residue pitch, but it is a reasonable compromise for the current study.

With this reduced harmonic, we first search the entire reduced harmonic list to see if there is another reduced harmonic existing at the same time that has one-half or one-third of the frequency. If there is no such tone, we take our original reduced harmonic to be the fundamental of the note, else we take the lowest reduced harmonic found as the fundamental. We can then race through and pick out harmonics for this fundamental just by locating reduced harmonics that exist at the same time and which have frequencies that are close to the predicted frequency of the harmonic in question.

Once the harmonics are selected, the note can be tested for viability. The first test is whether

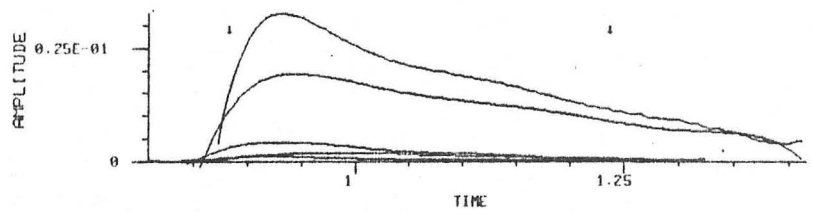
the fundamental is at all strong. We require the fundamental to be of substantial strength and quality. This is, again, a departure from human perceptual performance. If the fundamental is strong, we examine the strengths of the harmonics that are not multiples of two and not multiples of three. The 1st, 3rd, 5th, and 7th are examples of harmonics that are not multiples of two. The 1st, 2nd, 4th, 5th, 7th, and 8th harmonics are examples of harmonics that are not multiples of three. This is to try to determine whether the fundamental is a spurious trace and the note is really two or three times higher than we are hypothesizing. We threshold the ratio of the sums of the qualities for these selected harmonics with the sum of the quality for the remaining harmonics. This seems to be an adequate technique, although it occasionally eliminates useful notes.

We require also that the harmonics be *dense*. That is, for two or more harmonics, we require that the note possess all but one harmonic for acceptance, unless it is only odd harmonics, in which case it must possess all the odd harmonics up to the highest harmonic in the hypothesized note. A note consisting of just one harmonic, the fundamental, we require to be quite strong for acceptance.

We then merge notes that have very nearly the same frequency and overlap considerably in time. These can be produced by having a very long note. The initial segmentation based upon the musical harmony of the piece is made, some errors in segmentation result. The most common form of this kind of error is that a long note can get broken into smaller pieces. These pieces must be glued back together at some point. We have chosen to do so after the note hypothesis has been formed.

The data representing the note is then reduced to just four numbers: the pitch, the beginning time, the ending time, and the quality (amplitude over error score). The beginning and ending times are obtained by producing an overall amplitude profile for the note based on the polynomial representations of the amplitude curves for each of the harmonics. This overall profile is subjected to a threshold that assures that 95 percent of the energy is above the threshold. The times where the profile drops below this threshold are taken to be the beginning and ending times of the note.

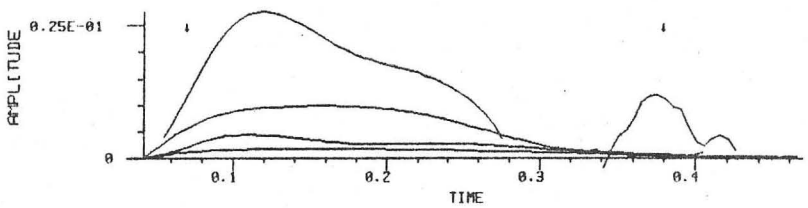
Figure 51 shows a representation of one of the notes inferred by this procedure. The curves on the plot represent the amplitude polynomials for each of the harmonics. The text in the lower part of the picture represents information on each of the harmonics. The first column is the beginning time, the second column is the ending time of the harmonic. These times are in tens of milliseconds. The next column is the average amplitude of the harmonic. The fourth column is the error score of the harmonic. Sometimes there is not a space between the figures in the third and fourth columns. The last column represents the average frequency of the harmonic. The isolated pitch figure at the bottom of the plot represents the weighted average pitch of the tone, which is derived by dividing down the average pitches of the harmonics, weighting them with the quality of the reduced harmonic, and averaging them.



SP	EP	AMP	SCR	PIT
87	141	324014	240	
85	141	20269	493	
80	132	192	47	741
84	123	438	21	984
87	127	109	44	1231

Trimmed notes PITCH IS 246.6307

FIGURE 51. A representation of a note hypothesis from a two-part piano piece. The curves in the upper figure are the polynomial approximations to the harmonic amplitude curves. Each curve represents the amplitude as a function of time of one harmonic. The numbers below the plot give the details of each harmonic. The first column is the starting time in hundredths of a second. The second column is the ending time. The third column is the amplitude of the harmonic (times 100,000), the fourth is the error score. The last column is the average frequency of the harmonic. There are two small arrows above the curve which delimit the region where 90% of the energy lies.



SP	EP	AMP	SCR	PIT
5	27	274818	331	
4	33	980	18	658
4	40	447	29	990
4	46	176	71	1329
34	42	1163166	336	
35	45	47	88	983

Trimmed notes PITCH IS 330.6625

FIGURE 52. This plot is like the above one, but points out that even at this late stage of the processing, noise traces can still be present in the data. Here, a bit of transient response from the following note overlapped this note enough to be absorbed as part of it. The noisy harmonic has a higher error score than the others, but it also has a very high amplitude, so it is not clear on what basis it can be eliminated.

Even at this late stage, imprecisions occur. Figure 52 shows one such error. There is a strong noise burst on the end of one of the harmonics. This burst is enough to cause the ending time of the note to be overestimated.

DERIVING THE MELODIES

Given this list of notes from the previous processing stage, we must now link them into melodies. For convenience, we do not attempt to handle the case where parts cross. To handle crossing parts correctly, we would have to identify the instrument involved, as well as examine the musical context in great detail.

We have decided upon a very simple algorithm for selecting melodic groupings. At this point in the algorithm, we make use of the assumption that there are no more than two independent voices in the piece. This way we can search for places where there are two notes sounding simultaneously and identify the voices positively. Any place that can be so identified is called an *island*. This island represents a place where there is no doubt as to the voices (upper or lower) a particular pair of notes belong to.

To finish the assignment, we use a global scoring algorithm. We assign a "score" to a particular assignment which is the sum of the magnitudes of the differences of the frequencies of adjacent notes in the melodies. We can then search all possible assignments of the unassigned notes and compare the various possibilities by comparing their scores. The assignment with the best (lowest) score is chosen.

Figure 53 shows the initial melodic assignment for a guitar duet. The score for the duet is shown in figure 60. What we see in this figure are the assignments based on the existence of islands in the piece. Each note is represented by a horizontal line. When a note is assigned to a voice, a "tail" is drawn at the end of the line which points up, denoting membership in the upper voice, or down, denoting membership in the lower voice. The dotted lines represent melodic connections made between notes which indicate melodic adjacency. In this figure, there are 8 unassigned notes.

With a small number of notes and a branching factor of two, it is reasonable to do an exhaustive search to determine the best melodic assignment. For this to be practical, the algorithm which determines the melodies once the notes are assigned to the voices must be fast. Fortunately, this can be done in a very simple manner. With the voices already assigned, we merely start at the beginning of the piece. We locate the first notes in the piece in each voice. We then locate the second notes in each voice simply by searching forward in time. We proceed through time in this manner, annexing notes onto their respective voices, until we exhaust the notes in the piece. This assignment is linear and can be made very fast by sorting the notes into time order. This sort only has to be done once.

Figure 54 shows the results of the melodic grouping for the guitar duet. Figures 55 and 56 show the same plots for the pseudo-violin duet whose score is shown in figure 58.

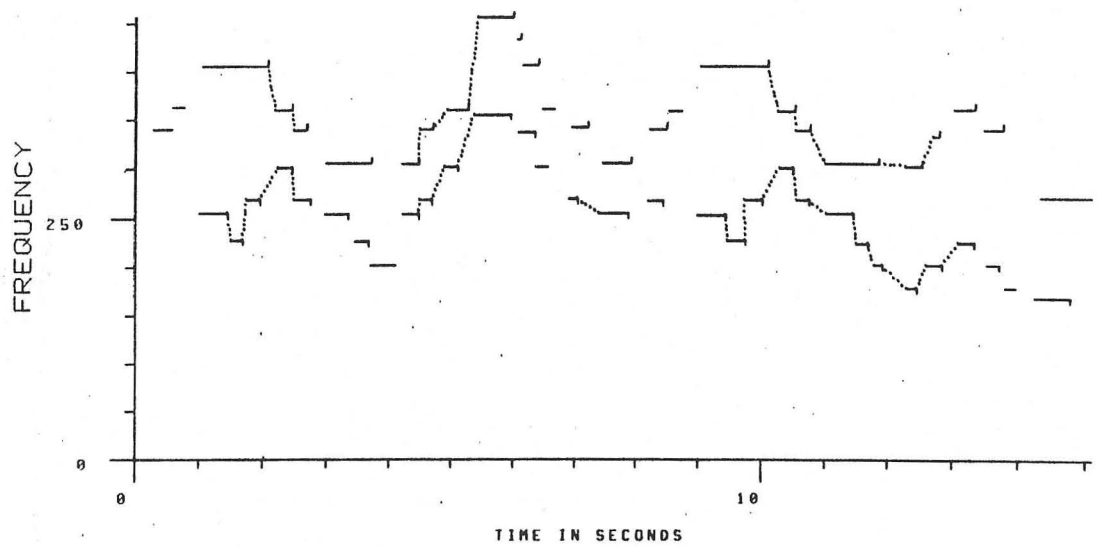


Figure 53. This shows the first stage of the melodic grouping. This is from a guitar duet. The score for this piece is shown in figure 60. Each note is specified by a horizontal line. Some notes have already been assigned to the upper or lower voice. There is a "tail" on each assigned note that points up or down, denoting membership in the upper or lower voice, respectively. Those notes that are assigned to voices in this plot were so assigned by finding pairs of notes that were sounding simultaneously. In such a case, the upper note will be assigned to the upper voice, and the lower note to the lower voice. The dotted lines indicate a melodic connection between adjacent notes of a melody.

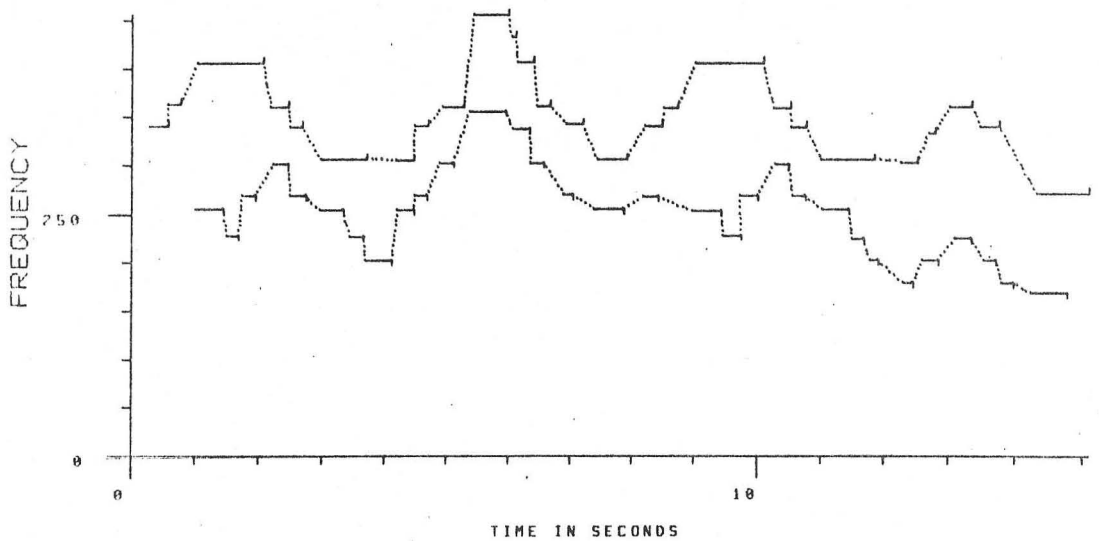


Figure 54. This shows the final melodic grouping. This is from a guitar duet. The score for this piece is shown in figure 60. As in the previous figure, each note is specified by a horizontal line. Some notes have already been assigned to the upper or lower voice. There is a "tail" on each assigned note that points up or down, denoting membership in the upper or lower voice, respectively. The remaining melodic membership was assigned by determining the voice assignment which minimized the sum of the magnitudes of the differences in frequency between each pair of adjacent notes in any proposed melodic assignment. Since the number of notes is small, this was done by a factorial search.

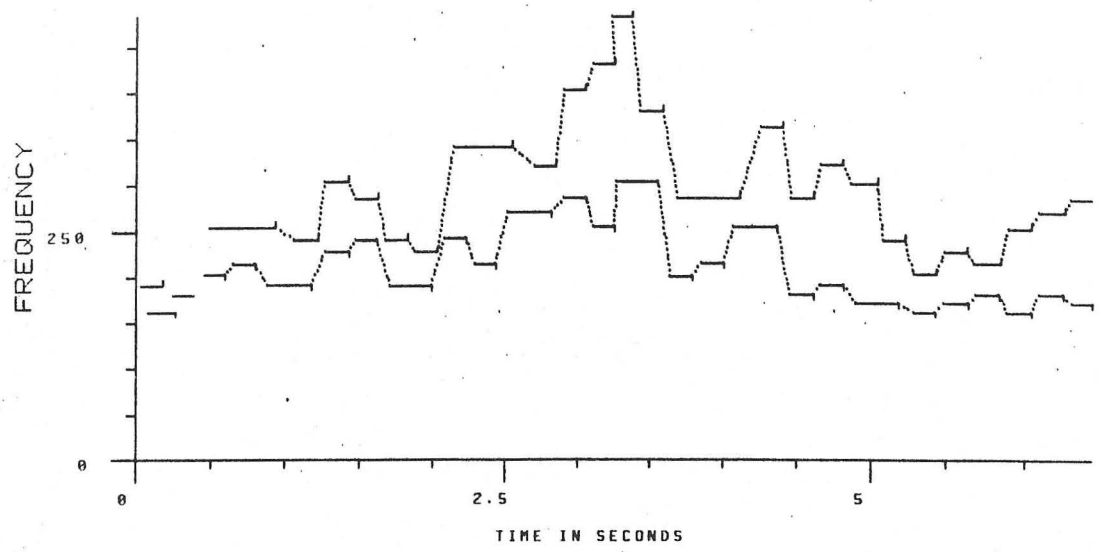


Figure 55. This shows the first stage of the melodic grouping for the pseudo-violin duet. The score for this piece is shown in figure 58. The format of this figure is like that of the previous two figures.

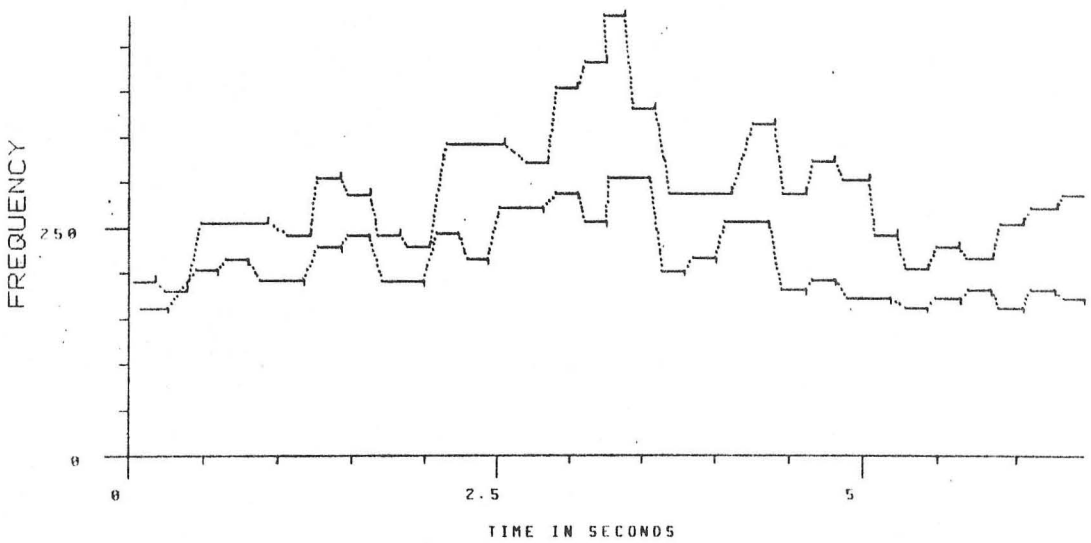


Figure 56. This shows the final melodic grouping for the pseudo-violin duet. The score for this piece is shown in figure 58.

ON MANUSCRIPTING

Once the melodies are determined, the manuscripting is just a matter of preparing input for Leland Smith's manuscripting program [1973]. Smith's program relieves us of having to consider the exact geometric and spacing details, but it does not guarantee that what is printed makes good musical sense. For instance, it is a convention that once an accidental occurs in a measure, the effect of the accidental persists throughout the measure. This means that we must keep track of each accidental and reset the flag at the end of the measure.

It is also a convention that a note of a certain duration shall only be written on an integral number of those durations into the measure. For instance, a syncopated note of three eighth-notes duration which begins after an eight rest at the beginning of a measure is usually not written as a dotted quarter. It is usually written as an eighth tied to a quarter. Thus we must build up each duration from an assemblage of notes connected by ties.

Still, compared with the difficulties involved in the low level tasks, this aspect of the problem is simple.

There is, of course, indeterminacy in a musical score. We can scale all the note representations by any number of factors of two and still make musical sense. A piece written in 4/4 can be written equivalently in 2/2 with little difficulty. We rely on the human to resolve the ambiguity in this case.

Although some work has been done on inferring the key and time signature of pieces [Longuet-Higgens and Steedman, 1971], we did not attempt to do so here. The reasons are that it would appear that any algorithm to do this must be dependent on the style, and that some of the pieces we were interested in were atonal pieces and thus had no key signature. It would be an interesting exercise to see if the key and time signatures could be inferred in general in some meaningful way.

Also not discussed here is the problem of tracking rallentando, accelerando, or other slow changes in tempo. This provides a special problem for the musical scribe. Detecting the beat, especially if any syncopation is involved, seems to be quite difficult. It is hard to define a strategy that will do this in any general fashion.

There is also the problem that the times and durations that the computer determines will be, in general, real numbers, whereas these must be converted to simple rational lengths for the score. We do this by asking the user what the smallest length note is that he will accept. All note positions and durations are forced into multiples of this length. This means that the user can ask for a quite grotesque score by giving a very short duration as the fundamental length.

This is not really a satisfactory arrangement, because we are generally less concerned with when very long notes end than when shorter notes end. Thus, to specify the duration of a note that is

slightly longer than a whole note down to the nearest 64^{th} note may not be exactly what is called for. Yet, if the composer wishes to specify a tone that continuously melts into a rapid, syncopated segment, that is exactly how he would write it. In other words there seems to be many options as to how to notate such cases, depending on the exact style of music involved and the ideas the composer is trying to embed in his piece. We have taken a somewhat neutral attitude here by attempting to do only an adequate job, rather than a superlative one in choosing among the printing alternatives here.

THE DFT AGAIN

A trial system using the discrete Fourier transform (DFT) was made before we realized that such a system was not capable of dealing with reverberation or vibrato. Although we do not attempt to deal with vibrato here, the ability can be worked into the current framework without too much difficulty. This is not true for the DFT. In any case, let us present the results of low-level analysis using the DFT.

The DFT-based system made one complete pass through the sound waveform and applied a 4096-point DFT every 10 milliseconds. At a sampling rate of 25600, the DFT window was 160 milliseconds wide. Since a second-order weighting function was used, the effective width of the window was less than half of this. This is similar to the averaging period of the bandpass filters that were previously discussed. The magnitude of the DFT was computed. Peaks were detected in the spectrum and were interpolated to get the frequencies and amplitudes more accurately. The method of Rife and Vincent [1970] were used for the weighting and interpolation. In their terminology, method II was used with a class-III weighting function of second order.

The first measures of two pieces were done. The guitar duet, whose score is shown in figure 60, and the pseudo-violin duet, whose score is shown in figure 58. Figures 57a and 57b correspond to the guitar duet, figures 57c and 57d to the pseudo-violin duet.

