Center for Computer Research in Music and Acoustics May 1985

Department of Music
Report No. STAN-M-27

ON THE AUTOMATIC TRANSCRIPTION OF PERCUSSIVE MUSIC -- FROM ACOUSTIC SIGNAL TO HIGH-LEVEL ANALYSIS

by

W. Andrew Schloss

Research sponsored by

National Science Foundation
and
System Development Foundation

CCRMA
DEPARTMENT OF MUSIC
Stanford University
Stanford, California 94305

Center for Computer Research in Music and Acoustics May 1985

Department of Music
Report No. STAN-M-27

ON THE AUTOMATIC TRANSCRIPTION OF PERCUSSIVE MUSIC -- FROM ACOUSTIC SIGNAL TO HIGH-LEVEL ANALYSIS

by

W. Andrew Schloss

This dissertation is concerned with the use of a computer to analyze and understand rhythm in music. The research focuses on the development of a program that automatically transcribes percussive music, investigating issues of timing and rhythmic complexity in a rich musical setting. Beginning with a recording of an improvised performance, the intent is to be able to produce a score of the performance, to be able to resynthesize the performance in various ways, and also to make inferences about rhythmic structure and style.

In order to segment percussive sound from the given acoustic waveform, automatic slope-detection algorithms have been developed and implemented. Initially, a simple amplitude envelope is found by tracing the peaks of the waveform. This provides a data reduction of about 200:1 and is useful for obtaining an overview of the musical material. The data are then segmented by repeatedly performing a linear regression over a small moving window of the envelope data, moving the window one point at a time over the envelope. The linear regressions create a sequence of line segments that "float" over the data and allow segmentation by carefully set slope thresholds.

The slope threshold determines the attacks. Once the attacks are determined, the decay time-constant, *tau*, is determined by fitting a one-pole model to the amplitude envelope. From the value of *tau*, a decision can be made as to whether a given stroke is damped or undamped. This corresponds to the method of striking the drum. Once the damped/undamped decision is made, a portion of the original time waveform is sent to a "stroke-detector" that determines how the drum was struck in greater detail.

At this point, enough information about the performance has been obtained to begin a higher-level analysis. Given the timing information and the patterns of strokes, it is possible to track tempo automatically, and to try to make inferences about the meter. These two issues are in fact quite deep, and are the focus of a body of work that involves detection of "macro-periodicity," that is a repetition rate over longer periods of time than signal processing would normally yield. Also included in this thesis is an historical and theoretical overview of research on rhythm, from several perspectives.

This thesis was submitted to the Department of Hearing and Speech Sciences and the Committee on Graduate Studies of Stanford University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This research was supported by the National Science Foundation under Contract NSF MCS 80-12476 and MCS 82-14350 and System Development Foundation under Grant SDF #345. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of Stanford University, any agency of the U. S. Government, or of sponsoring foundations.

© Copyright 1985

by

W. Andrew Schloss

Acknowledgements

There are so many people to thank, without whom this work would not exist, that I have trouble presenting their names. Professor Earl Schubert, my advisor, would have to be first on the list. I have met few people who combine a truly open mind with a careful, critical mind in a way such as Dr. Schubert. Next, John Chowning, director of CCRMA (Center for Computer Research in Music and Acoustics), has created the unique interdisciplinary environment here, and the opportunity to work in it is priceless.

The two most important people in the daily work of the thesis were Julius Smith and Bernard Mont-Reynaud. Julius has been invaluable for his apparently infinite supply of ideas, explanations and wizardry in signal processing, and also for considerable help with TEX—without him I would have certainly stood still. Bernard Mont-Reynaud has provided numerous insights, and his leadership on the National Science Foundation grant that supported my work in part has been quite important. Bernard developed many of the ideas reported here, and had an enormous influence on the direction of the work.

Many other friends and coworkers at CCRMA have at some point rescued me from dark corners. In particular, David Jaffe, John Gordon, Tovar, Bill Schottstaedt, Chris Chafe, and John Strawn should be mentioned. Scott Foster and A.J. Rockmore from Systems Control, Inc., were collaborators during early work in signal processing.

I would also like to thank Professor Fredric Lieberman of the University of Washington for planting the seed that began this research. When I was a graduate student in ethnomusicology at the University of Washington (before coming to Stanford), I took his class entitled "Transcription and Analysis." It was a torturous experience, which first motivated me to dream of using a machine to do the work. Since that time, my horizons have widened considerably, but the ethnomusicological point of view has remained with me.

Lastly, I would like to thank my parents for their support and encouragement during the long road to completion.