In each piece, the left-hand figure has a point for every peak in the DFT that was found. The vertical axes are labeled in Hertz and represent the frequencies of the spectral peaks. In the right-hand figures, the points in adjacent time slots have been linked together into lists. The head of each list is marked on the plot by a small vertical stroke. Isolated vertical strokes are lists of one element. We can see in the pseudo-violin duet that some harmonics which actually belong to different notes have been merged because of their proximity in frequency. We can see this in the lower plots (57c and 57d) in the fundamental frequency of the first two notes in the lower voice. These two notes actually occupy the first and second 200-millisecond windows of the piece. In 57d, we see that the two harmonics have been linked together, because the peak in the DFT representing these harmonics moved smoothly from one frequency to the next at about 200 milliseconds into the piece. This can be dealt with later by noticing that the frequency has a quantum jump over the duration of the harmonic.

This method might be viable for non-reverberant, non-vibrato cases, although for the guitar piece, some method would have to be developed to recover the missing harmonics, such as the second harmonic of the A3 (220 Hz) at about 1400 milliseconds (figure 57b). The second and third harmonics of the note only appear briefly in the DFT.

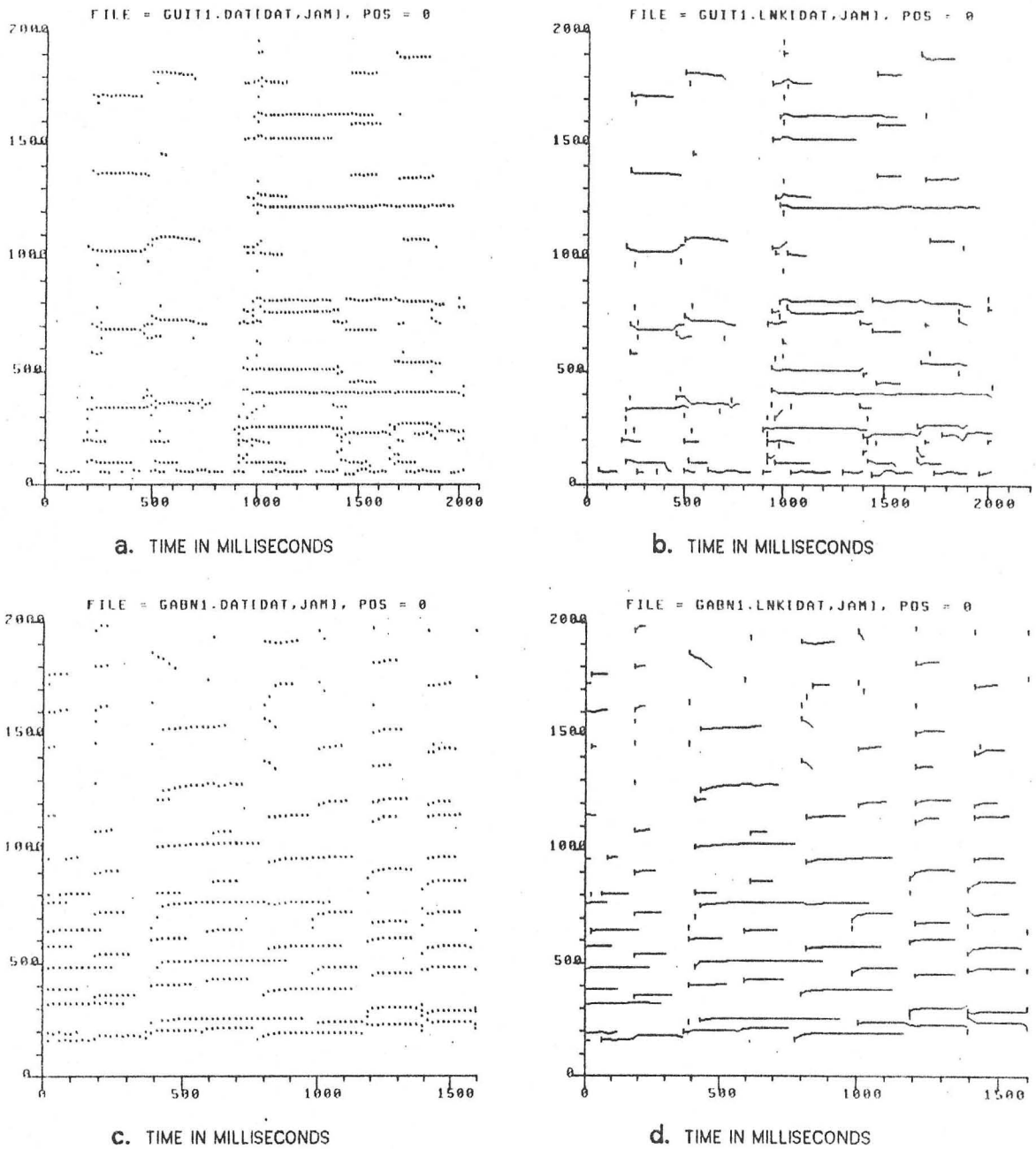


FIGURE 57. These are the results of an experimental system using only the discrete Fourier transform as the low-level routine. Every 10 milliseconds, a new DFT was computed. In the left figures (a and c), each point represents a peak in the DFT. All horizontal axes are in milliseconds, the vertical axes are in Hertz. The right figures (b and d) have been processed to link peaks in adjacent time windows. A vertical stroke denotes the beginning of a list of consecutive peaks. The piece that produced the top plots (a and b) is the guitar duet whose score is shown in figure 60. The piece that produced the bottom plots (c and d) is the pseudo-violin duet whose score is shown in figure 58. In each case, only the first measure of the piece is shown here. The transform was 4096 points (160 milliseconds long) and a second-order time window was used. The method of Rife and Vincent [1970] was used to interpolate the peaks. We can see, especially on the guitar piece, that harmonics of notes known to exist are often missing. Although there is not an exactly analogous illustration, we can compare this with the results of the programs using bandpass filtering in figure 53.

YES, BUT DOES IT WORK?

INTRODUCTION

In this section we present the results of our work, a critical review of its faults, and some ways that a future system might better be constructed.

One of the pieces shown was entirely synthetic, essentially untouched by the disturbing properties of transmission through the air. This was done for debugging purposes. The other piece was performed by the author and recorded at home on a cheap Sony tape recorder. Both pieces were composed by the author. They are both segments of larger pieces. They were chosen because they both exhibit properties that make them compatible with the restrictions we have imposed on the kind of music that will be accepted for analysis.

In discussion possible improvements, we deal with each stage of the analysis separately. We outline a possible two-step filtering scheme that uses wide band filters to determine the strongest sinusoid in a given frequency region, then a narrow band filter to extract that sinusoid individually.

A rating scheme for notes is suggested which is somewhat like that applied to individual amplitude and frequency traces. This would allow comparison of note hypotheses and a similar sort of maximizing search would be possible.

Other improvements include changing the tempo to compensate with the performer's tempo changes, and many other fine points.

SOME EXAMPLES

Here we present two examples to show the operation of the system as a whole. The first example is synthetic and was both synthesized and analyzed entirely within the computer. This was the piece the programs were debugged on. The utility of working with a synthetic piece is that one knows exactly when each note begins and ends, exactly what the pitches of the notes are, and exactly what their amplitude and frequency functions are. It is a little unreal in that there is no digitizing noise, no room noise, no spurious sounds from box and string resonances, and no room reverberation. The second piece, however, possesses all these problems.

Even though the synthetic piece has no noise, it is still not a trivial example. It is non-trivial because the tones were generated from the analyses of actual violin tones by use of the heterodyne filter which preserves all the highly time-variant properties of the tones. Another reason why the piece is non-trivial is that it is quite fast. Quarter-notes occur at 160 per minute, making the length of each eighth-note only 200 milliseconds. Since the note is staccato, its *effective* length is even shorter. These short notes spell death to most signal-processing techniques because there is little or no steady-state portion of the signal. Transient responses are strongly excited.

Figure 58 shows the original score of the synthetic piece. This piece was synthesized for pseudo-violin, using the analysis data of an actual violin. It sounds a little strange because only the analysis data of an Eb4 was used to synthesize all the notes. When you resynthesize a note off the original frequency, the timbre of the tone is altered, sometimes quite a bit, although the spectral shape and the transient behavior is identical at either frequency.

Figure 59 is the final output of the transcription programs. As is easily seen, all the notes are present, they begin at the correct times, and they are at the right pitches. The note lengths, however, have been consistently underestimated. This is because the segmentation algorithm threshold was set quite high to eliminate noise traces and consequently eliminated some good data. Any more comprehensive system should go back and, knowing the pitch and rough duration, analyze specifically for the time limits of each note. Knowing the pitch of the note and all the simultaneously sounding notes would enable us to perform this analysis.

Figure 60 shows the original score of a guitar duet. This piece is somewhat slower than the previous one. The eighth-notes are of about 250 milliseconds duration, for an overall tempo of 120 quarter-notes per minute.

Figure 61 shows the final output of the transcription programs for this piece. Again, the durations are consistently underestimated. There is one note missing toward the end of the piece. This was lost due to one harmonic being coincident with the other note sounding at that time. and a second harmonic being lost due to noise. The remaining harmonics were not strong enough to infer a note at that position.

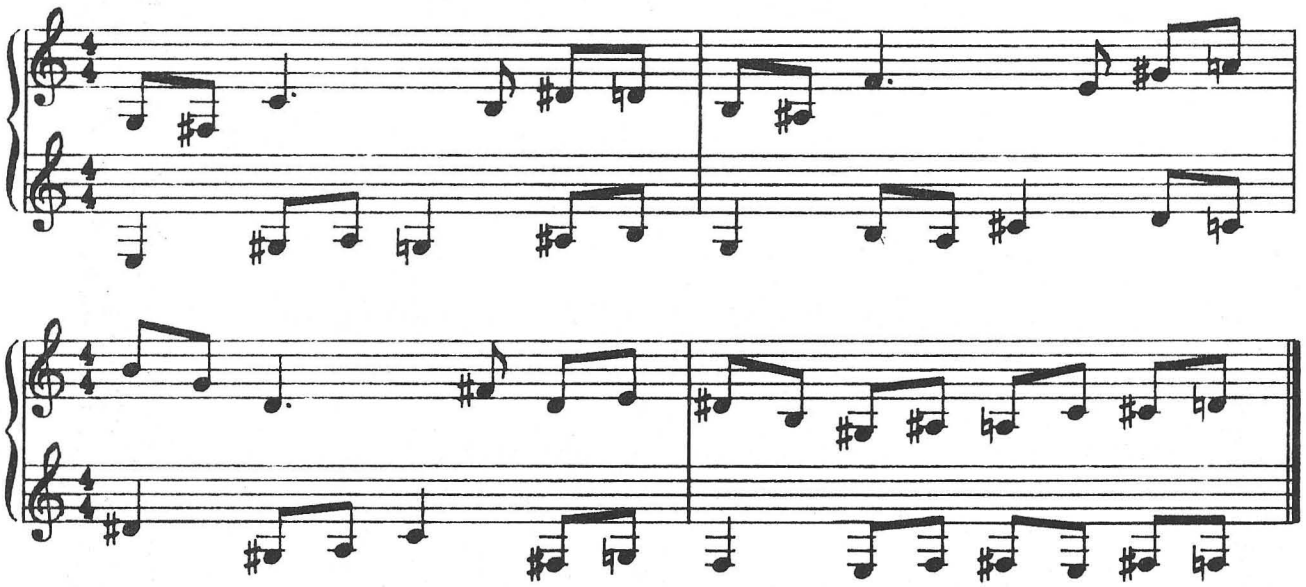


FIGURE 58. The original score for a pseudo-violin duet. The tempo is rather fast. There are 160 quarter-notes per minute, or about 200 milliseconds for each eighth-note. Since this piece goes below G3, this score could not .have been played on actual violins. With computer synthesized violin-like tones, we have no such restrictions.

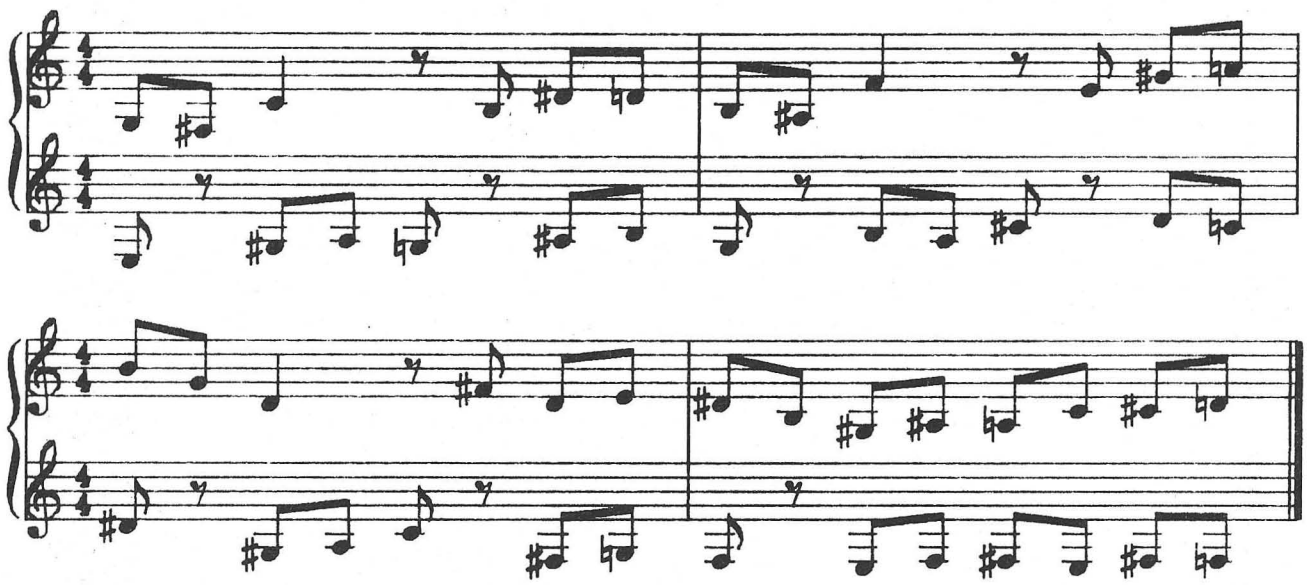


FIGURE 59. This is the score produced by the computer. The lengths of the longer notes are consistantly underestimated. This is because the threshold for noise rejection is set so high that the tail ends of the notes are lost.

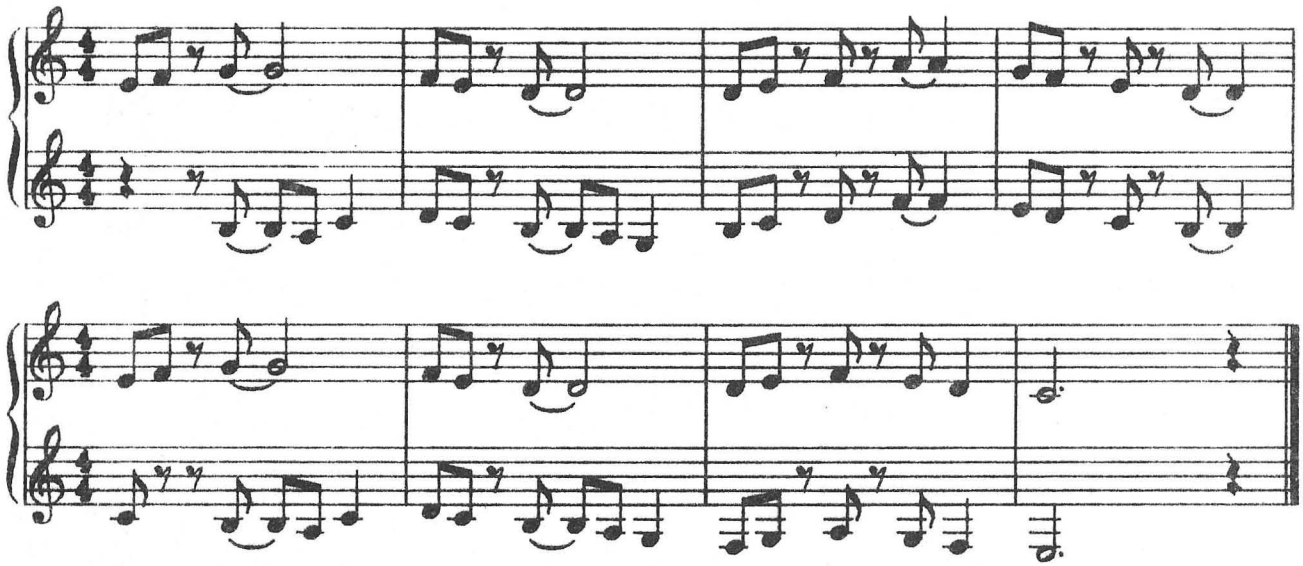


FIGURE 60. The original score for a guitar duet. The tempo is 120 quarter-notes per minute, or about 250 milliseconds for each eighth-note.

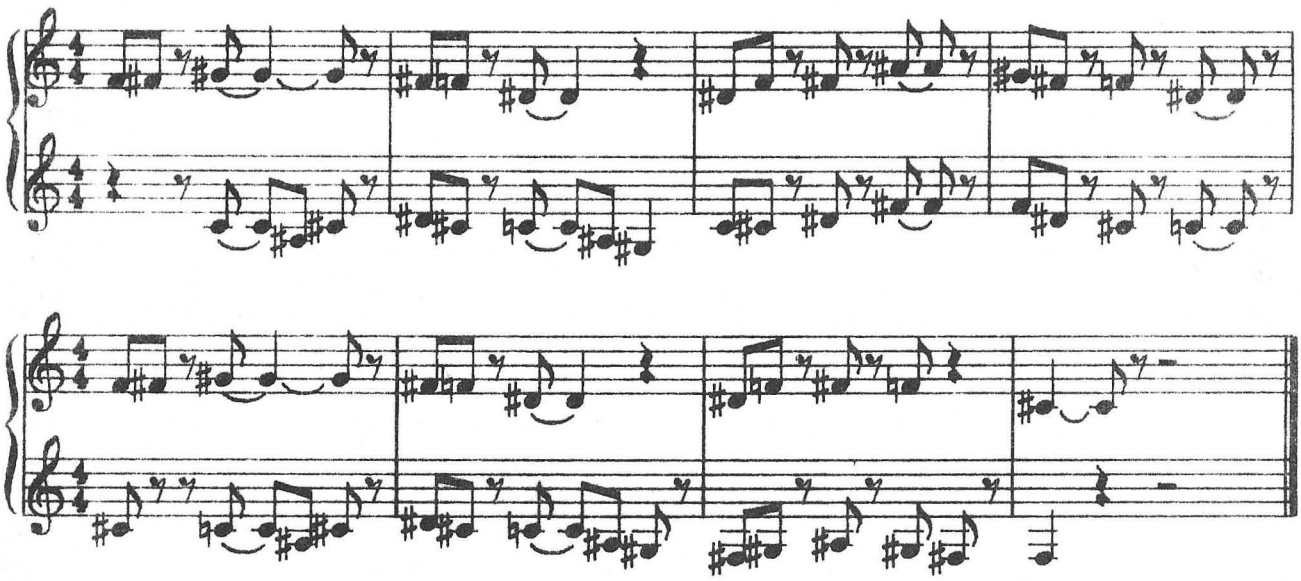


FIGURE 61. This is the score produced by the computer. Again, the lengths of the longer notes are consistantly underestimated. Also, there is a note missing in the last measure. The most conspicuous change, however, is due to the fact that the guitar was mistuned somewhat high. The literal-minded computer faithfully reports the score here one half-step high throughout. The intervals between consecutive notes is correct in terms of the number of half-steps the interval represents. Please note that this is not good musical notational style. This should be notated in the key of Db, which would make all the accidentals dissapear. We retain this notation because it is simple, general, and can represent 12-tone pieces as well as tonal pieces, although the representation is quite clumsy in many cases.

This points up another deficiency of the program that infers the notes from the harmonics: when a harmonic is used to infer a note and that note is accepted, that harmonic is removed from the list of harmonics. This means any subsequent note that might also use that harmonic must do without it. The program was arranged in this manner to help eliminate the problem of hypothesizing a note based on each harmonic present. This way, we hypothesize the lowest one, and remove all the harmonics from further consideration. Clearly, some compromise could be arranged.

One hopeful sign is that this guitar piece was recorded in a noisy environment, with poor equipment and no special care taken in type of tape used, type of tape recorder, type of microphone, microphone placement, or any of a number of considerations that define good recording technique. The only consideration was that the recorder was not saturated during the recording.

In fact, the guitar was not tuned to A4=440 Hz. for the recording. The result of this is that all the pitches were about 2 percent higher than concert. The program rounded this upward and printed the score uniformly one half-step higher throughout. This shows the literal-minded nature of the computer. We did nothing to correct this mistuning. A more comprehensive program would notate this piece in the key of C# or Db. We made no attempt to do so here. We might expect that doing this for *a capella* vocal work would result in the score slowly drifting from the original key. The program is arranged so that this would be notated as a sudden shift in key by one half-step.

WHAT NEXT?

After this exposition, we ask the question *how can we do this better?* As it turns out, constructing the programs to actually demonstrate the concepts of the system were very enlightening as to how it all should have been done. We shall examine the system one piece at a time to give a presentation as to how this task can be done better and what the weakest parts of the current implementation are.

PREDICTION AND FILTERING

Since most of the computer time for the task is used by prediction and filtering, we might look to see how they might be improved. One could imagine a two-level search strategy something like the following:

First, a bank of wide-band (third-octave perhaps) filters is applied. If the energy in the output of the filter is too small, that frequency band is not analyzed further. A filter of this wide bandwidth will, in general, pass several sinusoids at once. A pitch detector is applied to the output of the filter. There exist pitch detectors that will detect the pitch of the strongest sinusoidal component in the signal. This gives us the frequency of one of the sinusoids that is passed by the filter.

Once we obtain this frequency, we may apply a more narrow band filter to exactly this frequency as well as to integral multiples and integral fractions of this frequency, so as to capture the subharmonics and harmonics of the sinusoid. We may progressively narrow the band of the filter until it is clear that no sinusoid is present at this frequency or until we get a good estimate of its frequency. Once we know a sinusoid is present at a particular frequency and what bandwidth filter is necessary to extract it, we may sweep forward and backward in time, searching for the true extent in time of this sinusoid.

There are various complications that may occur which should be noted. First, another note may sound at some other time that would require us to make the filter much more narrow. We can tell this by noting that the output of the pitch detector suddenly becomes garbage when there is still plenty of energy at that frequency.

If another sinusoid suddenly were to appear at very nearly that same frequency, we could notice the sudden phase change, which would manifest itself as a spike in the frequency trace. The total energy in the filter output would presumably increase, unless the sinusoids cancel each other out. They may also beat.

Another thing that may happen is that there may be vibrato on the sinusoid, which would imply that its frequency is constantly changing. We may track the frequency by making the filter frequency follow the frequency estimate from the pitch detector. This has stability

problems. We must introduce some smoothing so that instabilities do not occur. We must force the filter to stay within certain bounds, such that excursions outside these bounds will be taken to mean that the trace is noisy and that either nothing is present or a more narrow filter must be used. Let us note that the problem of tracking the frequency of a single (monophonic) periodic signal is one that has been addressed extensively by the speech community. Some groups consider this to be a solved problem. We believe that there is still work to be done in the case of a noisy environment, as we have in this case. Even if the piece is recorded in a very quiet room, there is always the "noise" consisting of the vibrations of the strings that are not being played.

We persist in using bandpass filters rather than DFT or other signal-processing techniques on the grounds that the filter gives us a great deal of flexibility, it can deal with reverberant environments, it preserves time information, and can handle continuously-changing frequencies. This last feat cannot be performed with the DFT simply by looking for a peak at a certain place. Only a time-variant (adaptive, in this case) filter can deal successfully with vibrato.

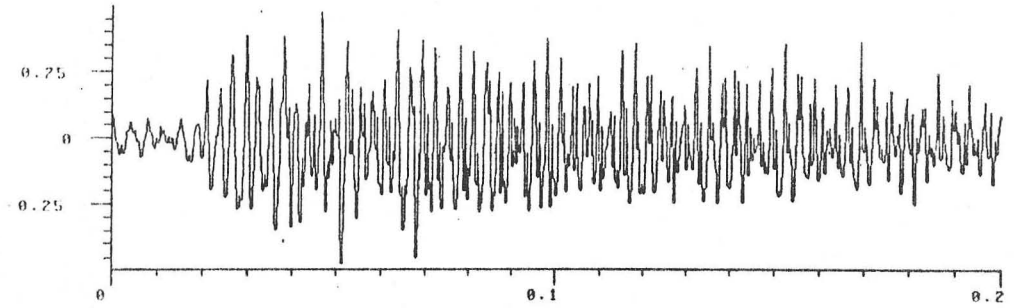
These procedures, we believe, can accomplish the low-level tasks well in somewhat less compute time, providing much more power.

To show how this might work, we have computed some test cases using a 200 millisecond segment of a two-part piano piece. The notes being played during this segment are a D4 at about 294 Hz (3.4 milliseconds period) and an F4 at 349 Hz (2.86 milliseconds period). Figure 62a shows the waveform of the signal itself. Figure 62b shows the discrete Fourier transform of the waveform. We can see the notes and their harmonics clearly (plus a lot of other stuff). Figure 62c shows the cepstrum of this waveform. As we might expect, the cepstrum of this polyphonic piece is a mess. The peaks do not seem to correspond to the periods of anything that we know is present in the signal. Figure 62d shows the autocorrelation of the waveform, and figure 62e shows the optimum-comb applied at a place in the middle of the waveform segment. These last two plots show significant activity at multiples of the periods of the notes that are present. We notice that the peaks coincide at about 17 milliseconds. This is because D4 and F4 form a minor third. This implies that their frequency ratio is about 5/6. Indeed, 5*3.4 milliseconds is 17, and 6*2.86 milliseconds is about 17.16 milliseconds.

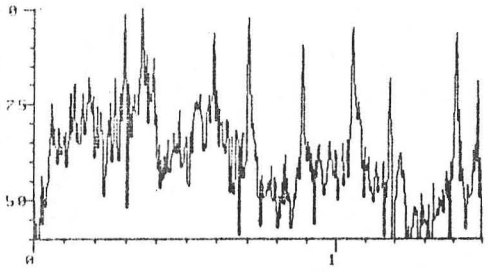
The next figure, number 63, shows the same sequence of plots for the filtered waveform. The waveform in figure 63a was filtered with a 4th order Butterworth bandpass filter with 3dB points at 170 Hz and 230 Hz. The filtered waveform is shown in figure 63a. As can be seen from the successive plots, we seem to have isolated a signal at about 174 Hz. This is a subharmonic of the F4 which is probably caused by a lower string resonating.

We can see that the amplitude of the filtered signal is somewhat low. This may be our only clue for eliminating this signal from consideration later.

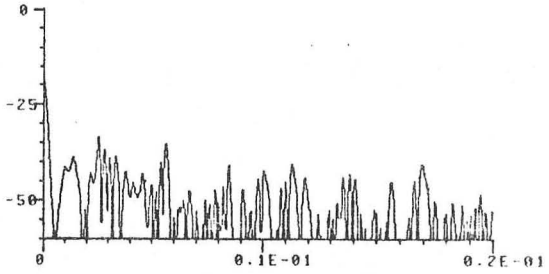
Figure 64 shows the original piano waveform filtered by a similar filter with 3dB points at 255 Hz and 345 Hz. We see the D4 shining through on the subsequent plots.



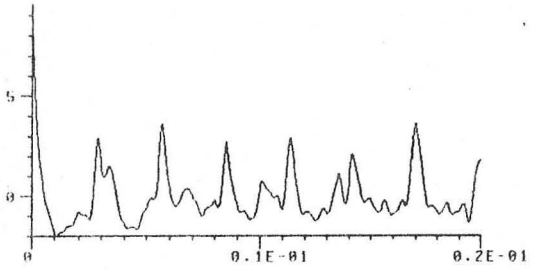
a. TIME IN SECONDS



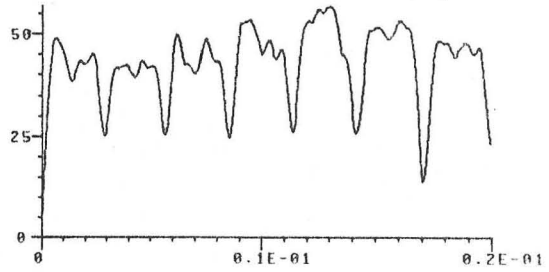
b. FREQUENCY IN KHZ.



c. TIME IN SECONDS



d. TIME IN SECONDS



e. TIME IN SECONDS

FIGURE 62: This and the following three figures examine a 200 millisecond segment from the middle of a two-part piano piece. Present at this time are a D4 at 294 Hz (3.4 milliseconds period) and an F4 at 349 Hz (2.86 milliseconds period). Figure 62a shows the sound waveform itself. Figure 62b shows the discrete Fourier transform of this segment of sound. We can see the peaks corresponding to the notes quite clearly. Figure 62c shows the cepstrum of this segment. As we might expect, the peaks in the cepstrum do not seem to have any obvious meaning. Figure 62d shows the autocorrelation of the music waveform. We can see peaks corresponding to the subharmonics of the two notes present. At just over 17 milliseconds, the peaks line up. This is because D4 and F4 form a minor third which implies a frequency ratio of nearly 5/6. In fact, 5*3.4 milliseconds is about 17 milliseconds and 6*2.86 milliseconds is about 17.16 milliseconds. Figure 62e shows the optimum-comb applied to this waveform. We can see that it corresponds greatly to the inverse of the autocorrelation with the exception that the minimum at 17 milliseconds is more pronounced than the maximum in the autocorrelation at 17 milliseconds. Neither is very prominant, compared to the other features in the plots.

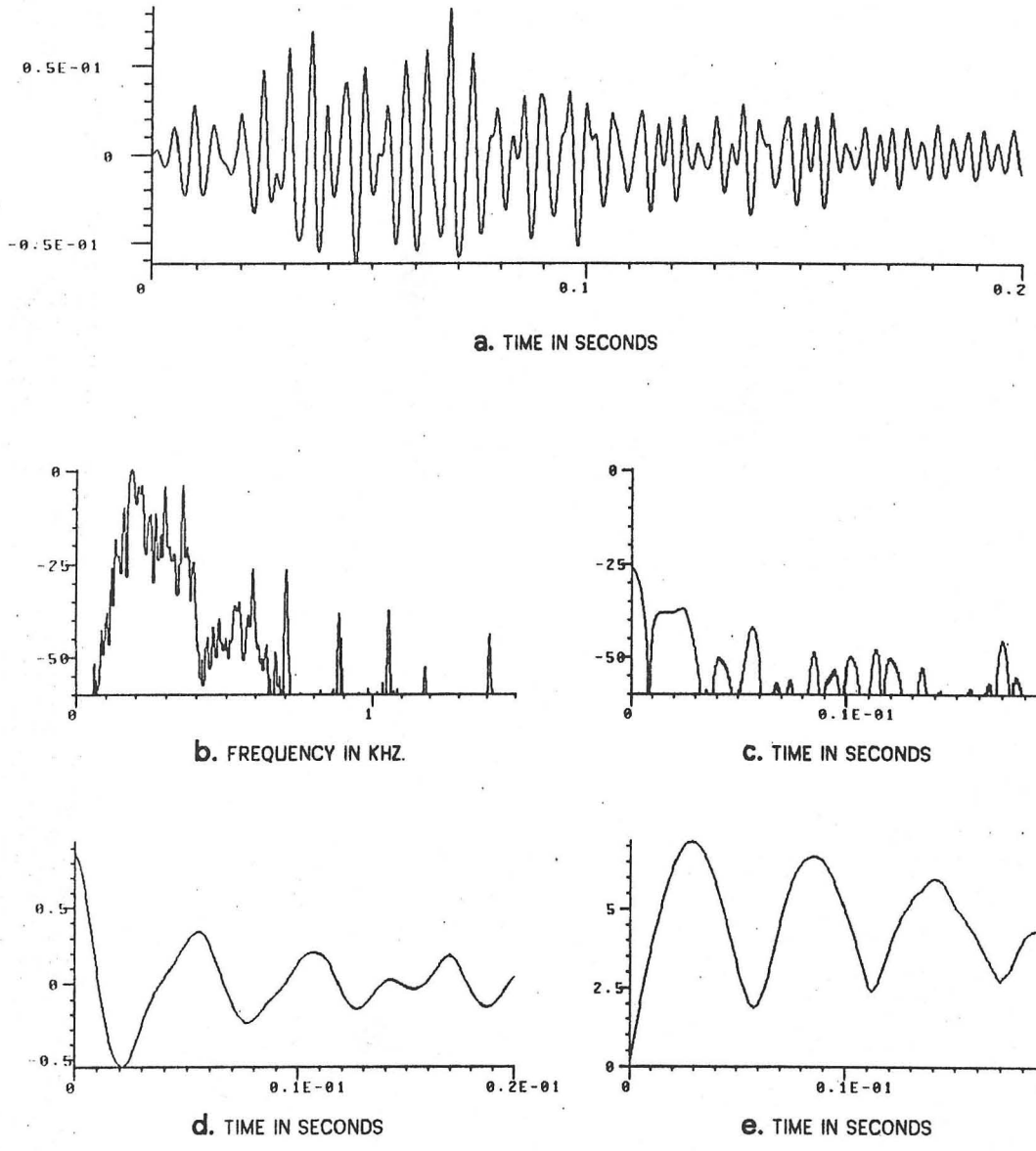


FIGURE 63: The upper plot shows the waveform of figure 62a filtered by a 4th order Butterworth bandpass filter whose 3 dB points were at 170 Hz and 230 Hz. Again, figure 63b is the discrete Fourier transform of the waveform shown in figure 63a, figure 63c is the cepstrum, figure 63d is the autocorrelation, and figure 63e is the optimum-comb. We can see that the autocorrelation and the optimum comb seem to have detected a frequency at about 5.8 milliseconds. This is about 174 Hz, or an F3. This is a subharmonic of the F4 that is being played. It is quite likely that this represents a spurious resonance of one of the lower piano strings.

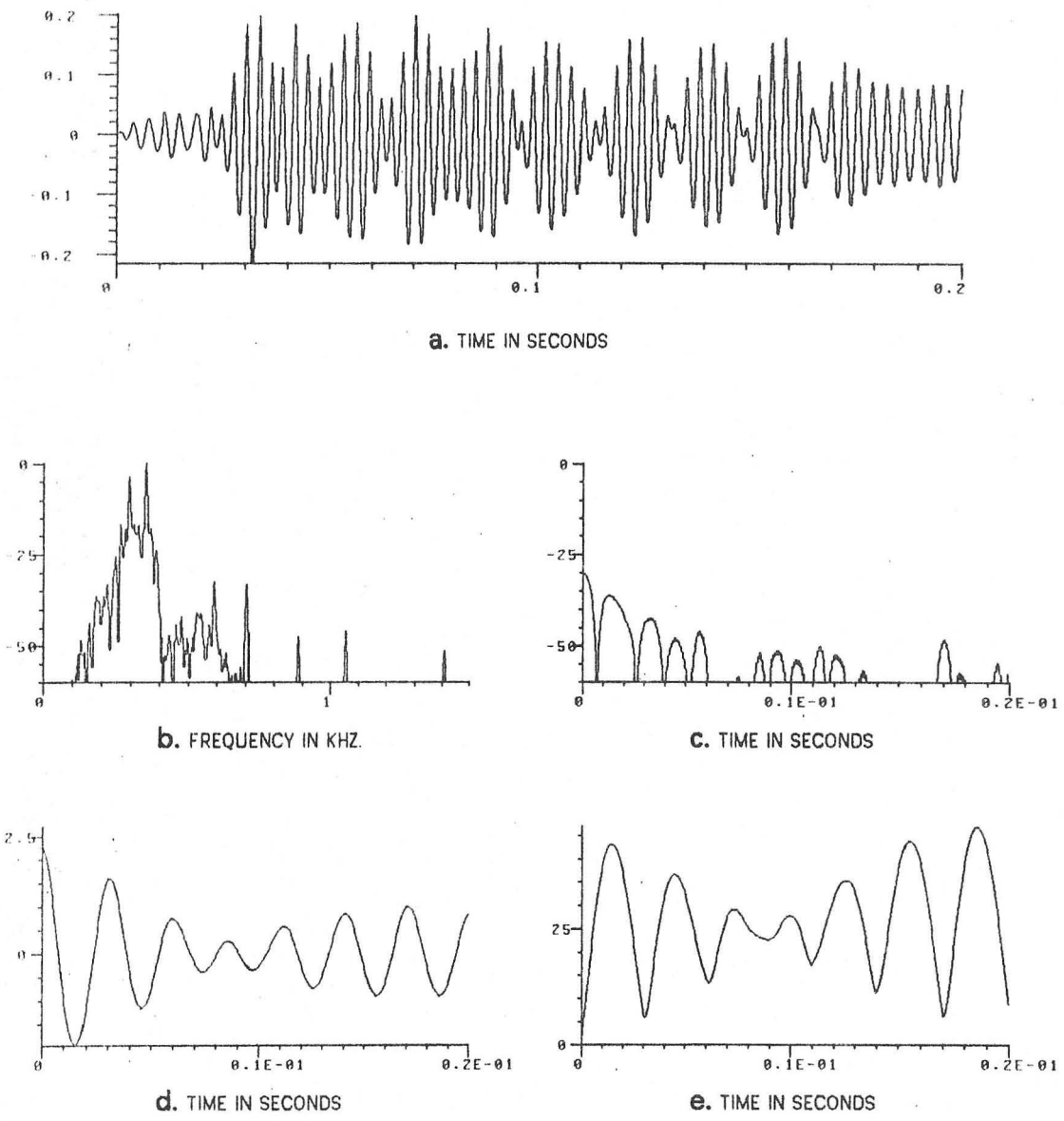


FIGURE 64: The upper plot shows the waveform of figure 62a filtered by a 4th order Butterworth bandpass filter whose 3 dB points were at 255 Hz and 345 Hz. Again, figure 64b is the discrete Fourier transform of the waveform shown in figure 64a, figure 64c is the cepstrum, figure 64d is the autocorrelation, and figure 64e is the optimum-comb. We can see that the autocorrelation and the optimum comb seem to have detected a frequency at about 3.4 milliseconds. This corresponds well to the period of the D4 that is being played.

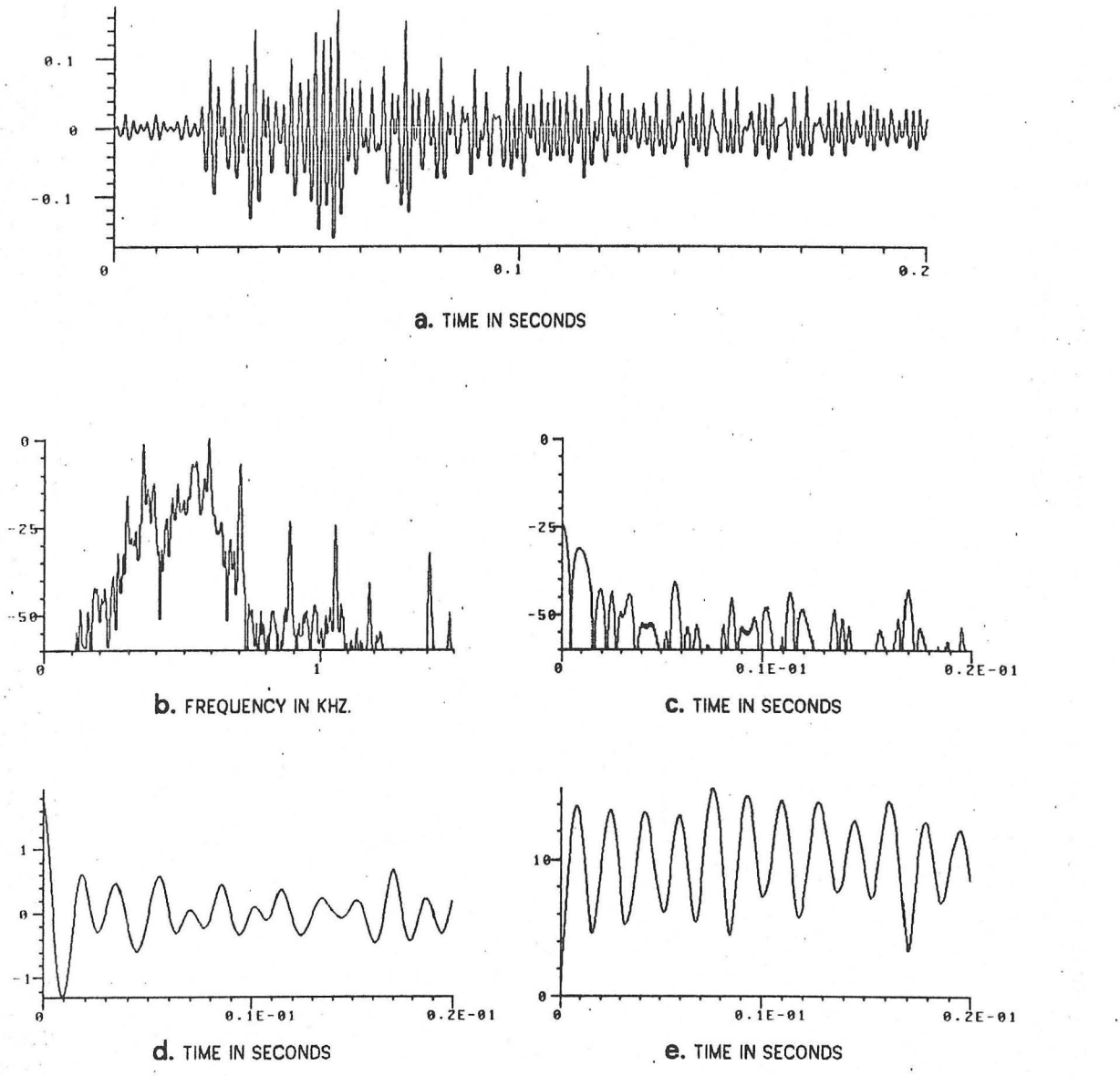


FIGURE 65: The upper plot shows the waveform of figure 62a filtered by a 4th order Butterworth bandpass filter whose 3 dB points were at 425 Hz and 575 Hz. Again, figure 65b is the discrete Fourier transform of the waveform shown in figure 65a, figure 65c is the cepstrum, figure 65d is the autocorrelation, and figure 65e is the optimum-comb. We can see that the autocorrelation and the optimum comb seem to have detected a frequency at about 1.7 milliseconds. This corresponds well to the period of the second harmonic of the D2 that is being played.