This thesis is in memoriam to my dog, Squid.

Table of Contents

Introduction. 1

Chapter 1. Overview. 3
- 1.1. The Problem 3
- 1.2. Input to the System 5
- 1.3. Types of Output 6

Chapter 2. Background and Theory. 9
- 2.1. Introduction 9
- 2.2. Previous Research in Automatic Transcription 10
  - 2.2.1. The Melograph 10
  - 2.2.2. Moorer's Thesis 11
  - 2.2.3. Piszczalski and Galler 14
  - 2.2.4. The Auditory Transform 16
  - 2.2.5. Recent Work at CCRMA 17
- 2.3. Timing Studies 19
  - 2.3.1. Lower Level 20
  - 2.3.2. Perceptual Attack Time 23
  - 2.3.3. Higher Level—Timing in Musical Contexts 24
- 2.4. Rhythmic Structure and Meter 30
  - 2.4.1. Theoretical Studies 30
  - 2.4.2. Computer Simulation Studies 34
- 2.5. Steps Toward a Global Theory of Rhythm 37
  - 2.5.1. The Proposed Paradigm 38
  - 2.5.2. Anatomy of the Meso-period 44
- 2.6. Issues Peculiar to Percussive Music 47
  - 2.6.1. Acoustical Considerations 47
  - 2.6.2. Ensemble Considerations 50

Chapter 3. Methods. 57
- 3.1. Introduction 57
- 3.2. Approach to the Low-Level Analysis 58
  - 3.2.1. Envelope Derivation 58
  - 3.2.2. Slope Detection 61
  - 3.2.3. Segmentation Rules 61
  - 3.2.4. Setting Segmentation Parameters 63
  - 3.2.5. Perceptual Attack Time 65
  - 3.2.6. High-Pass Filtering to Facilitate Segmentation . . . 66
  - 3.2.7. Source-identification 72
  - 3.2.8. On Pitch Detection 86
- 3.3. Approach to the High-level Analysis 88
  - 3.3.1. Important Durations 89
  - 3.3.2. Accents and Anchor Points 91
  - 3.3.3. Rational Approximation to Metric Unit 93

3.3.4. Tempo Line 96
3.3.5. Tempo Line Refinement 97
3.3.6. Determining Normalized Rhythmic Values 99
3.3.7. The Musical Map 104

Chapter 4. Conclusions. 107
4.1. Summary 107
4.2. Implications 108
4.3. Future Research 109
4.3.1. Polyphony 109
4.3.2. Analysis of Style via Timing Studies 110
4.3.3. Synthesis of Percussive Sounds 110
4.3.4. Intelligent Editor of Musical Sound 111
4.3.5. Interactive Performance 111

Appendix A. Contents of Tape of Musical Examples. 113

References. 114

On The Automatic Transcription of Percussive Music
—From Acoustic Signal to High-Level Analysis

By
W. Andrew Schloss

Program in Hearing and Speech Sciences
Stanford University, Stanford, California 94305

In this dissertation, an example of a real improvised performance is transformed from the acoustic waveform to a (modified) Western notation representation. A myriad of decisions are made en route, some of which could be research topics in themselves. They are addressed here as they relate to the problems at hand.

An effort is made here not only to automate a process that well-trained musicians can already do, but to shed light on the many musical/acoustical/theoretical issues that are involved or invoked in this process. Thus, the broad motivation is a desire to understand the entire process better; the final result is not the only concern. This thesis represents an attempt not only to perform the task of automatic transcription (though this is not a trivial task), but also to use that process to lead to an understanding of rhythm perception and production, and to understand more fully the relationship between composer and performer.

The material within covers a great deal of territory and crosses numerous boundaries. It cannot be contained in a single discipline. Some material has been included that is peripheral to the main effort of transcription (parts of Chapter 2, for example) because one aim of this work is a better understanding of percussive music in general.

Those readers familiar with signal processing may not be familiar with ethnomusicology, a psychoacoustician might not have thought about rhythm, a drummer may not know about cognitive psychology, and so on. It is hoped that the material presented will be of interest to a diverse group of readers.

Itinerary of Topics

In Chapter 1, an overview of the transcription system is presented, including the form of input and output, and furnishing a general flowchart. Chapter 2 provides historical background, first of previous transcription efforts, then on issues of time perception and musical timing, from low to high level. There follow some theoretical treatments of rhythm, and finally, related issues in ethnomusicology, both theoretical and practical, are discussed. Chapter 3 is a detailed description of how the current transcription system actually works. Lastly, in Chapter 4 we try to draw conclusions about the work, and suggest directions for future research.