Likewise with figure 65, the 3dB points are 425 Hz and 575 Hz. We get indications of a signal of period about 1.7 milliseconds, which corresponds roughly to the second harmonic of the D4.

We hope that these examples show that the 2-level search procedure described above has potential.

INTERMEDIATE LEVEL PROCESSING

One of the most important techniques that should be incorporated into the intermediate-level routines is the ability to consult the original sound waveform again to verify details, such as the exact beginning and ending times of harmonics. Since the intermediate-level routines know what frequencies are hypothesized to be present, they are optimally suited to determine how a sinusoid to be verified should be extracted.

We could envision a system which formulated many hypotheses before beginning to eliminate them. The current approach is myopic, in that it formulates a note hypothesis from the harmonic data and decides then and there whether to accept it. We should formulate the N strongest hypotheses at each point in time and find a rating system to decide among them. These hypotheses then might serve as guides to returning to the original sound file and searching for missing harmonics.

In the current programs, the filter bandwidth is a constant small size. This means that the timing information, such as when the harmonic starts, ends, and its exact amplitude envelope, is not terribly accurate. It has been greatly smoothed. If we used variable filter bandwidths, such that the widest filter was used that successfully extracted the harmonic, some of this time resolution might be regained. This would allow us to use this detailed time information in the intermediate-level processing. For instance, we could easily distinguish a spurious resonance by noting that its onset corresponds to some time after the onset of another stronger note in the piece. We might be able to distinguish notes at octaves by the onset times. The detailed frequency variations will help with that also, especially since one is likely to have different vibratos. We might also think about using the detailed amplitude envelope of the harmonics. In plucked or struck instruments, the time of the initial maximum that each harmonic attains soon after the beginning of the note could be used as a cue that these harmonics belong to the same note. One must be a bit careful, in that generally the high harmonics of a plucked string occur first, followed by the fundamental.

ON IDENTIFICATION

It is theorized that the attack portion of the tone is a very important cue for human identification of the instrument. It is possible that by increasing the time resolution of the low-level routines, machine identification of the instrument will be possible. It is clear that identification, human or otherwise, cannot be done on the average amplitudes of the harmonics alone. For instance, with two instruments playing at octaves, the harmonics overlap entirely,

such that each pair of harmonics will either add or cancel to some degree. This produces a completely unique spectrum. Either we must theorize that the human can recognize that this is the octave conjunction of two instruments, or that the human can somehow separate the individual contributions of the instruments, or we must admit the possibility of factors other than the harmonic amplitudes being used. In John Grey's dissertation [Grey 1975], three cues for timbre were strongly suggested. One was the bandwidth of the signal, which roughly means the number of harmonics present. Another factor was the type of noise burst at the beginning of the tone. A third factor was roughly related to the overall attack time of the tones in question. Two of these three cues are in the attack portion of the tone.

CONCLUSIONS

In this dissertation, we have examined the problem of the transcription of musical sound by digital computer. A series of programs were developed using many signal-processing and artificial-intelligence techniques which accomplish the task of automatic musical transcription on a limited basis. Most of the limitations were introduced for convenience and for the purpose of finishing the dissertation in a finite time. In fact, straightforward extensions of the techniques used in these programs would allow elimination of many of the restrictions.

The overall plan of the system was as follows: First, an attempt was made to determine the *harmony* of the piece through the use of a periodicity detection algorithm. This gave us *root frequencies* whose multiples were guaranteed to represent the frequencies of all the sinusoids present in the signal. Narrow bandpass filters were then centered on these frequencies to try to extract each of the harmonics of each of the tones present separately. A pitch detection algorithm was used at the output of each filter to determine if there was any periodicity at that frequency. A rating of each filter output was made which represents the quality of the filter output. This rating was used to choose the "best" signals to use to infer the notes. The notes were inferred by choosing high quality signals and then finding harmonics around them to form a complete note. The note hypotheses were compared and the best ones selected. A melodic grouping algorithm divided the notes into upper and lower voices. The melodic information was then formatted and delivered to the manuscripting program which produced the final hard-copy score output.

The restrictions imposed on the music were as follows:

All tones are nearly periodic. This eliminates drums, gongs, and other percussive instruments. We have not dealt with the problem of detecting and tracking *wide-band* and *inharmonic* signals which these instruments represent.

All frequencies are nearly piecewise-constant. This eliminates trills, vibrato, and glissando. This was just so that we could use filters at fixed frequencies. The programs can be upgraded to use adaptive filters which chase the tone around as the pitch changes.

The fundamental of a note will not overlay a harmonic of another note sounding simultaneously. We do not understand at this time all the factors that are involved in human separation of notes with these characteristics. We do not understand why people "fuse" the harmonics of an instrument into a single percept, but distinguish two separate instruments which are playing in unison. Perhaps if the frequencies and attacks were exactly synchronized, people could not so distinguish them. We must do further experiments in human perception to gain insight into these processes.

The piece contains no more than two voices. This was done for convenience. There is no

inherent limitation which necessitates this. The melodic grouping algorithm, however, is not set up to track voices which cross.

Other limitations. Notes must be longer than 80 milliseconds in duration. This is because we distinguish between transient response from the bandpass filters and signals by assuming that the transient response will die out in less than 80 milliseconds. The use of variable-width filters can help distinguish this better. We also require that the fundamental frequency of a tone be present. This is because we do not have a convenient way of assigning a rating to an entire note right now. Presumably such a measure could be made. For the same reason, we require that the harmonics be *dense*, that is, have no missing harmonics, unless *all* the even harmonics are missing, as in the case of the clarinet.

With these restrictions in mind, examples were processed through the programs with relatively good results. The computer usage was enormous. This system can hardly be called practical at this time.

We feel the main contribution of this thesis is the knowledge that this task can be done by computer and it seems likely that it can be advanced to a relatively high level by simple extensions of the procedures developed here.

APPENDICES

APPENDIX A: THE HETERODYNE FILTER

INTRODUCTION

This appendix is devoted to implementation details and a critical evaluation of the heterodyne filter. The filter has been run on a series of synthetic tones which demonstrate its powers and its weaknesses well. For implementation details, we have chosen ALGOL as a vehicle for communication of algorithms. This is not necessarily directly useable on everyone's system, but we hope the implementation will be a simple matter of conversion.

A CRITICAL TEST

To empirically test the performance of the filter, we have chosen a periodic waveform with harmonics such that each harmonic is some fraction of the previous harmonic. We have placed an overall amplitude envelope on the test signal that consists of a line segment for the attack and constant amplitude for the steady-state. It is interesting to vary the time of the attack and see how the output of the filter behaves. In each of the cases shown, three smoothings were applied, each smoothing done by averaging over about one period of the signal. We present the results of these tests in figures A1 through A12. All of the figures except A6 have each harmonic 70 percent of the amplitude of the previous harmonic. Figure A6 has each harmonic 50 percent of the previous harmonic. We experiment with a 505 Hz signal and a 101 Hz signal. The first two figures, A1 and A2, show simple cases where the attack time is several periods long. In A1, the attack time is 25 periods, and in A2, the attack time is 10 periods. In each figure, there are four plots. The upper left plot is a perspective drawing showing the amplitudes of all the harmonics as derived by the heterodyne filter. In each case, we analysed up to the 10^{th} harmonic. The first harmonic is in the rear, the 10^{th} harmonic is in the fore. The upper right plot in each figure is a similar plot for the frequencies of the harmonics, except that the first harmonic is in the front and the 10^{th} harmonic is in the rear. The lower left plot is a pair of graphs showing the amplitude (upper) and frequency (lower) contours for just the first harmonic (fundamental) by itself. The lower right pair of graphs is the same thing for the 10^{th} harmonic. This is so that we can see exactly when the frequency trace stabilizes. In general, it takes a few periods before the frequency curve settles down. There is some confusion at the end of each plot due to the edge effects. For any practical situation, the tone should be surrounded by silence of length at least 4 periods on each side.

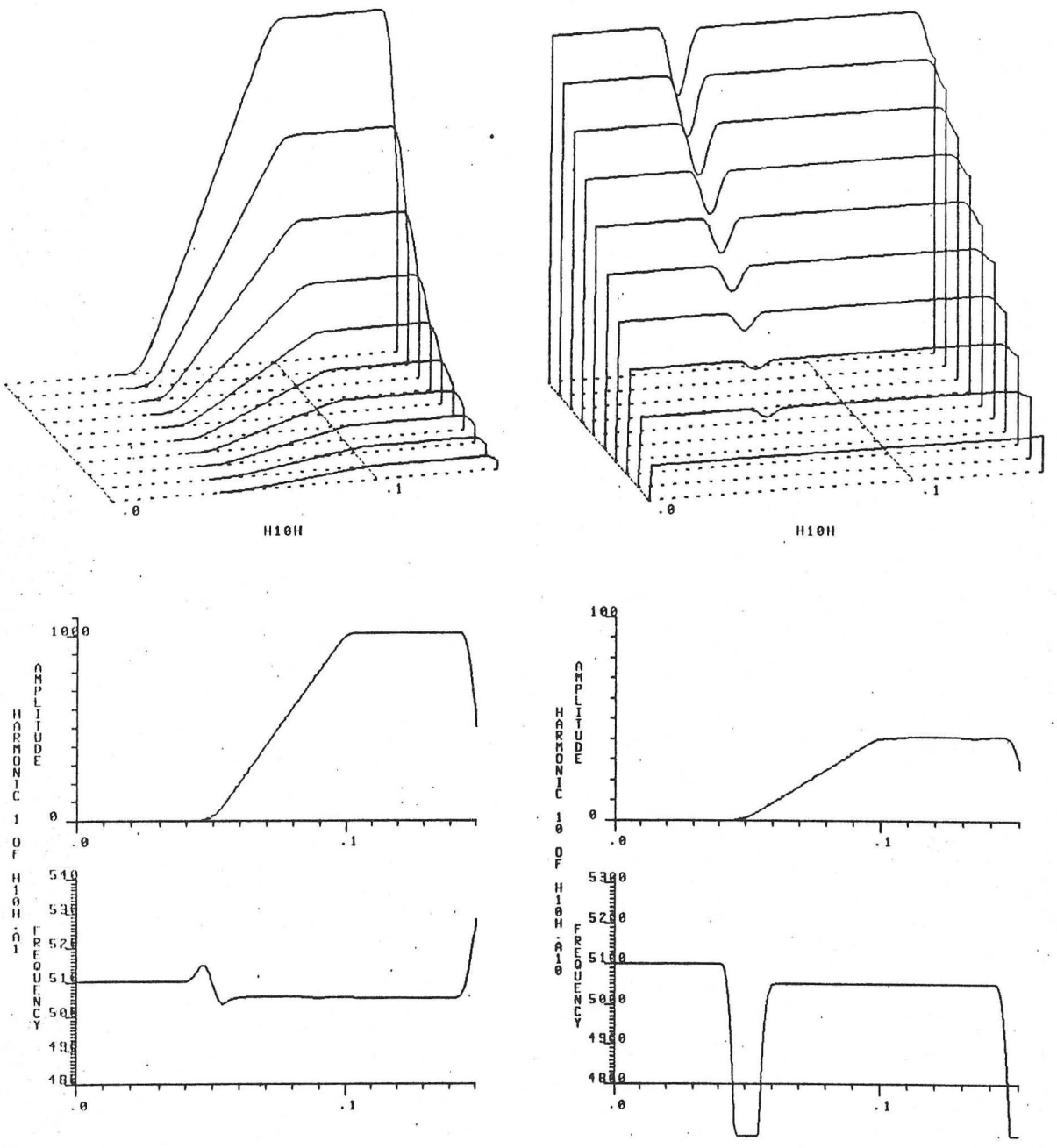


FIGURE A1. This shows the output of the heterodyne filter for a synthesized input signal which consists of a 505 Hz signal with a 50 millisecond linear attack on each harmonic. Each harmonic is 70% of the amplitude of the previous harmonic. The upper left figure shows a perspective plot of the amplitudes of the harmonics as determined by the heterodyne filter. The upper right plot shows the frequencies of the harmonics. The lower left pair of plots show the amplitude and frequency of the first harmonic, the lower right pair show the amplitude and frequency of the 10th harmonic. There is error in the frequency plots around the attack and the ending, but the amplitude plots seem to be accurate except for a slight rounding of the ends of each line segment. If we set the amplitude of the fundamental to 1.0, then the harmonic amplitudes are as follows: 1.0, 0.7, 0.49, 0.343, 0.240, 0.168, 0.118, 0.082, 0.058, 0.040 These plots were generated using a program which was written by John Grey for his dissertation.

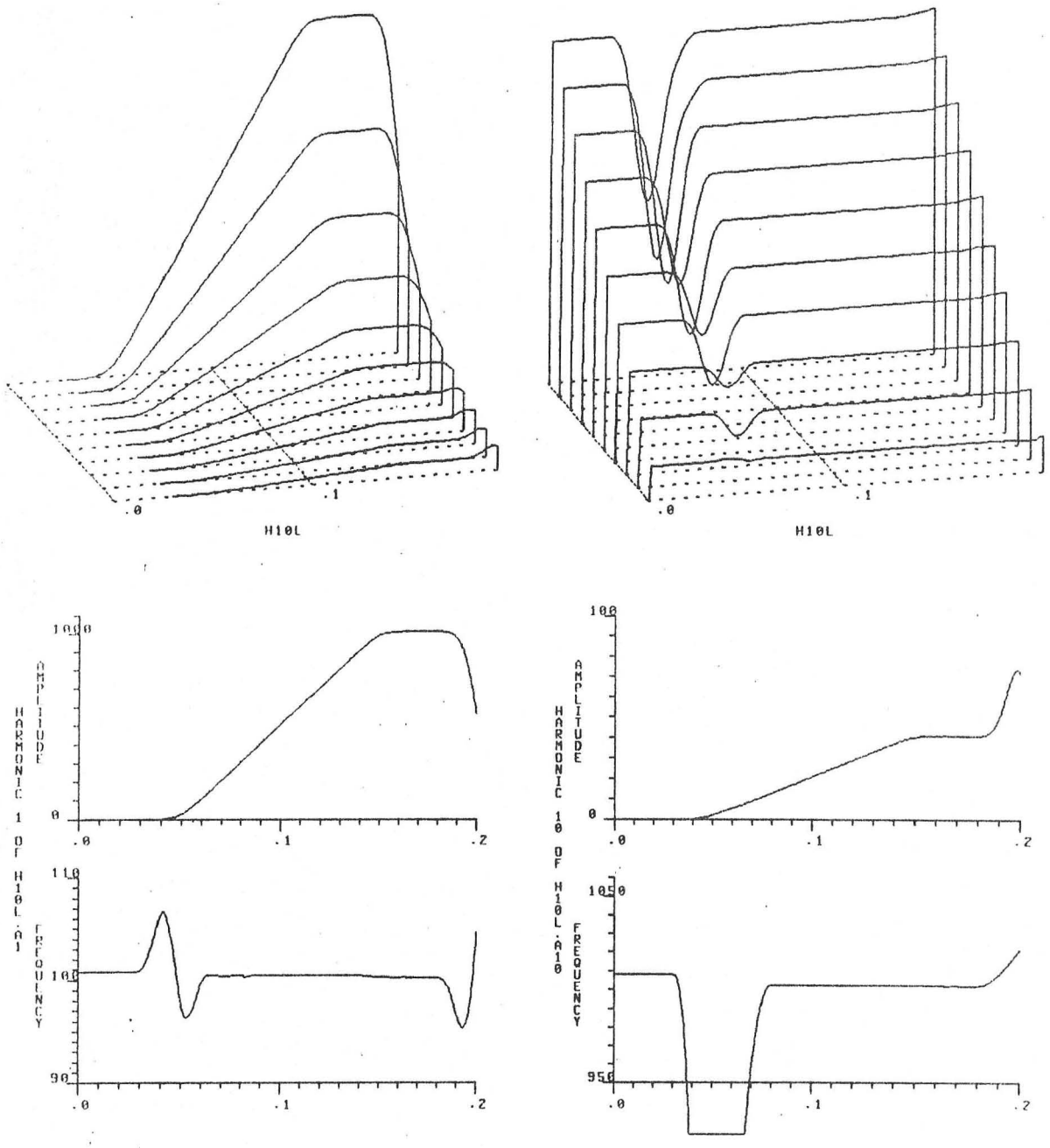


FIGURE A2. This shows the output of the heterodyne filter for a synthesized input signal which consists of a 100 Hz signal with a 100 millisecond linear attack on each harmonic. Each harmonic is 70% of the amplitude of the previous harmonic.

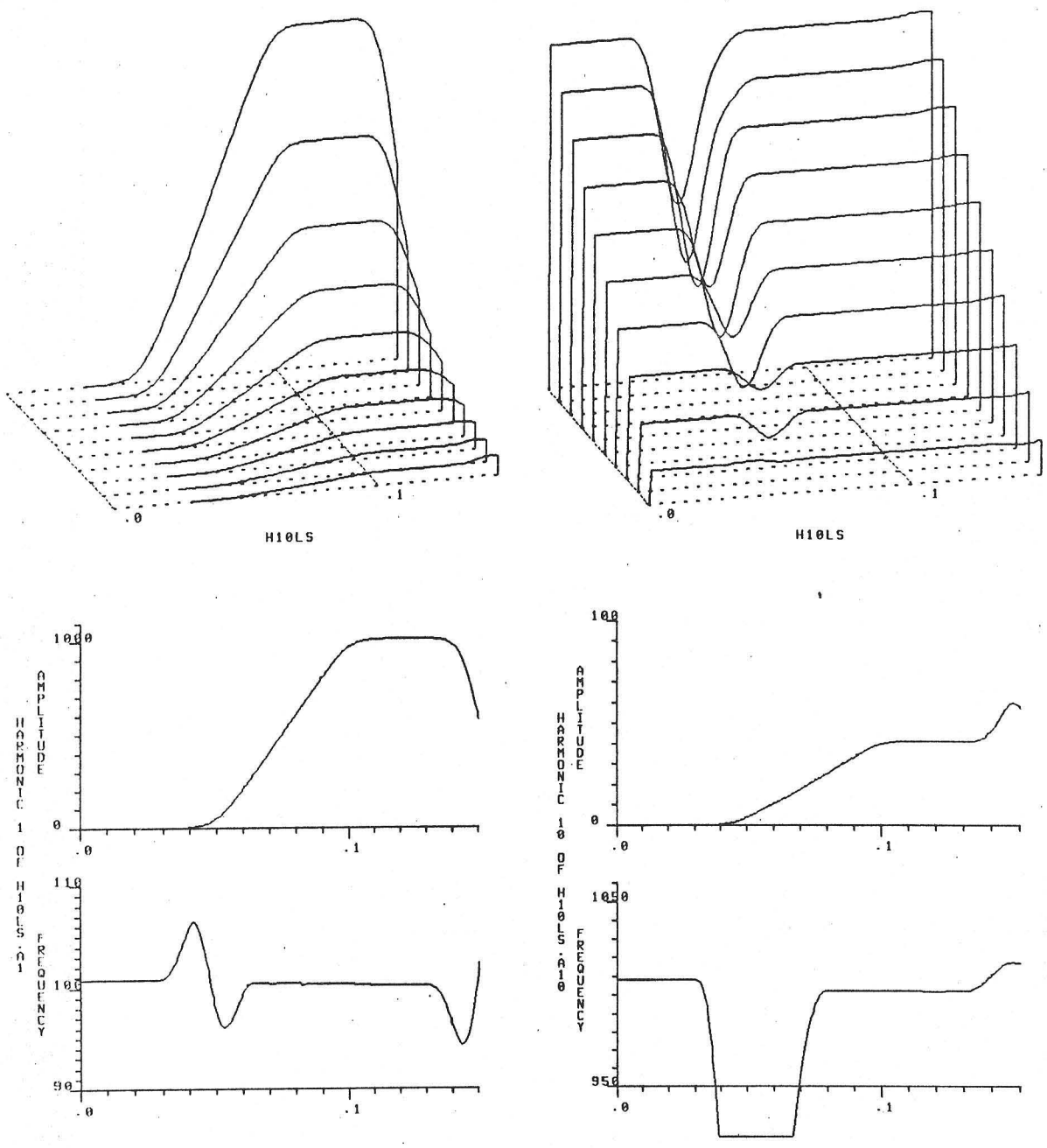


FIGURE A3. This shows the output of the heterodyne filter for a synthesized input signal which consists of a 100 Hz signal with a 50 millisecond linear attack on each harmonic. Each harmonic is 70% of the amplitude of the previous harmonic. Since this is a slightly faster attack than the previous figure, we see the attack portion of the frequency curves is somewhat more distorted.

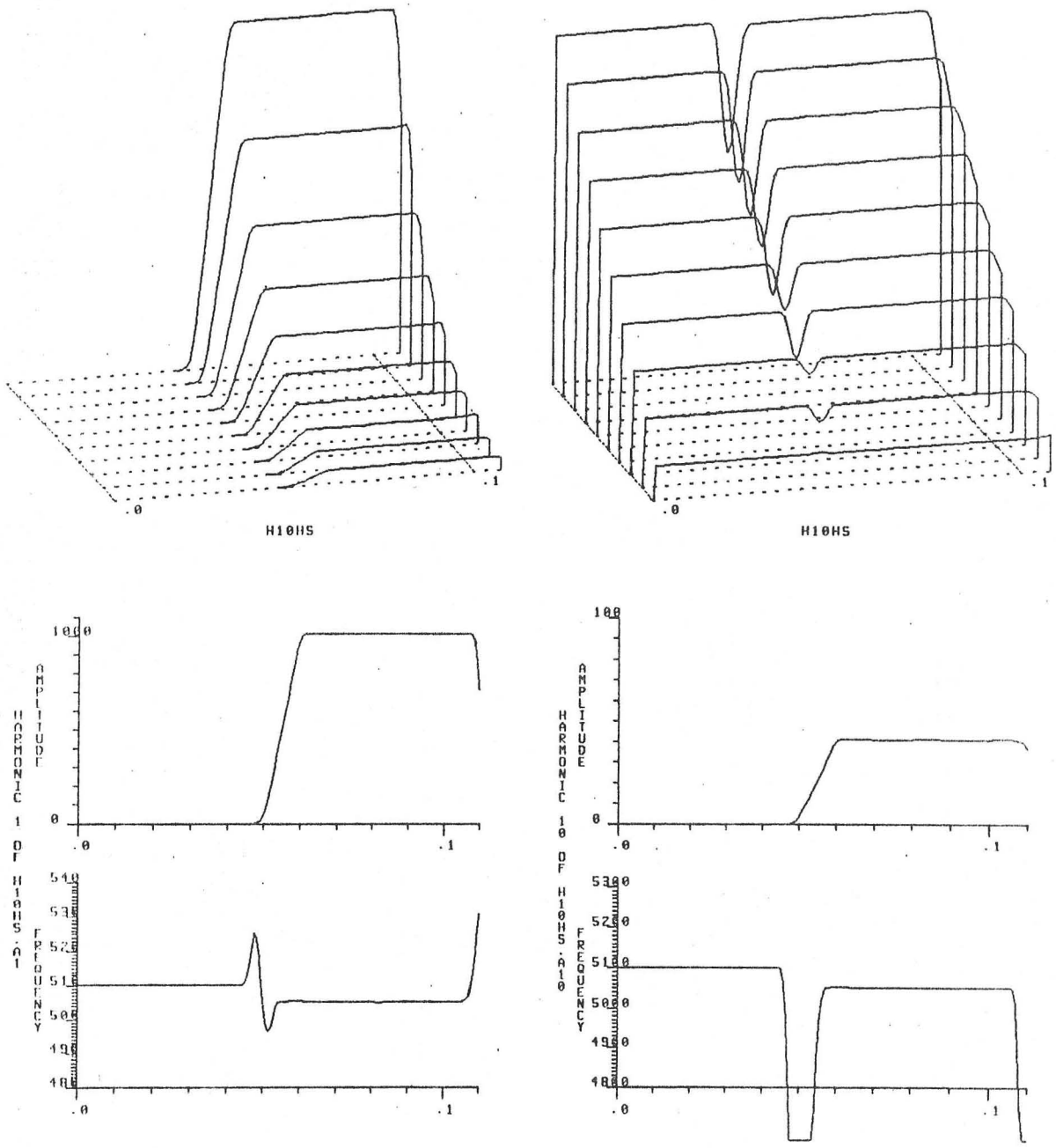


FIGURE A4. This shows the output of the heterodyne filter for a synthesized input signal which consists of a 505 Hz signal with a 10 millisecond linear attack on each harmonic. Each harmonic is 70% of the amplitude of the previous harmonic. The attack portion of the tone lasts only 5 periods, which is quite fast. As we would expect, the frequency trace for the first few periods is not accurate.

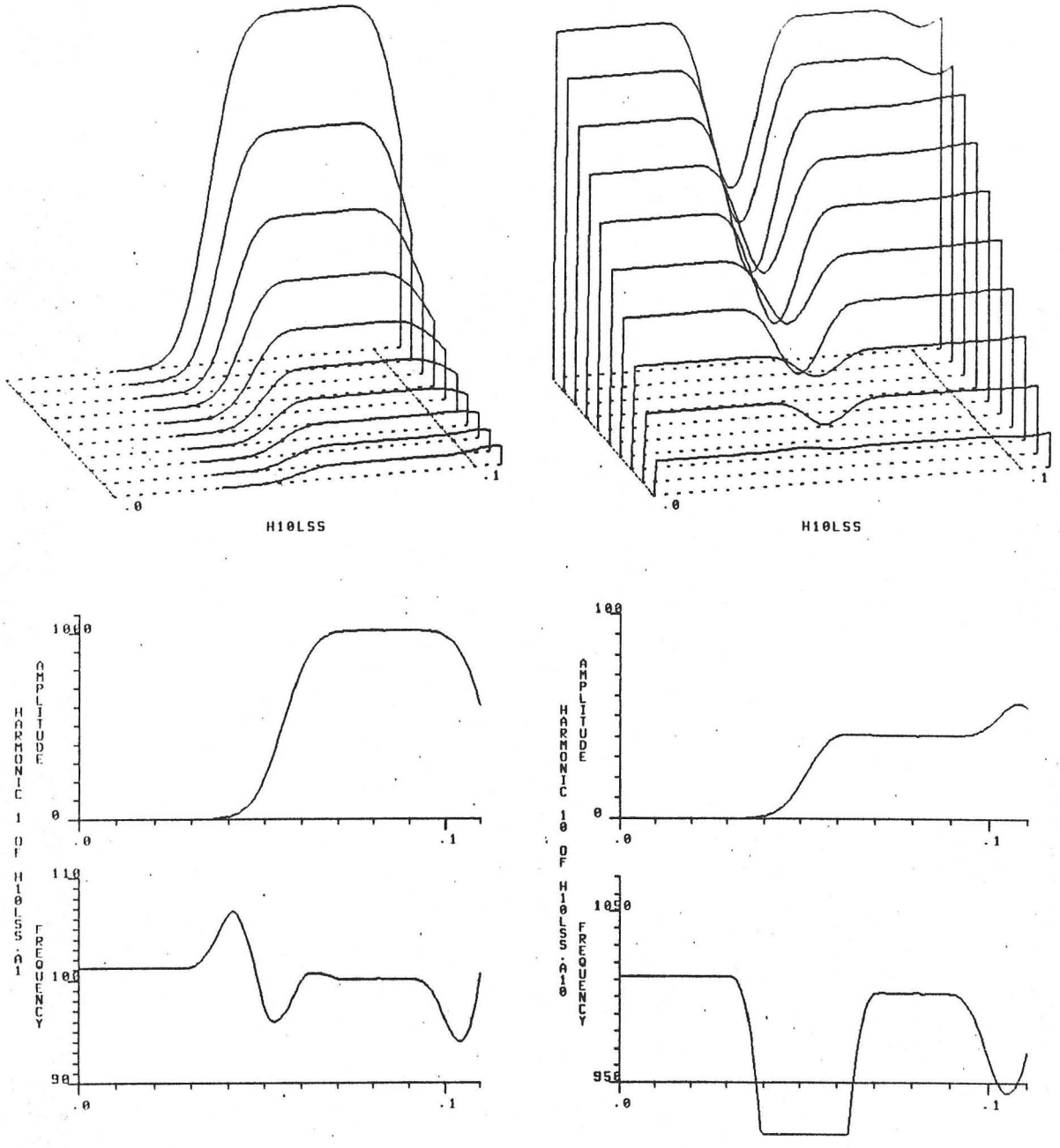


FIGURE A5. This shows the output of the heterodyne filter for a synthesized input signal which consists of a 100 Hz signal with a 10 millisecond linear attack on each harmonic. Each harmonic is 70% of the amplitude of the previous harmonic. The attack portion of the tone lasts only 1 period, which is extremely fast. The heterodyne filter cannot track the frequency during the attack portion and throughout some of the steady-state portion. The amplitude curves, however, are not grossly in error.

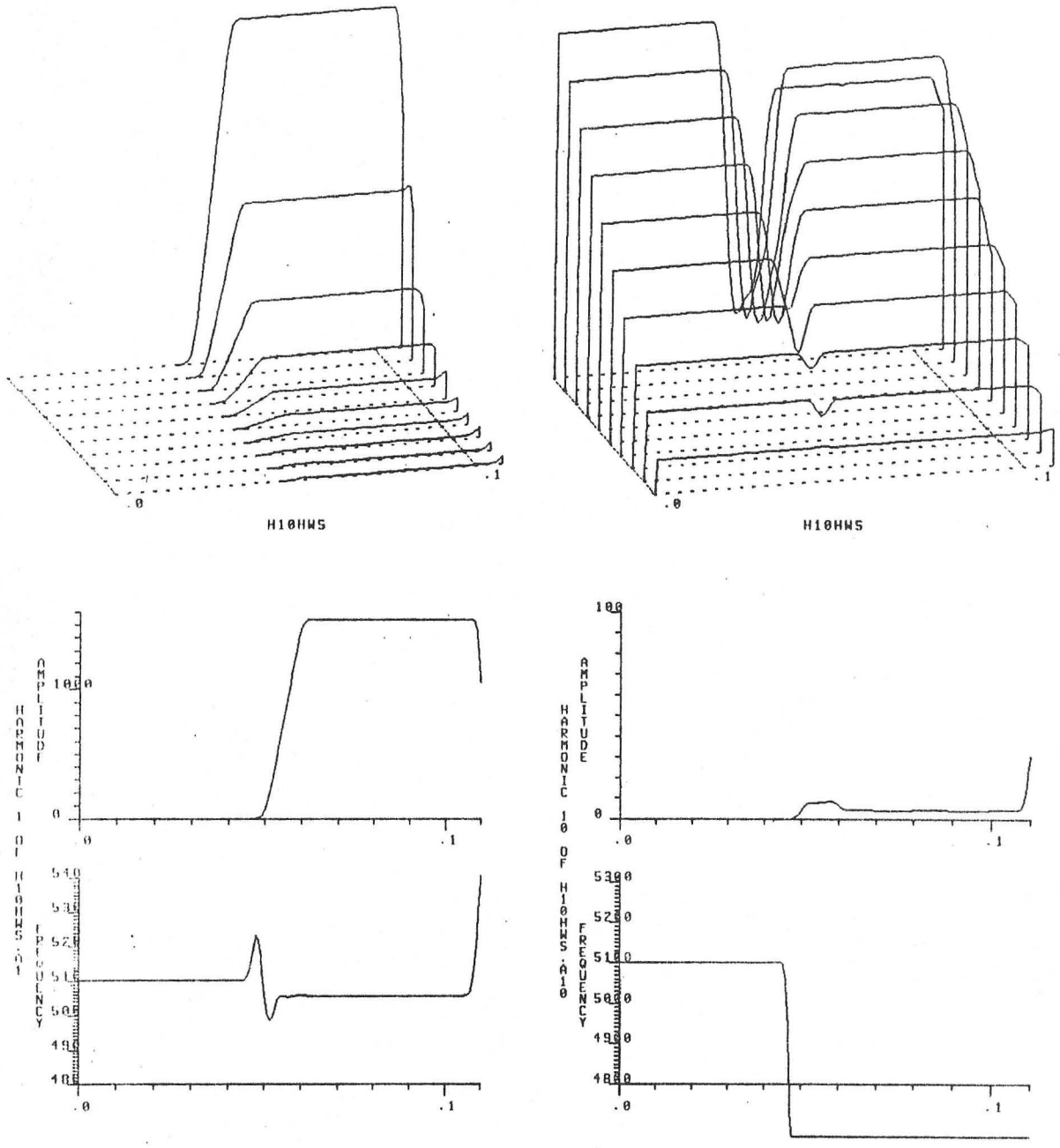


FIGURE A6. This shows the output of the heterodyne filter for a synthesized input signal which consists of a 505 Hz signal with a 10 millisecond linear attack on each harmonic. Each harmonic is 50% of the amplitude of the previous harmonic. This case is similar to figure A4, but the higher harmonics are much weaker. In fact, the 10th harmonic is so weak that its frequency cannot be successfully tracked. The plots are correct, however, up to the 9th harmonic. The relative amplitudes of the harmonics in this case are as follows, setting the amplitude of the first harmonic to 1 for convenience: 1.0, 0.5, 0.25, 0.125, 0.063, 0.031, 0.016, 0.008, 0.004, 0.002. Thus, the amplitude of the 10th harmonic is only 1/20th of the amplitude of the 10th harmonic in the previous figures.

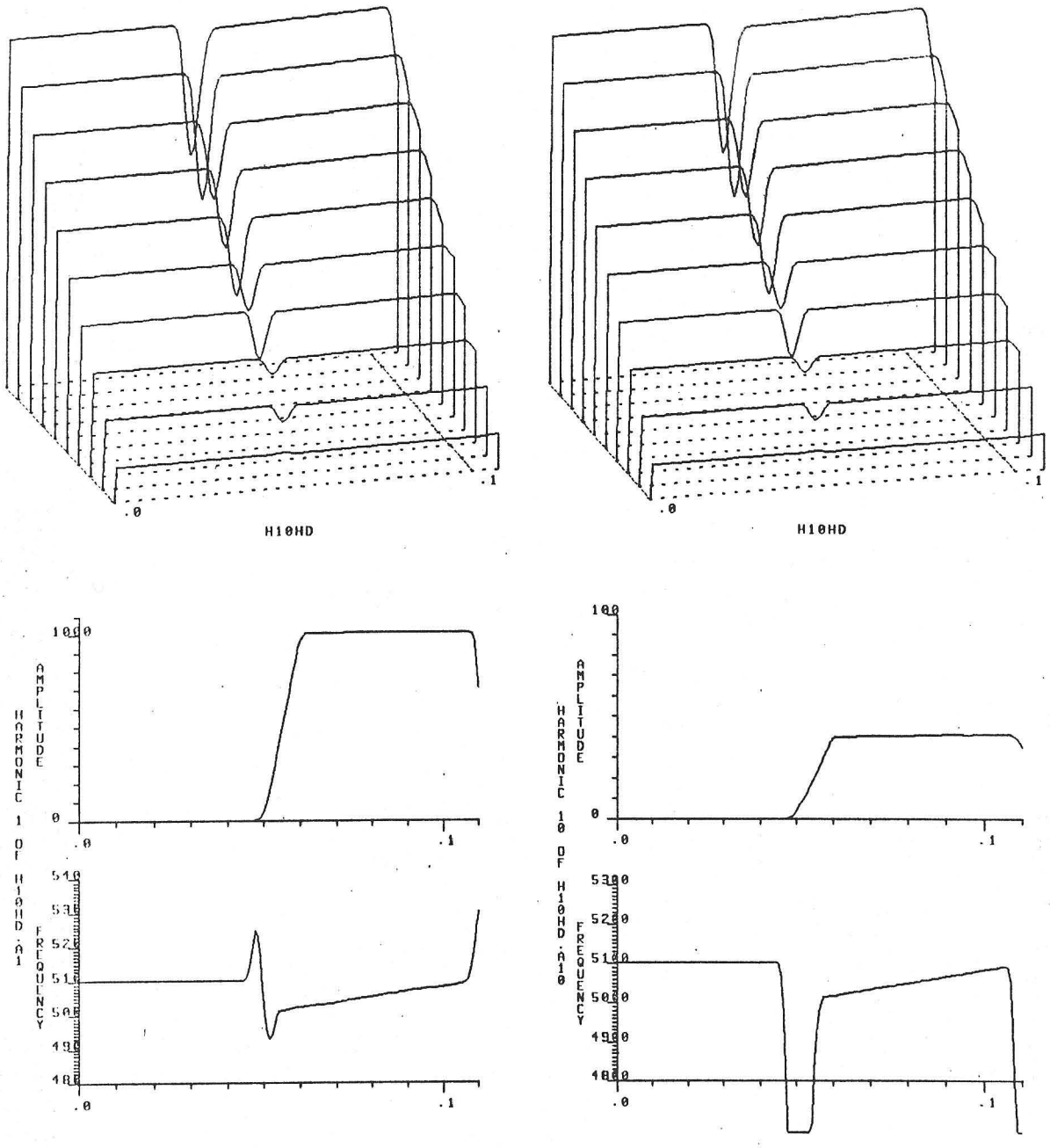


FIGURE A7. This shows the output of the heterodyne filter for a synthesized input signal which consists of a 500 Hz signal with a 10 millisecond linear attack on each harmonic. Each harmonic is 70% of the amplitude of the previous harmonic. Here we slew the frequency of the note from 500 Hz to 505 Hz over the duration of the tone. This is a 1% change, less than a quarter-step. Even in this case, the heterodyne filter seems to track acceptably.