Chapter 1

Overview

1.1. The Problem

Several attempts have been made to transcribe music automatically, beginning with an acoustic signal and terminating in some kind of visual representation describing the musical events. In Chapter 2 we describe some of these attempts. Certainly there are many levels at which this process can be considered to succeed or fail, in technical, and ultimately musical, terms. When any sound is heard, from the simplest sine wave, to a rich orchestral climax, to a car crashing through a plate-glass window, what reaches the ear is in effect simply a single-channel signal. This single-channel signal, which is the algebraic sum (superposition) of all contributing sound sources and is therefore a single function of pressure vs. time, contains many levels of information and can thus be parsed in many meaningful ways by the auditory system, depending on context, experience, and need. The auditory system, even in the untrained listener, exhibits more and more amazing analytical power as one attempts to duplicate what at first might seem trivial listening tasks.

Of course, the issues begin with the auditory system, but extend naturally to cognitive and perceptual processes. It is not enough to decipher *what* was played; some sense has to be made of *how* the material is constructed. This is tantamount to a musical analysis, because the listener, either implicitly or explicitly, perceives structure in the music. A trained listener may begin to make the structure explicit; an untrained listener probably cannot, but he or she perceives it in some form nevertheless. This is equally true of notated and non-notated (improvised) music.

The process of notating music is dependent upon a significant level of musical analysis: there is no way to notate music of a particular style without having a theory of that style. Thus, any theory of music is heavily laden with conventions about the styles it describes. In fact, a theory may distort a listener's percept of what is being played. One hears what maps onto one's musical paradigm, in a kind of "global categorical perception." We do not want to "squash" the music into Western notation; we hope instead that Western notation, when carefully derived, can provide us with powerful insights into the music.

One of the difficulties with any attempt to transcribe music has been described by Charles Seeger [Seeger, 1958] as the dichotomy between "prescriptive" versus "descriptive" music notation. Prescriptive notation is the notation we are accustomed to seeing in Western music, that is, the score. In the score, innumerable conventions are assumed, and the performer implicitly includes these unspoken conventions in every performance. The task of many musicologists is to clarify some of these implicit conventions, or to infer what they might have been at the time that the music was written. Descriptive notation, on the other hand, should ideally assume nothing, and try to present a map of every "salient" feature of the musical event. This is almost impossible, because such a map, if one could actually pin down all the features, would be so dense as to be nearly impossible to parse. It would tell you "more than you wanted to know." Thus, musical knowledge is needed to represent the data in a meaningful way.

In this thesis, the effort is somewhat different from those previously cited, because the musical input is restricted to percussive music and the analysis concentrates mainly on timing issues and rhythm. Different problems are thus introduced; since the sounds are not always periodic, as in other studies, more attention is paid to modeling attack characteristics. Also, instead of pitch, the important dimension (apart from timing) is "stroke characteristic." That is, one must identify what kind of stroke, on what drum, corresponds to each attack.

Once the timing and stroke characteristics have been determined, a notelist* is created. Following this, the rhythm is analyzed. Finally, armed with the analysis data of real performances, inferences can be made about style and structure, periodicity, and the use of specific deviations from a canonical time-base for musical purposes.

The analysis system is being applied, in the current research, to music for which there is no score. Presumably, if there is already a score, then this tool is useful not as the producer of a score, but as a means by which one can compare the score with the actuality of a particular performance, resulting in investigations into

---

* The notelist is an uninterpreted list of begin-times, durations, amplitudes, and any other important features or parameters of the music.