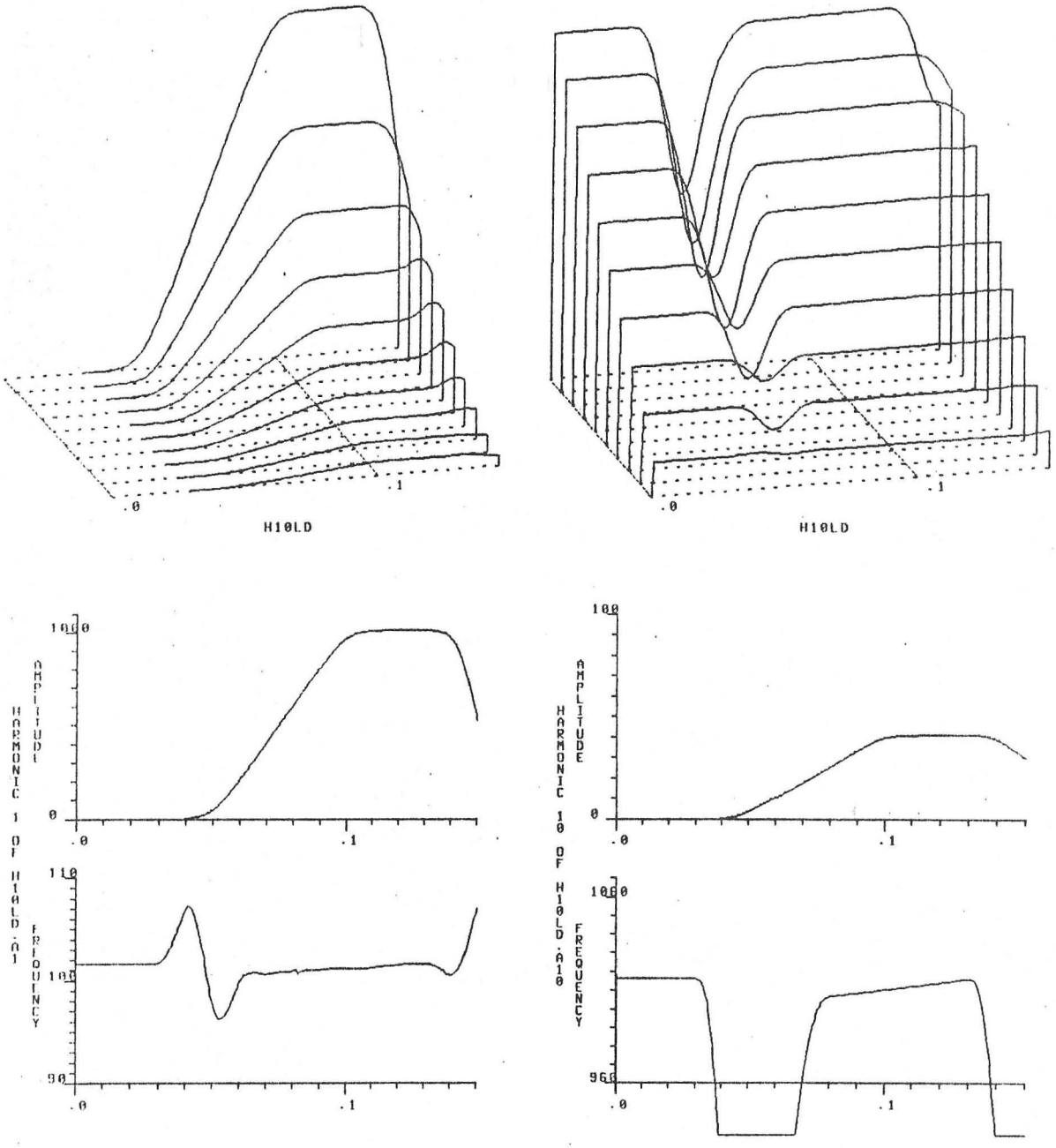


FIGURE A8. This shows the output of the heterodyne filter for a synthesized input signal which consists of a 100 Hz signal with a 50 millisecond linear attack on each harmonic. Each harmonic is 70% of the amplitude of the previous harmonic. Here we slew the frequency of the note from 100 Hz to 101 Hz over the duration of the tone. This is a 1% change, less than a quarter-step. As in the previous figure, the heterodyne filter seems to track acceptably.

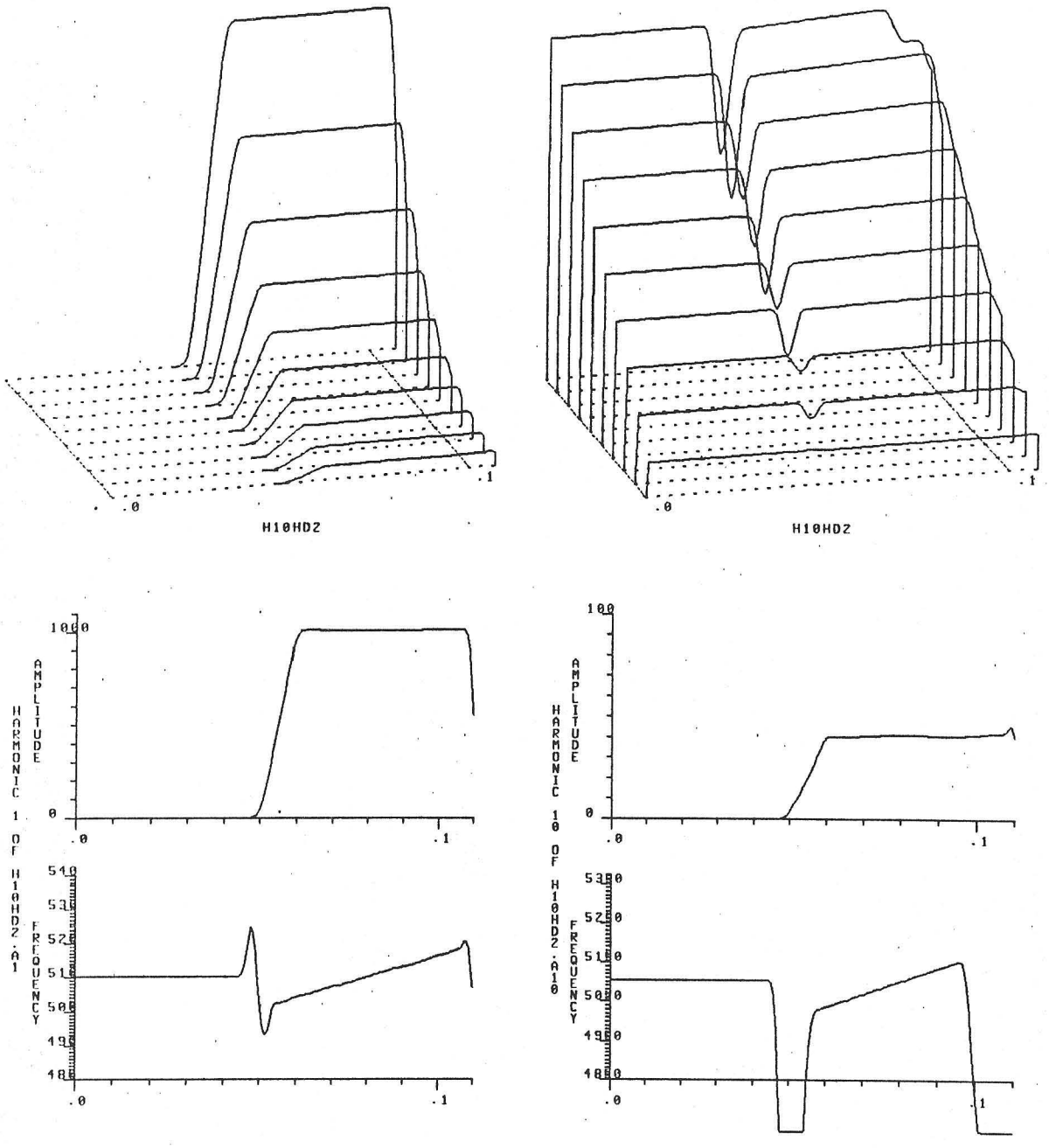


FIGURE A9. This shows the output of the heterodyne filter for a synthesized input signal which consists of a 500 Hz signal with a 10 millisecond linear attack on each harmonic. Each harmonic is 70% of the amplitude of the previous harmonic. Here we slew the frequency of the note from 500 Hz to 510 Hz over the duration of the tone. This is a 2% change, slightly less than a quarter-step. This seems to be about the limit of the allowable frequency change. Some of the frequency traces for the higher harmonics are not tracking properly.

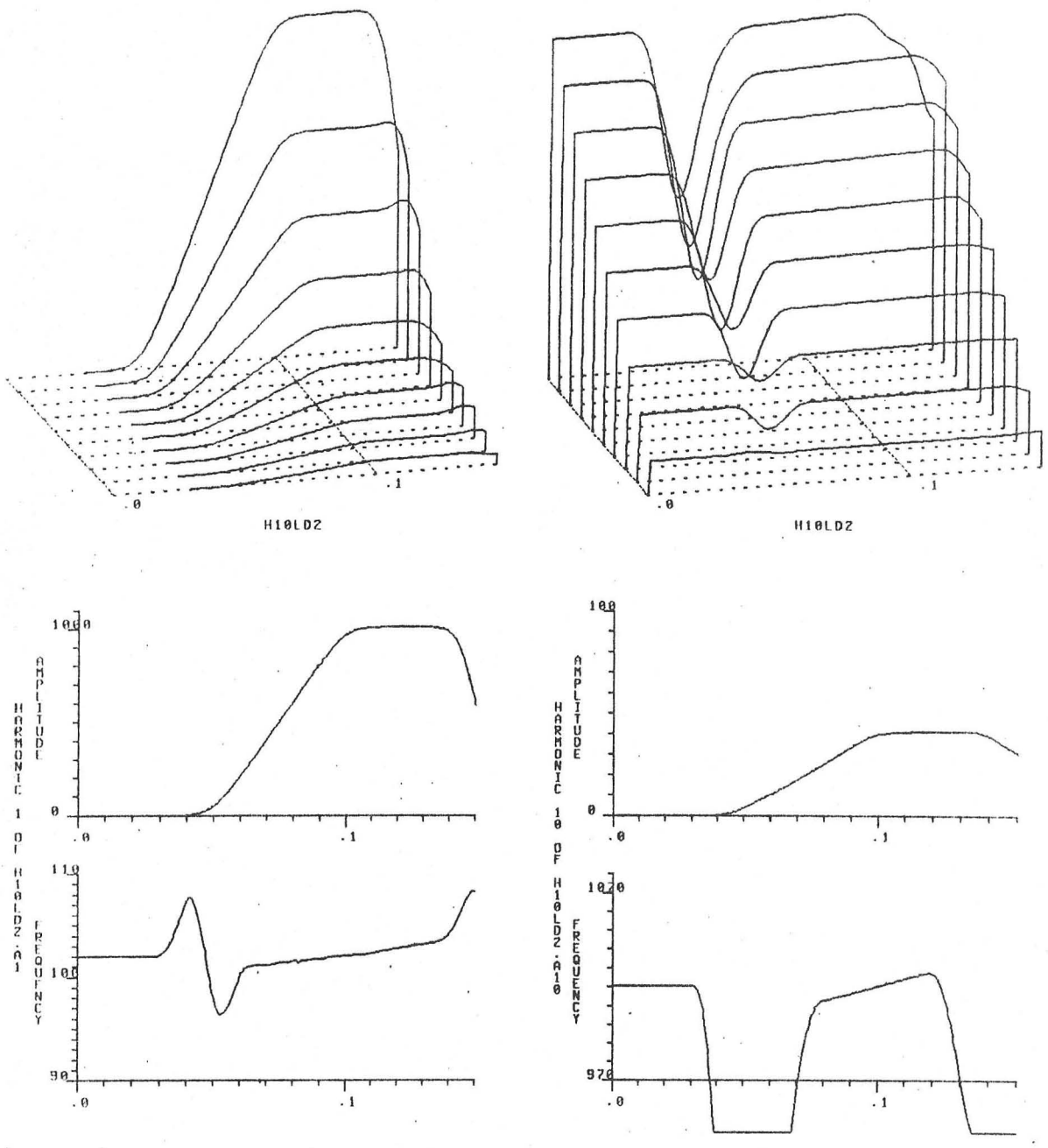


FIGURE A10. This shows the output of the heterodyne filter for a synthesized input signal which consists of a 100 Hz signal with a 50 millisecond linear attack on each harmonic. Each harmonic is 70% of the amplitude of the previous harmonic. Here we slew the frequency of the note from 100 Hz to 102 Hz over the duration of the tone. This is a 2% change, slightly less than a quarter-step. This seems to be about the limit of the allowable frequency change. Some of the frequency traces for the higher harmonics are not tracking properly.

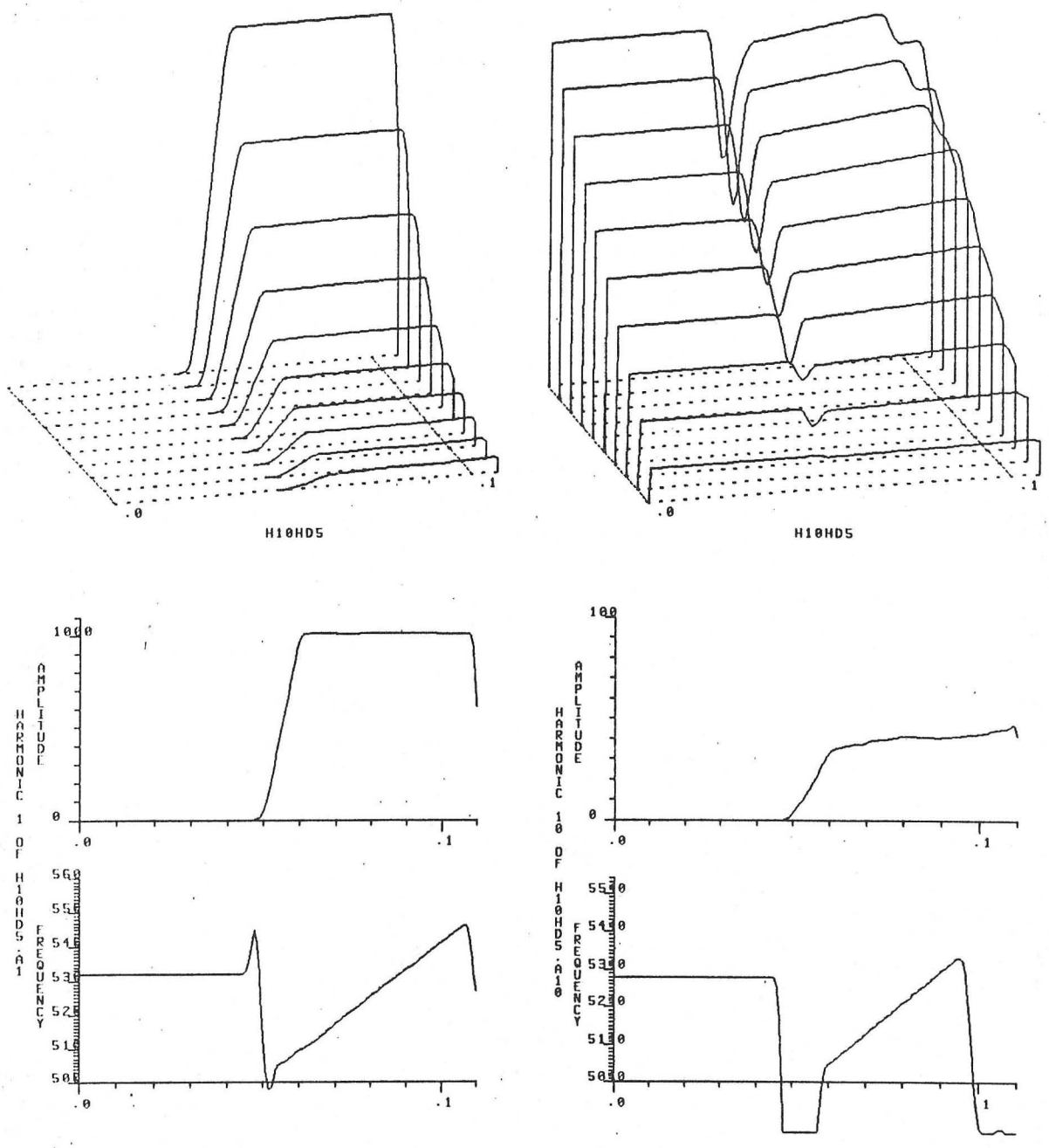


FIGURE A11. This shows the output of the heterodyne filter for a synthesized input signal which consists of a 500 Hz signal with a 10 millisecond linear attack on each harmonic. Each harmonic is 70% of the amplitude of the previous harmonic. Here we slew the frequency of the note from 500 Hz to 525 Hz over the duration of the tone. This is a 5% change, almost a half-step. This exceeds the bounds that the heterodyne filter can accept. Note that at the 10th harmonic, it drops down and starts tracking the 9th harmonic.

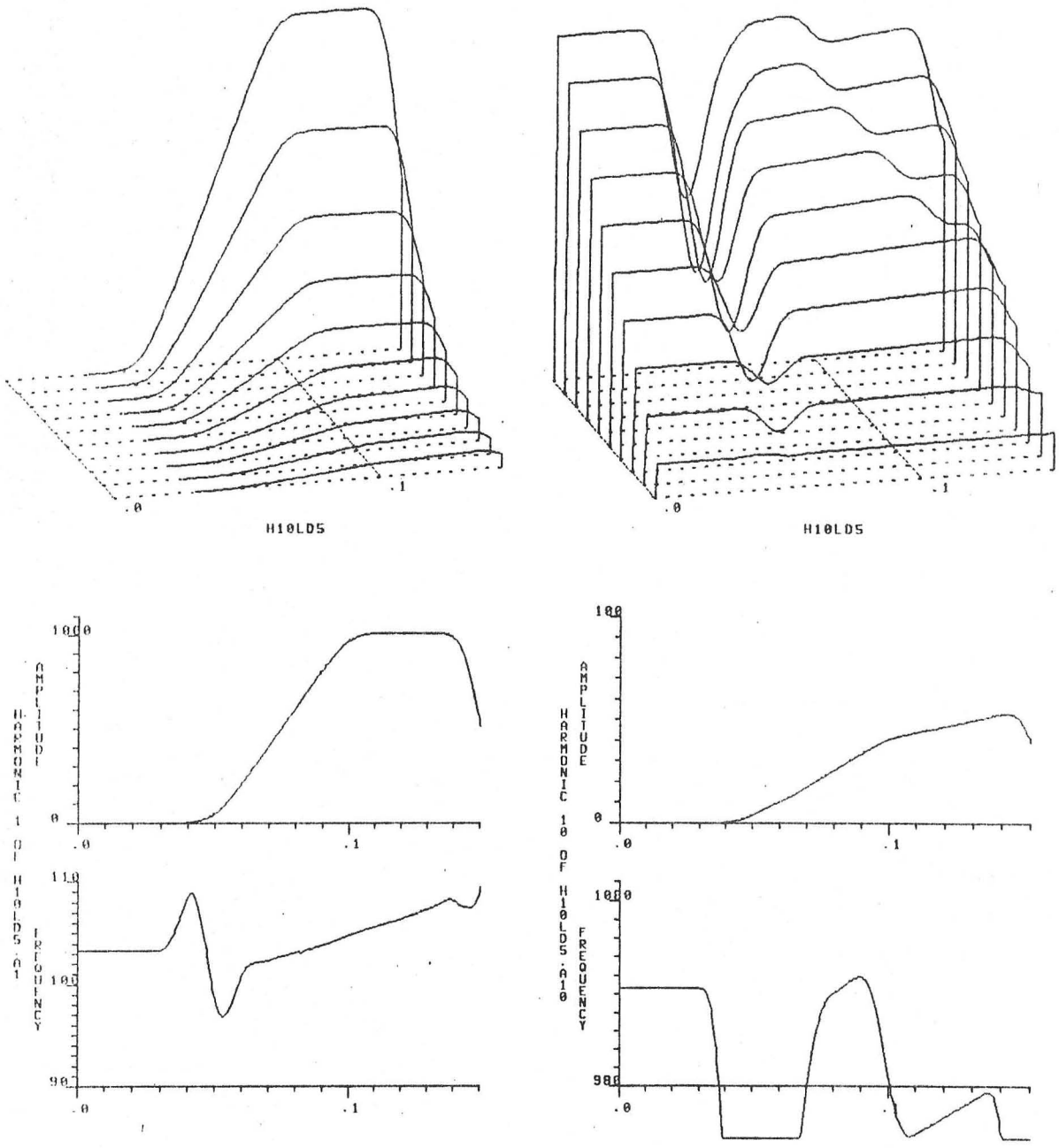


FIGURE A12. This shows the output of the heterodyne filter for a synthesized input signal which consists of a 100 Hz signal with a 50 millisecond linear attack on each harmonic. Each harmonic is 70% of the amplitude of the previous harmonic. Here we slew the frequency of the note from 100 Hz to 105 Hz over the duration of the tone. This is a 5% change, almost a half-step. This exceeds the bounds that the heterodyne filter can accept. Note that at the 10th harmonic, it drops down and starts tracking the 9th harmonic. Notice that amplitude distortion is beginning also.

In figures A3 and A4, we shorten the attack time to exactly 5 periods. This causes the frequency trace to lag behind the amplitude curve. In figure A3, we see that for the 10^{th} harmonic, the frequency curve is about 25 milliseconds late in stabilizing. In figure A5, the attack time has been shortened to one period. This is an extreme case and causes the frequency trace to be greatly in error, especially in the higher harmonics. In figure A6, we see the case where each harmonic is only 50 percent as great as the previous harmonic. Here, the 10^{th} harmonic is so weak that it cannot be traced at all. It is a typical form of behavior for the frequency curve to drop down to the frequency of the next lower harmonic when the amplitude of the harmonic is too weak.

In figures A7 through A12, we experiment with changing the frequency of the tone while the analysis proceeds. Figures A7 and A8 show a 1 percent change through the note, figures A9 and A10 show a 2 percent change, and figures A11 and A12 show a 5 percent change. We can see the failure start to set in in the 10^{th} harmonics with a 2 percent change. With a 5 percent change, the top several harmonics do not track properly, especially with the lower tone. When the frequency deviates this far, we can no longer guarantee absence of "leakage" between adjacent harmonics.

IMPLEMENTATION

There are several things that can be done to simplify the computation of the heterodyne filter. The first is to use "sliding" summations rather than computing the entire summation at each point. This is an old and well known trick that has great use here. The only problem is the accumulation of roundoff error. Although not included in the program that follows, one feature that was included in our own program was resetting all the sums every 1000 samples.

In converting the phase angle at each sample into a continuous phase function, it is somewhat difficult in the presence of noise to avoid occaisional jumps of multiples of π. Schafer [1969] gave an algorithm for "unwrapping" the phase in this manner. Unfortunately, his algorithm is not effective in the presence of large amounts of noise. Our approach has been to use the angle sum and difference formulae to compute not the angle, but the *difference* of the angle with the angle at the previous sample point. This works as follows:

(A1) $$\sin(\phi_{n\alpha}) \leftarrow \frac{a_{n\alpha}}{\sqrt{a_{n\alpha}^2 + b_{n\alpha}^2}}$$

(A2) $$\cos(\phi_{n\alpha}) \leftarrow \frac{b_{n\alpha}}{\sqrt{a_{n\alpha}^2 + b_{n\alpha}^2}}$$

(A3) $$\sin(\Delta\theta) \leftarrow \sin(\phi_{n\alpha})\cos(\phi_{n,\alpha-1}) - \cos(\phi_{n\alpha})\sin(\phi_{n,\alpha-1})$$

(A4) $$\sin(\Delta\theta) \leftarrow \cos(\phi_{n\alpha})\cos(\phi_{n,\alpha-1}) + \sin(\phi_{n\alpha})\sin(\phi_{n,\alpha-1})$$

(A5) $$\Delta\theta \leftarrow \tan^{-1}\left(\frac{\sin(\Delta\theta)}{\cos(\Delta\theta)}\right)$$

(A6) $$\theta_{n\alpha} \leftarrow \theta_{n,\alpha-1} + \Delta\theta$$

Where $a_{n\alpha}$ is the *real* part of the heterodyne filter output at the α^{th} point, for the n^{th} harmonic, as shown in equation (21) in the text,

$b_{n\alpha}$ is the *imaginary* part of the heterodyne filter output at the α^{th} point, for the n^{th} harmonic, as shown in equation (22) in the text,

$\theta_{n\alpha}$ is the phase angle at the α^{th} point, for the n^{th} harmonic, subject to the initial conditions $\theta_{n0}=0$,

$\phi_{n\alpha}$ is the *principal value* of the phase angle, $\theta_{n\alpha}$, at the α^{th} point, for the n^{th} harmonic, and

$\Delta\theta$ is $(\theta_{n\alpha}-\theta_{n,\alpha-1})$, the difference of the phase angles of this point and the previous point, as computed by the sine sum and difference formulae.

This may look like a succession of tautologies, but the result is a nice continuous phase with few discontinuities. The only jumps occur where the amplitude goes to near zero, where the phase is then just the phase of the noise, which is, of course, random.

This method gives, in general, a much smoother phase than Schafer's method.

A HETERODYNE FILTER PROGRAM

```
BOOLEAN PROCEDURE HET(INPUT,AMP,FREQ,CLOCK,FUND,HARMONIC,
        AVWIDTH,NSMOOTHS,N,M);
REFERENCE REAL ARRAY INPUT,AMP,FREQ;
VALUE REAL CLOCK,FUND;
VALUE INTEGER HARMONIC,AVWIDTH,NSMOOTHS,N;
REFERENCE INTEGER M;
BEGIN
```

COMMENT This program takes an array of sound samples in INPUT of length N (INPUT[1:N]), the fundamental frequency of the tone, FUND, the sampling rate in samples per second, CLOCK, the number of the harmonic under analysis, HARMONIC, the number of smoothings, NSMOOTHS, the width of the window used to compute the slope of the phase, AVWIDTH, and returns the amplitude of the harmonic as a function of time, AMP, and the frequency of the harmonic as a function of time, FREQ, and the number of valid points in AMP and FREQ, M. M is set to the input data length, N, minus the length of the period of the fundamental frequency in samples. A typical call might be HET(I,A,F,20000,155,3,25,3,10000,M). This would take from array I, put the amplitude in A, the frequency in F, sampling rate would be 20000 samples per second, the fundamental frequency would be 155 Hz., we would analyse for the 3rd harmonic (465 Hz.), would average over 25 points for the frequency curve, would do 3 smoothings, would take 10000 points (.5 seconds) out of I, and would place the number of output points in M;

```
INTEGER PERIOD;
REAL DANGLE,ANGLE,CS,SN,LCS,LSN;
REAL SUMT,SUMT2,SUMF,SUMFT,TIME;
REAL TIMINC,TSAMP,HFREQ;
REAL CSUM,SSUM,TEMP,PI,TWOPI;
REAL ARRAY FSAVE,FTSAVE[1:AVWIDTH],SINTAB[0:5000];
INTEGER I,J,K,L,INDEX;
LABEL EXIT;
```

COMMENT At first, we merely set up some constants and then load the sine table. This table could, of course, be set up once and for all beforehand, rather than be set up each time. Further, the table could be set up using the sine recursion formula at one multiply per point rather than calling the sine routine (generally 7 multiplies);

```
        PERIOD←CLOCK/FUND;
        BEGIN
            REAL ARRAY SNSAVE,CSSAVE[1:PERIOD+1];
            FUND←CLOCK/PERIOD;
            PI←3.1415926536;
            TWOPI←2*PI;
            ANGLE←0;
            DANGLE←5000*HARMONIC*FUND/CLOCK;
            FOR I←0 STEP 1 UNTIL 5000 DO
                SINTAB[I]←SIN(TWOPI*I/5000);
COMMENT The sine table should be computed beforehand,
    just once for all the harmonics;
            CSUM←0;
            SSUM←0;

            IF PERIOD+AVWIDTH<N THEN
            BEGIN
                HET←TRUE;
                GO TO EXIT;
            END;
```

COMMENT Here we actually do the heterodyne filter. It consists of multiplying the input signal by the sine and the cosine of the frequency of analysis (HARMONIC*FUND) and averaging over one period of the fundamental frequency. This is done by a sliding average. SN and CS represent the SINE and COSINE at the expected frequency of the harmonic (HARMONIC*FUND). We keep the sum of the input stream times the SINE in SSUM, and the sum of the input stream times the COSINE in CSUM. SNSAVE and CSSAVE are just to avoid doing a multiply to update SSUM and CSUM.;

```
J←1;
FOR I←1 STEP 1 UNTIL N DO
BEGIN
    INDEX←ANGLE;
    SN←SINTAB[INDEX];
    INDEX←INDEX+1250;
    IF INDEX≥5000 THEN INDEX←INDEX-5000;
    CS←SINTAB[INDEX];
    ANGLE←ANGLE+DANGLE;
    IF ANGLE≥5000.0 THEN ANGLE←ANGLE-5000.0;
    IF I>PERIOD THEN
    BEGIN
        CSUM←CSUM-CSSAVE[J];
        SSUM←SSUM-SNSAVE[J];
        COMMENT Subtract off the point past the
            end of the window. This saves doing
            the entire summation each time;
    END;
    CSSAVE[J]←INPUT[I]*CS;
    CSUM←CSUM+CSSAVE[J];
    SNSAVE[J]←INPUT[I]*SN;
    SSUM←SSUM+SNSAVE[J];
    IF I>PERIOD THEN
    BEGIN
        AMP[I-PERIOD]←CSUM;
        FREQ[I-PERIOD]←SSUM;
    END;
    J←J+1;
    IF J>PERIOD THEN J←1;
END;
```

COMMENT Now we smooth the curves by averaging over a window around the period of the fundamental. This places a new zero of transmission at each harmonic except the one under analysis. Generally, three smoothings are recommended. Quite often, unacceptable ripple will be present in the output without these smoothings. These smoothings are to be preferred over a standard low-pass filter because they place an explicit zero of transmission at the other harmonics. The variable L below denotes the width of the average. It starts out at 0 and grows to LENGTH. This means that it takes one period for the average to get started, which means that you will not get zeros of transmission at the other harmonics until the smoother has a chance to "warm up". If you have N smoothings, you must wait N periods for good results. Each tone to be analysed should have several periods of silence around it to get these filters started;

```
M←N-PERIOD;
FOR K←1 STEP 1 UNTIL NSMOOTHS DO
BEGIN
    LENGTH←PERIOD+(K MOD 3)-1;
    COMMENT We filter at the period, the period
        plus one sample, and the period minus
        one sample. This is a "shotgun" approach
        to help when the frequency is slightly
        different from what we expect it to be;
    J←1;
    L←0;
    SSUM←0;
    CSUM←0;
    FOR I←1 STEP 1 UNTIL M DO
    BEGIN
        IF I>LENGTH THEN
        BEGIN
            SSUM←SSUM-SNSAVE[J];
            CSUM←CSUM-CSSAVE[J];
        END
        ELSE L←L+1;
        COMMENT L is the width of the averaging
            interval, 1≤L≤LENGTH;
        SSUM←SSUM+AMP[I];
        CSUM←CSUM+FREQ[I];
        SNSAVE[J]←AMP[I];
        CSSAVE[J]←FREQ[I];
        COMMENT We must save copies of the
            inputs to the smoothing routines
            because we overwrite these
            quantities in the next steps;
        AMP[I]←SSUM/L;
        FREQ[I]←CSUM/L;
        J←J+1;
        IF J>LENGTH THEN J←1;
    END;
END;
```

COMMENT Now we convert to magnitude and phase form. To assure that the phase remains continuous, even during noisy parts, we compute the change in the angle by the difference-of-sines formula. We keep around the SINE and COSINE from the previous step and produce the angle increment by the arctangent of the difference in angle from the last sample to this one. Here we assume a procedure of value REAL which takes the arctangent of a number which is a fraction. We assume, then, that ATAN(NUM/DEN)=ATAN2(NUM,DEN), except that the case DEN=0 is handled properly in ATAN2 (that is, it returns plus or minus π/2, depending on the quadrant). We enter with AMP and FREQ containing the quadrature components of the harmonic. When we exit this section, AMP contains the amplitude of the harmonic and FREQ contains the phase of the harmonic.;

```
LSN←AMP[1];
LCS←FREQ[1];
AMP[1]←SQRT(LSN↑2+LCS↑2);
LSN←LSN/AMP[1];
LCS←LCS/AMP[1];
COMMENT LCS and LSN will be the cosine and sine
    of the phase angle at the previous sample;
FREQ[1]←ATAN2(LSN,LCS);
FOR I←2 STEP 1 UNTIL M DO
BEGIN
    SN←AMP[I];
    CS←FREQ[I];
    AMP[I]←SQRT(SN↑2+CS↑2);
    SN←SN/AMP[I];
    CS←CS/AMP[I];
    COMMENT This makes SN and CS the sine and
        cosine of the phase angle at this point;
    NUM←SN*LCS-CS*LSN;
    DEN←CS*LCS+SN*LSN;
    COMMENT NUM and DEN are the sine and cosine
        respectively of the difference between
        the phase angle of the previous sample
        and the phase angle of this sample, as
        computed by the angle sum and
        difference formulae;
    FREQ[I]←FREQ[I-1]+ATAN2(NUM,DEN);
    LCS←CS;
    LSN←SN;
END;
```

COMMENT Now we compute the frequency from the phase by getting the slope of the phase. We do this, adding some additional smoothing in the process, by computing a least-squares fit of a straight line to the phase and using the slope of this line at each point as the difference of the actual frequency and the expected frequency of the harmonic. Again, the sums are computed by sliding averages.;

```
            SUMT←0;
            SUMT2←0;
            SUMF←0;
            SUMFT←0;
            TIME←0;
            TIMINC←AVWIDTH/CLOCK;
            TSAMP←1/CLOCK;
            HFREQ←HARMONIC*FUND;
            J←1;
            L←0;
            FOR I←1 STEP 1 UNTIL M DO
            BEGIN
                IF I>AVWIDTH THEN
                BEGIN
                    TEMP1←TIME-TIMINC;
                    SUMT←SUMT+TEMP1;
                    SUMT2←SUMT2+TEMP1↑2;
                    SUMF←SUMF-FSAVE[J];
                    SUMFT←SUMFT-FTSAVE[J];
                END ELSE L←L+1;
                SUMT←SUMT+TIME;
                SUMT2←SUMT2+TIME↑2;
                SUMF←SUMF+FREQ[I];
                TEMP1←FREQ[I]*TIME;
                SUMFT←SUMFT+TEMP1;
                FSAVE[J]←FREQ[I];
                FTSAVE[J]←TIME1;
                TIME←TIME+TSAMP;
                IF I≤2 THEN FREQ[I]←HFREQ
                ELSE FREQ[I]←HFREQ+
                    (L*SUMFT-SUMT*SUMF)/
                        ((L*SUMT2-SUMT↑2)*TWOPI);
                J←J+1;
                IF J>AVWIDTH THEN J←1;
            END;
        END;
        HET←FALSE;
EXIT:
END;
```

APPENDIX B:
ON DESIGNING DIGITAL FILTERS

INTRODUCTION

During the course of this thesis, digital filters of many different varieties were used. Since the basis of the low-level processing is the bandpass filter, it was important to have a way of designing digital bandpass filters very quickly. The only closed-form solutions for filter coefficients that are currently known are the classical analog designs, like the Chebychev, Butterworth, Lagrange, Bessel, and others. In this method, we first design a low-pass filter, and then transform it to get high-pass, bandpass, or bandstop filters. We chose to do this transformation in the continuous domain. The analog filter is then transformed to the digital domain by use of the bilinear transform. Of course, the 3dB frequencies must have been already 'warped' before transformation to digital.

PROCEDURE

We, of course, will not attempt to review all of analog circuit design theory here. Two appropriate references are Guillemin [1957] or Karni [1966]. Neither will we review the bilinear transform for the generation of discrete filters from continuous. For this information, see Oppenheim and Schafer [1975] or Rabiner and Gold [1975]. What we would like to discuss are the details of what we feel to be a convenient, stable technique for numerically evaluating the coefficients. All of the processing is done in factored form, that is, all the roots are kept separately as complex numbers. For an N^{th} order filter, we will have N such numbers. When we go to bandpass or bandstop, there will then be $2N$ such numbers, for each root in the original low-pass design will generate two roots in the bandpass or bandstop case.

Each of these filters accept the following as design information: the frequency of the 3dB point (in the bandpass and bandstop cases, the frequencies of both 3dB points are required), the order of the filter (in bandpass and bandstop cases, this number will be doubled), and the type of the filter. Currently, only Butterworth and Chebychev at .5 dB ripple, 1 dB ripple, 2dB ripple and 3dB ripple are allowed. It is a simple matter to add other kinds.

LOWPASS AND HIGHPASS

These are the simplest cases. For the lowpass, we just take the continuous filter design directly. For the highpass, we merely invert the roots. This is simply done by dividing the conjugate of the root by its magnitude squared. Remembering that it is highpass, we go directly to the digital conversion. Both filters are designed with their 3dB point at 1. They must be scaled to the proper frequency. This is done simply by multiplying all the roots by that frequency.

BANDPASS AND BANDSTOP

These are the most interesting cases, for each original root must create two roots in the transformed filter. This is done by means of the following transformations:

$$(B1) \quad p \rightarrow \frac{(s^2+\omega_0^2)}{s}$$

for the bandpass case and for the bandstop case:

$$(B2) \quad p \rightarrow \frac{s}{(s^2+\omega_0^2)}$$

Where p is the complex frequency variable of the original low-pass design.

s is the complex frequency variable of the transformed filter.

ω_0 is the geometric mean of the 3dB frequencies of the desired bandpass or bandstop filter

We can see what this does to each pole of the original design by just substituting the complex frequency of the original pole as p in the above equations and solving for s;

$$(B3) \quad s = \frac{A + \sqrt{A^2-4\omega_0^2}}{2}, \quad \frac{A - \sqrt{A^2-4\omega_0^2}}{2}$$

$$(B4) \quad s = \frac{1 + \sqrt{1-4A^2\omega_0^2}}{2}, \quad \frac{1 - \sqrt{1-4A^2\omega_0^2}}{2}$$

Where A is the complex frequency of one of the original low-pass poles.

Again, (B3) is for the bandpass case and (B4) is for the bandstop case. To compute these numbers, we may use arctangents and do it in magnitude-angle formulation, but we have found that the Cartesian coordinates give slightly more accuracy. To perform the complex square root, all we need to do is compute the square root of the length of $A^2-4\omega_0^2$ and compute the SINE and COSINE of one-half the angle of $A^2-4\omega_0^2$. This can be done as follows:

$$(B5) \quad C \leftarrow \frac{\alpha}{\sqrt{\alpha^2+\beta^2}}$$

$$(B6) \quad C_2 \leftarrow \sqrt{\frac{1+C}{2}}$$

$$(B7) \quad S_2 \leftarrow \mathrm{sgn}(\beta)\sqrt{\frac{1-C}{2}}$$

$$(B8) \quad k \leftarrow \sqrt{\sqrt{\alpha^2+\beta^2}}$$

$$(B9) \quad \sqrt{A^2-4\omega_\theta^2} = k(C_2+jS_2)$$

Where α is the real part of $A^2-4\omega_\theta^2$,
β is the imaginary part of $A^2-4\omega_\theta^2$,
C is then the cosine of the angle of $A^2-4\omega_\theta^2$,
C_2 is then the cosine of half the angle of $A^2-4\omega_\theta^2$,
S_2 is then the sine of half the angle of $A^2-4\omega_\theta^2$,
k is the magnitude of the square root of $A^2-4\omega_\theta^2$.
$\mathrm{sgn}(\beta)$ is +1 if $\beta \geq 0$ and −1 if $\beta < 0$

This is shown for the bandpass case, but may also be done for the bandstop case similarly.

TRANSFORMATION TO DISCRETE

After the transformation to the proper kind of filter, we may inspect for stability just by examining the real parts of the filter. We have found the filters designed this way all have negative real parts as high as 20th order.

We then group the conjugate poles together for lump-transformation to a digital second-order section. The remaining real pole, if any, will be transformed into a first-order section. We can also order the poles according to Q for what is hoped to produce minimum roundoff error.

After the transformation, we can normalize the response so that certain frequencies have a

magnitude transfer function of 1.0. For the low-pass and bandstop cases, we wish 0 frequency to be passed with gain 1.0. For the high-pass case, it is infinite frequency (π in discrete domain). For the bandpass case, it is ω_0, the geometric mean of the 3dB frequencies. We can get this scale factor by computing it as we go along, or by computing it at the end of all the transformations. It is simple to compute at the end and is guaranteed to give the correct results, so this is what was used in our program. We merely predict the transfer function at the critical frequency and multiply the first filter section input terms by the inverse of the predicted transfer function.

This completes the design of the filter. It is realized in cascade form as the conjunction of many second-order sections and possibly a single first-order section.