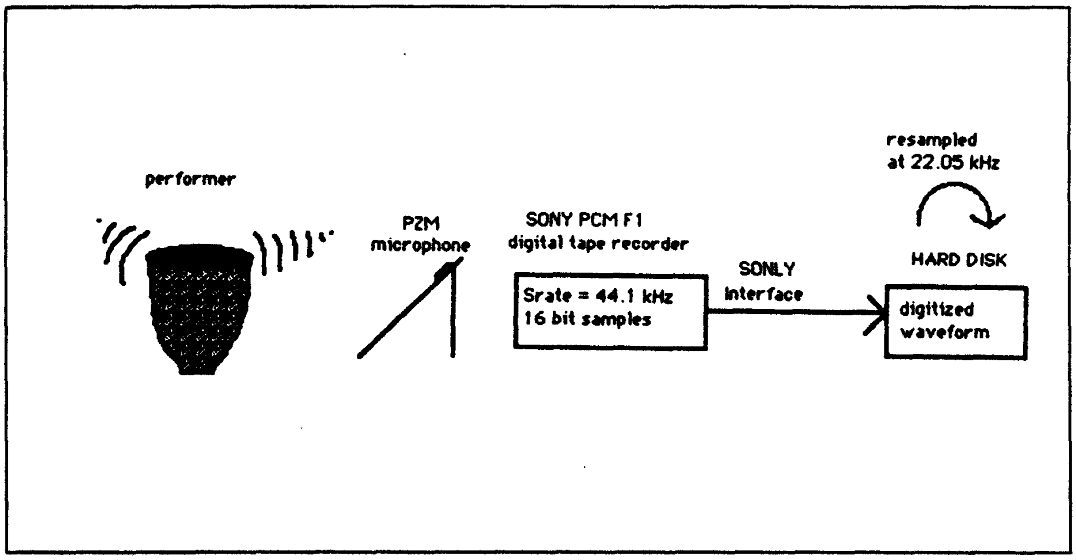

performer
PZM
microphone
SONY PCM F1
digital tape recorder
Srate = 44.1 kHz
16 bit samples
SONLY
interface
resampled
at 22.05 kHz
HARD DISK
digitized
waveform

Figure 1.1. The recording setup.

performance practice. However, since much of the world's music is improvisatory, the way is open for transcription and analysis tools to be applied to a large body of music, both for systematic study and for recreation of desirable performances.

Finally, this work may be seen as part of a larger scheme: that of an intelligent editor of music. I hope that this thesis will provide a few contributions towards this effort. Automatic segmentation is a crucial first step towards the efficient manipulation of musical material, and globally, the position of musical events in time gives the most effective way to maneuver within a piece of music. Details on the automatic segmentation method are presented in Chapter 3 of this thesis.

1.2. Input to the System

The input material may take the form of acoustic or synthesized signals. The use of synthesized waveforms affords greater control of test data, and provides a precise preliminary check on analysis results. The acoustic data consist of digital recordings of several different drums in a very dry room, using a Crown PZM microphone attached to a plexiglass reflector .8 meters from the drumheads, at the same height as the drumhead (see Fig. 1.1).

The signal is digitized by the Sony PCM F1 digital tape recorder and stored on a Betamax cassette at a sample rate of 44.1 kHz, in 16-bit samples. Later, the musical material of interest is read from the PCM-encoded tape and written on a hard disk via the SONLY interface (designed by Phil Gossett at CCRMA), which allows direct transfer of the digitized signal without converting it to analog and back. Once the data are written on the disk, they can be manipulated by the standard soundfile software at CCRMA, or by special processing routines written as necessary. In this thesis, all algorithms are written in SAIL, and run on the FOONLY F4 computer, which emulates to a DEC PDP-10. Because the signal processing is done on a general-purpose computer, it requires approximately a ten-to-one compute-time to real-time ratio, but most of it could be easily modified to run in real-time on an array processor. To alleviate the compute-bound processes, the signal is resampled at 22.05 kHz before further processing. Resampling lowers the bandwidth of the signal, but has no effect on the efficacy of the procedures; high sampling rates are generally more important for synthesis than for analysis.

In difficult cases, the signal may be high-pass filtered to facilitate segmentation. This is done because the upper partials of a tone tend to die out more quickly than the fundamental. Since the upper partials have nearly the same onset time as the fundamental, this simplifies the work of the segmenter by reducing the problem of overlapping notes (see Chapter 3).

1.3. Types of Output

Intermediate level representations, such as time-domain waveforms, FFTs, envelope graphs, or parameter maps can be collected at all points in the system. Music notation is automatically created upon completion of an example. The system is interactive, and allows the user to change values of numerous parameters and thresholds in mid-execution. It is most useful to select a setting for some of these values using a small data set, and then let the system proceed uninterrupted on the body of data under consideration. Figure 1.2 shows the system overview.