BIBLIOGRAPHY

B.S. Atal, M.R. Schroeder, **Adaptive Predictive Coding of Speech Signals**, Bell Systems Technical Journal, Vol 49, pp1973-1986

B.S. Atal, S.L. Hanauer, **Speech Analysis and Synthesis by Linear Prediction of the Speech Wave**, J. Acoust. Soc. Am., Vol 50, pp637-655, February 1971

H. Backhaus, **Uber Geigenklange**, Zeitschrift fur Technische Physik, Vol 8, p510, 1927

H. Backhaus, **Uber die Bedeutung der Ausgleichsvorgange in der Austrik**, Zeitschrift fur Technische Physik, Vol 13, #1, pp31-46, 1932

Sidney Bertram, **Frequency Analysis Using the Discrete Fourier Transform**, IEEE Trans Audio and Electroacoustics, Vol AU-18, #4, Dec. 1970

J.W. Beauchamp, **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

J.W. Beauchamp, personal communication, summer 1974

G. von Békésy, **Zür Theorie des Hörens; die Schwingungsform der Basilarmembran**, Phys. Z., Vol 29, pp793-810, 1928

G. von Békésy, **Ueber die Nichtlinearen Verzerrungen des Ohres**, Ann. Phys., Vol 20, pp809-827, 1934

G. von Békésy, **Hearing Theories and Complex Sounds**, J. Acoust. Soc. Amer., Vol 35, pp588-601, 1963

Frans A. Bilsen, **Repetition Pitch; Its Implication for Hearing Theory and Room Acoustics**, in R. Plomp, G.F. Smoorenburg, eds., *Frequency Analysis and Periodicity Detection in Hearing*, A.W. Sijthoff, Leiden, the Netherlands, pp250-266, 1970

C. Bingham, M.D. Godfrey, J.W. Tukey, **Modern Techniques of Power Spectral Estimation**, IEEE Trans. Audio Electroacoustics, Vol AU-15, #2, pp91-98, June 1967

R.B. Blackman, J.W. Tukey, **The Measurement of Power Spectra**, New York, Dover, 1959

R.B. Blackman, **Linear Data Smoothing and Prediction in Theory and Practice**, Reading, Mass.: Addison-Wesley, 1965

E. de Boer, **On the Residue in Hearing**, Doctoral Dissertation, University of Amsterdam, 1956

B.P. Bogert, M.J.R. Healy, J.W. Tukey, **The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum and Saphe Cracking**, in *Proceedings of the Symposium on Time Series Analysis*, M. Rosenblatt, ed., John Wiley and Sons, Inc., New York, Chapter 15, pp209-243, 1963

Steven Frank Boll, **A Priori Digital Speech Analysis**, PhD Thesis, University of Utah report UTEC-CSc-73-123, March 1973

P. Broome, **A Frequency Transformation for Numerical Filters**, Proc. IEEE, Vol 52, #2, pp326-327, February 1966

William T. Cochran, James W. Cooley, David L. Favin, Howard D. Helms, Reg A. Kaenel, William W. Lang, George C. Maling Jr., David E. Nelson, Charles M. Rader, Peter D. Welch, **What is the Fast Fourier Transform**, Proc. IEEE, Vol 55, #10, pp1664-1674, October 1967

D.K. Faddeev, V.N. Faddeeva, **Computational Methods of Linear Algebra**, English translation by R.C. Williams, W.H. Freeman, San Francisco, 1973

Harvey Fletcher, E. Donnell Blackham, Richard Stratton, **Quality of Piano Tones**, J. Acoust. Soc. Amer., Vol 34, #6, pp749-761, June 1962

Allen Forte, **Tonal Harmony in Concept and Practice**, Holt, Rinehart, and Winston; New York, 1962

M.D. Freedman, **A Technique for Analysis of Musical Instrument Tones**, PhD dissertation, University of Illinois, Urbana. 1965

M.D. Freedman, **Analysis of Musical Instrument Tones**, J. Acoust. Soc. Am., Vol 41, p793, 1967

M.D. Freedman, **A Method For Analysing Musical Tones**, J. Audio Eng. Soc., Vol 16, #4, p419, October 1968

John E. Freund , **Mathematical Statistics**, Prentice-Hall, Inc., Englewood Cliffs, N.J., 390p, 1962

W.W. Gentleman, G. Sande, **Fast Fourier Transforms for Fun and Profit**, 1966 Fall Joint Computer Conference, AFIPS Proc., Vol 29, 1966

Bernard Gold, **Computer Program for Pitch Extraction**, J. Acoust. Soc. Amer., Vol 34, p916, 1962

Bernard Gold, Charles M. Rader, **Digital Processing of Signals**, McGraw-Hill Book Company, New York, 1969

Bernard Gold, Lawrence R. Rabiner, **Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain**, J. Acoust. Soc. Amer, Vol 46, #2, p442, 1969

Julius L. Goldstein, **An Optimum Processor Theory for the Central Formation of the Pitch of Complex Tones**, J. Acoust. Soc. Amer., Vol 54, #6, pp1496-1516, 1973

John M. Grey, **An Exploration of Musical Timbre**, PhD Dissertation, Department of Psychology, Stanford University, 1975

L.J. Griffiths, **Rapid Measurement of Digital Instantaneous Frequency**, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol ASSP-23, #2, pp207-221, April 1975

Ernst A. Guillemin, **Synthesis of Passive Networks**, John Wiley & Sons, Inc., New York, 741p, 1957

R.W. Hamming, **Numerical Methods for Scientists and Engineers**, New York: McGraw-Hill, 1962

C.M. Harris, M.R. Weiss, **Pitch Extraction by Computer Processing of High Resolution Fourier Analysis Data**, J. Acoust. Soc. Amer., Vol 35, p339, 1963

H.L.F. von Helmholtz, **Die Lehre von den Tonempfindungen als Physiologische Grundlage für die Theorie der Musik**, F. Vieweg & Sonn, Braunschweig, 1st Edition, 1863

L.A. Hiller, C. Bean, **Information Theory Analysis of Four Sonata Expositions**, J. of Music Theory, Vol 10, #1, pp96-138, 1966

L.H. Hiller, R. Fuller, **Structure and Information in Webern's Symphonie, Op. 21**, J. of Music Theory, Vol 11, #1, pp67, 1967

Adrian J.M. Houtsma, **What Determines Musical Pitch**, J. of Musical Theory, Vol 15, #1, pp138-157, 1971

Adrian J.M. Houtsma, Julius L. Goldstein, **The Central Origin of the Pitch of Complex Tones: Evidence from Musical Interval Recognition**, J. Acoust. Soc. Amer., Vol 51, #2(Part 2), pp520-529, 1972

F. Itakura, S. Saito, **Analysis Synthesis Telephony Based on the Maximum Liklihood Method**, in *Report of the 6th International Congress on Acoustics*, Y. Kohasi, ed., pp C17-C20, Paper C-5-5, August 1968

F. Itakura, S. Saito, **A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies**, Electron. Commun. Japan, Volume 53-A, Number 1, pp36-43, 1970

F. Itakura, S. Saito, **Digital Filtering Techniques for Speech Analysis and Synthesis**, in *Conf. Rec., 7th Int. Congr. Acoustics*, Paper 25 C 1, 1971

R. Jackson, **The Computer as a 'Student' of Harmony**, Tenth Congress of the International Musicological Society, 1967

Shlomo Karni, **Network Theory: Analysis and Synthesis**, Allyn and Bacon, Inc., Boston, 483p, 1966

J.F. Kaiser, **Design Methods for Sampled-Data Filters**, Proc. First Allerton Conference on Circuit and System Theory, pp221-236, November 1963

J.S. Keeler, **The Attack Transients of Some Organ Pipes**, IEEE Tran. on Audio and Electroacoustics, Vol AU-20, #5, pp378-391, December 1972

J.S. Keeler, **Piecewise-Periodic Analysis of Almost-Periodic Sounds and Musical Transients**, IEEE Tran. on Audio and Electroacoustics, Vol AU-20, #5, pp338-344, December 1972

F.F. Kuo, J.F. Kaiser Eds., **System Analysis by Digital Computer**, New York: Wiley, Chapter 7, 1966

Y.W. Lee, **Statistical Theory of Communication**, John Wiley & Sons, Inc., New York, 509p, 1960

Paul R. Lehman, **Harmonic Structure of the Tone of the Bassoon**, J. Acoust. Soc. Amer., Vol 36, #9, pp1649-1653, September 1964

N. Levinson, **The Wiener RMS Error Criterion in Filter Design and Prediction**, J. Math. Phys., Vol 25, #4, pp261-278, 1947. also in N. Wiener *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, MIT Press, Cambridge, Mass., 1966

H.C. Longuet-Higgins, M.J. Steedman, **On Interpreting Bach** in *Machine Intelligence VI* Bernard Meltzer and Donald Michie Editors, Edinburgh University Press, Edinburgh, pp221-241, 1971

David A. Luce , **Physical Correlates of Nonpercussive Musical Instrument Tones**, PhD thesis, MIT, 1963

David Luce, Melville Clark, **Duration of Attack Transients of Nonpercussive Orchestral Instruments**, J. Audio Eng. Soc., Vol 13, #3, p194, 1965

David Luce, Melville Clark, **Physical Correlates of Brass Instrument Tones**, J. Acoust. Soc. Am, Vol 42, p1232, 1967

J. Makhoul, J.J. Wolf, **Linear Prediction and the Spectral Analysis of Speech**, Bolt Beranek and Newman, Inc., Cambridge, Mass. Report number 2304, August 1972

John Makhoul, **Linear Prediction: A Tutorial Review**, Proceedings of the IEEE, Vol 63, #4, pp561-580, April 1975

J.D. Markel, **Digital Inverse Filtering - A New Tool for Formant Trajectory Estimation**, IEEE Trans. on Audio and Electroacoustics, Vol AU-20, #2, pp129-137, June 1972

J.D. Markel, **The SIFT Algorithm for Fundamental Frequency Estimation**, IEEE Trans. on Audio and Electroacoustics, Vol AU-20, #5, pp367-377, December 1972

J.D. Markel, A.H. Gray Jr., **A Linear Prediction Vocoder Simulation Based Upon the Autocorrelation Method**, IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol ASSP-22, #2, pp124-134, April 1974

Neil Joseph Miller, **Filtering of Singing Voice Signal from Noise by Synthesis**, PhD Thesis, University of Utah, Dept. Of Electrical Engineering, March 1973

Neil Joseph Miller, **Pitch Detection by Data Reduction**, IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol ASSP-23, #1, pp72-78, February 1975

James Anderson Moorer, **The Optimum Comb Method of Pitch Period Analysis of Continuous Digitized Speech**, IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol ASSP-22, #5, pp330-338, October 1974

M.P. Mussorgsky, M. Ravel, **Tableaux D'une Exposition**, Boosey & Hawkes, London, England, 151p, 1929

A. Michael Noll, **Cepstrum Pitch Determination**, J. Acoust. Soc. Amer., Vol 41, #2, p293, 1967

G.S. Ohm, **Ueber die Definition des Tönes, nebst daran geknupfter Theorie der Sirene und ahnlicher tonbildenden Vorrichtungen**, Ann. Phys. Cham., Vol 59, pp513-565, 1843

G.S. Ohm, **Noch ein Paar Worte über die Definition des Tones**, Ann. Phys. Chem., Vol 62, pp1-18, 1844

A.V. Oppenheim, R.W. Schafer, **Homomorphic Analysis of Speech**, IEEE Trans. on Audio and Electroacoustics, Vol AU-16, #2, pp221-226, 1968

A.V. Oppenheim, **Speech Analysis-Synthesis System Based on Homomorphic Filtering**, J. Acoust. Soc. Am., Vol 45, pp458-465, February 1969

A.V. Oppenheim, R.W. Schafer, **Digital Signal Processing**, Prentice-Hall, Englewood Cliffs, New Jersey, 608p, 1975

Sam E. Parker, **Analyses of the Tones of Wooden and Metal Clarinets**, J. Acoust. Soc. Amer., Vol 19, #3, pp415-419, May 1947

Walter Piston, **Harmony**, W.W. Norton & Company, Inc., New York, 374p, 1941

Richard Plomp, **The Ear as a Frequency Analyzer**, J. Acoust. Soc. Amer., Vol 36, p1628-1636, 1964

Richard Plomp, **Experiments on Tone Perception**, Institute for Perception RVO-TNO, Soesterberg: Netherlands, 1966

R. Plomp and A.M. Mimpen, **The Ear as a Frequency Analyzer, II**, J. Acoust. Soc. Amer., Vol 43, p764-767, 1968

R. Plomp, **Pitch, Timbre and Hearing Theory**, Internat. Audiol., Vol 7, pp322-344, 1968

Lawrence R. Rabiner, Bernard Gold, **Theory and Application of Digital Signal Processing**, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 762 p, 1975

Charles M. Rader, Bernard Gold, **Digital Filter Design Techniques in the Frequency Domain**, Proc. IEEE, Vol 55, #2, pp149-171, February 1967

D.R. Reddy, **An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave**, PhD Dissertation, Department of Computer Science, Stanford University, 1966

D.C. Rife, G.A. Vincent, **Use of the Discrete Fourier Transform in the Measurement of Frequencies and Levels of Tones**, Bell Syst. Tech. J., February 1970

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

R.J. Ritsma, **Existence Region of the tonal Residue, I**, J. Acoust Soc. Amer., Vol 34, pp1224-1229, 1962

R.J. Ritsma, **Existence Region of the tonal Residue, II**, J. Acoust Soc. Amer., Vol 35, pp1241-1245, 1963

R.J. Ritsma, F.L. Engel, **Pitch of Frequency-Modulated Signals**, J. Acoust. Soc. Am., Vol 36, pp1637-1644, 1964

R.J. Ritsma, **Frequencies Dominant in the Perception of the Pitch of Complex Sounds**, J. Acoust Soc. Amer, Vol 42, pp191-199, 1967

R.J. Ritsma, **Periodicity Detection**, in R. Plomp, G.F. Smoorenburg (Eds.), *Frequency Analysis and Periodicity Detection in Hearing*, A.W. Sijthoff, Leiden, the Netherlands, pp250-266, 1970

Enders A. Robinson, **Statistical Communication and Detection**, Hafner Publishing Company, New York, 362p, 1967

M.J. Ross, H.L. Shaffer, A. Cohen, R. Freudberg, H.J. Manley, **Average Magnitude Difference Function Pitch Extractor**, IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol ASSP-22, #5, pp353-361, October 1974

A.P. Sage, J.L. Melsa, **System Identification**, Academic Press, New York, 1971

F.A. Saunders, **Analysis of the Tones of a Few Wind Instruments**, J. Acoust. Soc. Amer., Vol 18, #2, pp395-401, October 1946

R.W. Schafer, **Echo Removal by Discrete Generalized Linear Filtering**, PhD thesis, MIT, 1969. Also MIT Technical Report 466

R.W. Schafer, L.R. Rabiner, **A Digital Signal Processing Approach to Interpolation**, Proceedings of the IEEE, Vol 61, #6, pp692-702, June 1973

J.F. Schouten, **The Residue, a New Component in Subjective Sound Analysis**, Proc. Kon. Nederl. Akad. Wetensch, Vol 43, pp356-365, 1940

J.F. Schouten, **The Residue and the Mechanism of Hearing**, Proc. Kon. Nederl. Akad. Wetensch, Vol 43, pp991-999, 1940

J.F. Schouten, **The perception of Pitch**, Phillips Techn. Rev., Vol 5, pp286-294, 1940

J.F. Schouten, **Five Articles on the Perception of Sound (1938-1940)**, Instituut voor Perceptie Onderzoek, Eindhoven, The Netherlands, 1960

J.F. Schouten, R.J. Ritsma, B. Lopes Cardozo, **Pitch of the Residue**, J. Acoust. Soc. Am., Vol 34, #8, part 2, pp1418-1424, September 1962

J.F. Schouten, **The Residue Revisited**, in R. Plomp, G.F. Smoorenburg, eds., *Frequency Analysis and Periodicity Detection in Hearing*, A.W. Sijthoff, Leiden, the Netherlands, pp250-266, 1970

M.R. Schroeder, **Improved Quasi-sterophony and Colerless Artificial Reverberation**, J. Acous. Soc. Am., Vol 33, p1061, 1961

M.R. Schroeder, B.F. Logan, **Colerless Artificial Reverberation**, J. Audio Eng. Soc., Vol 9 #3, p192, July 1961

M.R. Schroeder, **Natural Sounding Artificial Reverberation**, J. Audio Eng. Soc., Vol 10, p219, 1962

M.R. Schroeder, K.H. Kuttruff, **On Frequency Response Curves in Rooms**, J. Acoust. Soc. Am., Vol 34, p76, 1962

M.R. Schroeder, A.M. Noll, **Recent Studies in Speech Research at Bell Telephone Laboratories**, Proc. 5th International Congr. Acoustics, Liege, 1965

M.R. Schroeder, B.S. Atal, G.M. Sessler, J.E. West, **Acoustic Measurements in Philharmonic Hall (New York)** J. Acoust. Soc. Am., Vol 40, pp431-434, February 1970

M.R. Schroeder, **Parameter Estimation in Speech: A Lesson in Unorthodoxy**, Proceedings of the IEEE, Vol 58, #5, pp707-712, May 1970

A. Seebeck, **Beobachtungen über einige Bedingungen der Entstehung von Tönen**, Ann. Phys. Chem., Vol 53, pp417-436, 1841

A. Seebeck, **Ueber die Sirene**, Ann. Phys. Chem., Vol 60, pp449-481, 1843

A. Seebeck, **Ueber die Definition des Tönes**, Ann. Phys. Chem., Vol 63, pp353-368, 1844

A. Seebeck, **Ueber die Erzeugung von Tönen durch getrennte Eindrucke, mit Beziehung auf die Definition des Tönes**, Ann. Phys. Chem., Vol 63, pp368-380, 1844

William M. Siebert, **Frequency Discrimination in the Auditory System: Place or Periodicity Mechanisms?**, Proceedings of the IEEE, Vol 58, #5, pp723-730, May 1970

Richard C. Singleton, **On Computing the Fast Fourier Transform**, Comm. of the ACM, Vol 10, #10, pp647-654, October 1967

Richard C. Singleton, **ALGOL Procedures for the Fast Fourier Transform**, Comm. of the ACM, Vol 11, #11, pp773-776, November 1968

Richard C. Singleton, **An ALGOL Procedure for the Fast Fourier Transform with Arbitrary Factors**, Comm. of the ACM, Vol 11, #11, pp776-779, November 1968

Richard C. Singleton, **An ALGOL Convolution Procedure Based on the Fast Fourier Transform**, Comm. of the ACM., Vol 12, #3, pp179-184, March 1969, also *Remark on Algorithm 345 . . .*, Comm. of the ACM, Vol 12, #10, Oct. 1969

Leland C. Smith, **Editing and Printing Music by Computer**, Journal of Music Theory, pp292-308, Fall 1973

Guido F. Smoorenburg, **Pitch of Two-Tone Complexes**, in R. Plomp, G.F. Smoorenburg, eds., *Frequency Analysis and Periodicity Detection in Hearing*, A.W. Sijthoff, Leiden, the Netherlands, pp250-266, 1970

Man Mohan Sondhi, **New Methods of Pitch Extraction**, IEEE Trans. on Audio and Electroacoustics, Vol AU-16, #2, June 1968

W. Strong, M. Clark, **Synthesis of Wind Instrument Tones**, J. Acoust. Soc. Amer., Vol 41, p39, 1967

W. Strong, M. Clark, **Perturbations of Synthetic Orchestral Wind Instrument Tones**, J. Acoust. Soc. Amer., Vol 41, p277, 1967

E. Terhardt, **Frequency Analysis and Periodicity Detection in the Sensations of Roughness and Periodicity Pitch**, in R. Plomp, G.F. Smoorenburg, eds., *Frequency Analysis and Periodicity Detection in Hearing*, A.W. Sijthoff, Leiden, the Netherlands, pp250-266, 1970

José Manuel Tribolet, **Identification of Linear Discrete Systems With Applications to Speech Processing**, MS dissertation, Department of Electrical Engineering, MIT, 1974

Pierre Vicens, **Aspects of Speech Recognition by Computer**, PhD Dissertation, Department of Computer Science, Stanford University, 1969

M.R. Weiss, C.M. Harris, **Computer Technique for the High-Speed Extraction of Speech Parameters**, J. Acoust. Soc. Amer., Vol 35, p208, 1963

Norbert Wiener, **Extrapolation, Interpolation, and Smoothin**, American Scientist, Vol 62, #2, pp208-215, March-April 1974

J.H. Wilkinson, C. Reinsch, **Linear Algebra**, Volume II, Springer-Verlag, New York, 1971

Terry Winograd, **Linguistics and the Computer Analysis of Tonal Harmony**, J. of Music Theory, Vol 12, #1, p2-49, 1968

Yoram Yakimovsky, **Scene Analysis Using a Semantic Base for Region Growing**, PhD Dissertation, Department of Computer Science, C.S. report number STAN-CS-73-380, Stanford University, June 1973