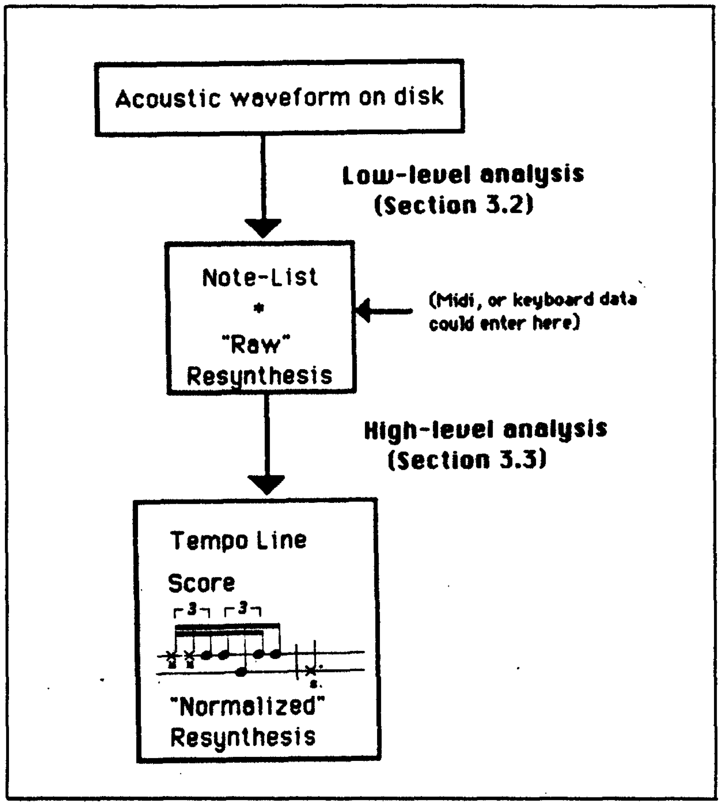

Acoustic waveform on disk
Low-level analysis (Section 3.2)
Note-List
*
"Raw" Resynthesis
(Midi, or keyboard data could enter here)
High-level analysis (Section 3.3)
Tempo Line
Score
"Normalized" Resynthesis

Figure 1.2. System overview. Auditory feedback is important at both low and high levels of analysis. The "raw" resynthesis means reconstruction of the music from the notelist, to see if the signal processing is correct. The "normalized" resynthesis means reconstruction of the music from the various versions of musical notation generated by the higher level analysis.

Chapter 2

Background and Theory

"The only true notations are the sound-tracks on the record itself."

— from Bartók, 1951

2.1. Introduction

To put the focus of this thesis into an historical and theoretical perspective, previous research in several areas will be outlined in this chapter. Also we occasionally go beyond reporting previous work, and introduce ideas that are new, or a synthesis of previous ideas. It is hoped that these elaborations will shed light on what has been a somewhat elusive topic.

In Section 2.2 we review actual attempts at automatic transcription, of which this thesis is one. The subtleties of the task, both at a perceptual (musical) and technical level, are numerous, and the problem is far from solved.

In the rest of the chapter, an attempt is made to orient the reader within the realm of rhythm, from the lowest to highest possible levels. Section 2.3 deals with studies of timing, which in themselves cover a wide range in approach, from psychoacoustical to cognitive, including systematic deviations from metrical patterns. Section 2.4 follows with theoretical investigations into the issues of metrical structure and hierarchies in music. In Section 2.5 we attempt to describe a tentative paradigm for categorizing music in a global perspective, showing how rhythm functions and how it is constructed in different world-music contexts. Finally, in Section 2.6, we briefly discuss some specific characteristics of the type of percussive music that is being analyzed herein.

2.2. Previous Research in Automatic Transcription

Almost all previous methods for automatic transcription were applied to non-percussive instruments. They are included here because many of the ideas and basic problems are the same. The earliest work predated modern digital technology and therefore was analog-based, staying "close to the signal," that is, no higher-level decisions were made. More recent work includes considerable high-level musical analysis, entering into the realm of Artificial Intelligence. We begin with the first researcher to deal with the problem systematically.

2.2.1. The Melograph

The late Charles Seeger (husband of the composer Ruth Crawford Seeger, and father of Pete and Peggy Seeger) occupies an important position in the field of musicology. He was one of America's earliest and most prominent ethnomusicologists, and he made significant contributions to the theoretical basis of a systematic study of world music.

Seeger, as early as the 1940's, saw the need for an *objective* representation of performed music. His earliest efforts led to a graph of fundamental frequency against time, and even today, a robust pitch-tracker remains of interest to the computer-music community. Seeger was convinced that a graphical representation of music was the answer to objective studies: "I have a feeling that before a hundred years are passed our present notation will look more like a graphic than a symbolic method of writing." [Seeger, 1951]. On the other hand, such a graphic representation, while it is an objective record of events in time, does not represent the *musical* abstractions that are so important.

Six years later, Seeger moderated his position somewhat, mentioning that traditional notation complements the graphical representation: "Our conventional notation will not serve—and we should no longer pretend it can serve—the need of a universal music sound-writing. To no one would I recommend abandonment of traditional techniques of writing music for the novel and still undeveloped graph. For the present, I would urge the two to be used side by side." [Seeger, 1957].

In order to represent more aspects of the music, Seeger developed his Melograph, which had three temporally aligned time-varying graphs: amplitude vs. time, fundamental frequency vs. time, and a high-resolution spectrograph. The spectrograph is really the output of a bank of band-pass filters that record the energy in each channel on a chart recorder. The amplitude is simply proportional to the voltage of the input signal. The fundamental frequency is found by scanning a bank of 1/3 octave filters, beginning with the lowest frequency, and when a maximum is found, it is assumed to contain the fundamental. Then the zero-crossings of the output of the filter are counted and plotted on the chart recorder. The problem with this method is that the fundamental may be very weak or absent; it would be more robust to try to determine the fundamental from looking at its harmonics, or using some other modern technique like the *cepstrum** to find the fundamental frequency, but this is a much less straightforward task.

The Melograph evolved over a period of twenty years; the most developed version was the model C, completed circa 1970. The model C is more accurate than previous models, and though mostly an analog device, it does supply output available in digital form, which could be subjected to further analysis [Hood, 1971]. The melograph is intended to be applied to a monophonic (one-voice) input, and is rather sensitive to noise. See Fig. 2.1 for an example of Melograph output.

This instrument gives the researcher a very useful overview of his data. It makes no attempt to interpret anything; this is left to the user. The process of interpreting the output of the machine is laborious and prone to error. Subsequent music transcription attempts that address the problem of representing the music in a more abstract way are described below.

2.2.2. Moorer's Thesis

James Moorer's doctoral thesis [Moorer, 1975] is an important contribution to the field, although many restrictions were placed on the material to be analyzed. Moorer occupies a central position in the "lore" of computer music, and his dissertation is still timely nine years after it was written, because he makes an early attempt to deal with polyphony.

Moorer's work is quite different in orientation from this thesis. He does not try to deal with tempo variation, instrument identification, notes of less than 100 msec. in duration, and inharmonic spectra, all of which are dealt with here. Also

---

* The cepstrum is defined as the inverse discrete Fourier transform of the log magnitude spectrum of the time data.

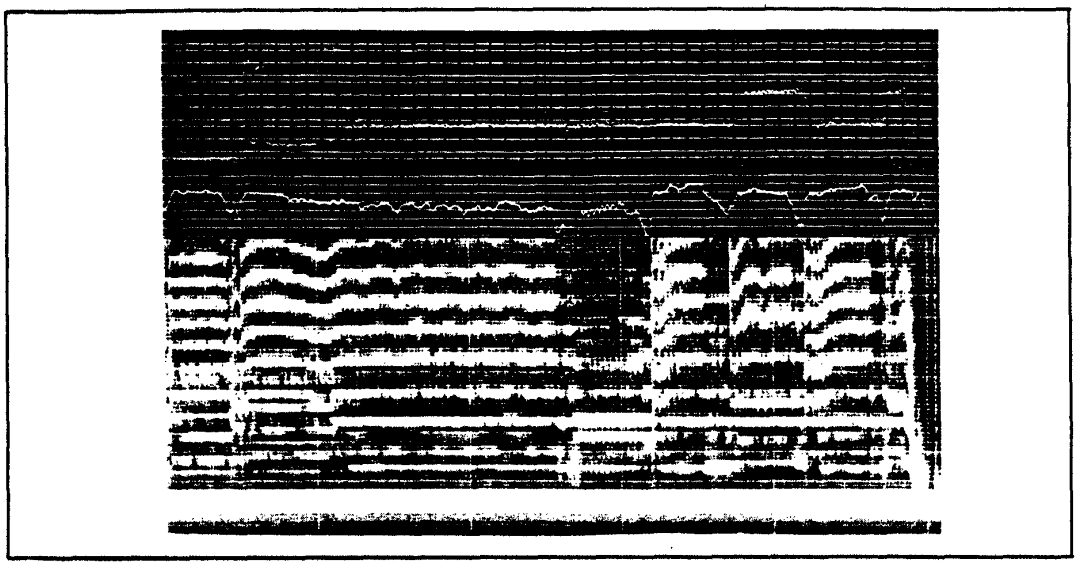


Figure 2.1. An example of the Melograph output for an excerpt of the Chinese *sona* (double reed aerophone). Amplitude trace on top, fundamental frequency trace in the middle, and spectrograph in the bottom half. From [Hood, 1971].

he disallows glissandi, fast trills, and vibrato; however, he is dealing with the transciption of polyphonic music, which is still a largely unsolved problem. In his pioneering attempt, he also restricts input to two simultaneous voices, and he disallows unisons, octaves, twelfths, and some other intervals because the harmonics of two instruments playing at these intervals will overlap, greatly complicating the disambiguation of these cases.

Moorer's thesis is valuable not only for what methods it employs in the transcription effort, but in its review of methods that were tried and found not to be useful for musical analysis. For instance, Moorer does not recommend the Discrete Fourier Transform (DFT), and its more computationally efficient analog, the Fast Fourier Transform (FFT), because he says they are distorted by any change due to either amplitude or frequency, and thus will give misleading results in the presence of vibrato or room reverberation.* Two speech analysis techniques, the cepstrum (the inverse DFT of the log magnitude of the DFT of the input), and the linear predictor, are also not recommended because of their apparent problems in dealing with multiple sources (polyphony).

---

* The DFT and FFT are obviously workhorses in any music analysis situation, but they have to be carefully extended to succeed in complex musical situations.

Moorer also describes the heterodyne filter that was used to produce John Grey's plots in his dissertation on musical timbre called *An Exploration of Musical Timbre* [Grey, 1975]. The heterodyne filter is basically an adaptation of the DFT; its problem is that it is too sensitive to frequency and amplitude variation (present in most musical signals), and if the partials wander out of a channel, it is difficult to interpret the results. Thus, the heterodyne filter is most useful in analyzing single isolated tones from various musical instruments, but not for continuous music or more complex situations.

Of the methods actually used in Moorer's research, the first applied to the data is a technique related to the autocorrelation function, which Moorer calls the "optimum comb." Its derivation is as follows: For each data point x_n in the waveform of k samples, form the sum

$$\sum_{i=0}^{k-1} \| x_{n+i} - x_{n+i-m} \|,$$

looking for minima over m in this sum. This is really a preprocessor, and is used to find the basic periodicities, or "harmony" of the data, mostly to reduce the extent of further processing. The idea is that one can find a "least common divisor" of harmonics using this method, which helps in setting bandpass center frequencies later.

The main signal processing method is actually a carefully adapted band-pass filtering routine, set to examine multiples and subharmonics of the basic periodicity found by the optimum comb. Then the results of the band-pass filters are analyzed and the notes that could produce the original sinusiods are inferred. This is not an obvious procedure, in that there is not a one-to-one correspondence between the original sinusoids and the supposed source.

Moorer is not specific about how he goes from the notelist to the score; assuming that the tempo is constant makes this task more manageable. Tempo fluctuations inevitably make the time values found by the lower level analysis extremely unreliable to notate without further processing before converting to musical notation.

After finishing his thesis, Moorer explored the possible use of the phase vocoder [Moorer, 1978], which is similar to the heterodyne filter except that the output of each channel is not simply the sinusoid that falls into it, but rather a time-varying parametric representation for each channel. This means that the phase vocoder is less vulnerable to vibrato and tremolo than the heterodyne filter, and therefore presumbly might have been his preferred method for transcription if it had been implemented at an earlier date. However, the phase vocoder presents problems when used for quasi-periodic signals like percussion, because it is difficult to "line up" the channels of the phase vocoder with the partials of the percussive tones [Gordon and Strawn, 1984].

2.2.3. Piszczalski and Galler

Piszczalski and Galler have been publishing articles on automatic transcription since 1977. Their approach is basically pragmatic, and is loosely tied to perceptual theories. In their first article [Piszczalski and Galler, 1977], they concentrate mostly on the recorder and flute. They describe three levels of their analysis system applied to these instruments. The three levels are actually quite similar to Moorer's work cited in the previous section, viz:

1. Convert the time-waveform to an amplitude-time-frequency representation; in this case, a 3D plot based on successive FFT's of the signal.

2. Infer from intermediate processing what musical "notes" can best account for the frequency, amplitude and time values derived in step 1.

3. Represent the events hypothesized in step 2 in musical notation.

The most interesting part of this system is step 2, as there is some intelligence built into this level, mainly in trying to account for the harmonics as they combine to make a continuous tone; for example, if the fundamental has a drop in amplitude but the second harmonic is continuous, then there should be no segmentation at the point where the fundamental drops.* There are inevitably some frequency variations detected, and the value derived by averaging the frequencies over the duration of a given note is taken as the musical pitch.

The example is then scored from the pitch and time information. No attempt is made to deal with tempo fluctuation; in their examples, they try to provide

---

* That is, there may not be a new attack at the point where the fundamental swells to its original amplitude; the program should be looking at more than just the fundamental when trying to detect new attacks (segmentation).

themselves with a constant tempo (a very unlikely situation in a real performance). This assumption of constant tempo is typical of the early transcription attempts; there is enough to deal with in just getting reasonable results in pitch-tracking and segmentation from the acoustic signal.

In Piszczalski and Galler's method just described, the problem of a weak or missing fundamental is not addressed. They subsequently developed their system and made it more robust by adding a scheme to infer the pitch from a set of harmonics [Piszczalski and Galler, 1979]. The method is basically to look at adjacent pairs of harmonics, and calculate the ratio of their frequencies. If this ratio is close to a low-integer ratio, it is used to estimate the fundamental frequency. The idea is to infer the likely harmonic numbers; the value of the fundamental should be uniquely defined by its harmonics. Since the calculation is done over several pairs of harmonics, the estimate gains reliability. In fact, the method saves all pairs in a table, but there is no combinatorial scheme applied to all the ratios; they are simply weighted according to their relative amplitude, and the best value is chosen.

This method works satisfactorily in most of their test cases. It fails in the following cases:

1. Substantial ringing of a previous note (that is, any amount of polyphony, inadvertent or intentional).

2. The fundamentals are low in frequency, because the frequency ratios will often imply the wrong fundamental. (The strong higher harmonics will have closely-spaced ratios.)

3. The duration is very short (this is due to the time-resolution limit of 40 msec. from the lower-level processing).

It is also hard to imagine this method working effectively for inharmonic tones, but of course the sense of pitch elicited by inharmonic tones diminishes as the partials depart from integral multiples of the fundamental. In the case of percussive instruments, the subject of pitch perception is still quite open, and the transcription should reflect this by notating pitch where it is perceptible.

Piszczalski and Galler later implemented the above method into their transcription system, which certainly improved pitch-tracking for instruments with weak fundamentals. Also, to reduce the compute-bound bottleneck of the spectral estimation calculations, they abandoned the standard FFT for a charge coupled device

(CCD) performing a chirp Z-transform (CZT). This operation can be done in real-time, and since this would be the most computationally intensive part of the system, it makes possible more exploration of graphic representations of the time-varying spectra, but it does not change the basic approach. Polyphonic input is still not addressed, and there is no way to deal with varying tempi.

2.2.4. The Auditory Transform

In his MS thesis [Stautner, 1983], Stautner tries to develop and implement an analysis method that is closer to the way we hear. Tuning his system to several "rule-of-thumb" auditory parameters such as critical bandwidth, nerve firings, and the *jnd* (just-noticable difference) of loudness, Stautner creates 3D plots that are carefully tuned short-time Fourier Transform log magnitudes (STFT). These plots are visually informative, and he shows how the sound can be resynthesized from the plot, but to get more useful information from the data, the analysis must be extended. First, a second stage of STFT magnitude analysis at the output of each analysis channel is computed. It turns out that, though the Auditory Transform resolves the first few harmonics, the higher partials will fall together into single channels, which results in temporal beating. This will result in a "pitch periodogram" for the second spectral analysis, allowing one to infer fundamental frequency. This method resembles the cepstrum, though Stautner does not refer to it as such.

The last level of analysis is in the application of *principal components analysis* to search for the optimal model for the spectral data. This analysis yields several orthogonal components that relate to physical features of the sound. The sound can be resynthesized successfully from the principal components analysis. The method of principal components is rather computationally intensive, and though it leads to some interesting results, one wonders if there is not a simpler way to achieve the same goal.

When the above analysis is carried out on a section of music played on the Indian *tablá* (a set of two tuned drums), the detection of several salient features is demonstrated. It turns out that the first six principal components seem to correspond to certain aspects of performance; however, the task of automating the translation of these features from an observation to an algorithm involves some kind of pattern-recognition, and is non-trivial. In fact, it has not been done except to detect onsets in the music. Resynthesis was done to debug the onset detection, and seemed to be correct. However, this is a long way from a complete analysis.

2.2.5. Recent Work at CCRMA

Several researchers at CCRMA and at Systems Control, Inc., including the author, working on a joint National Science Foundation grant, have been dealing with the problem of automatic transcription. Their early efforts are described in [Foster, et al., 1982] and [Chafe, et al., 1982]. The former paper concentrates on lower-level problems, the latter paper on high-level issues. Here we give a summary of some of the work done.

Foster, et al. developed a segmentation method based on an autoregressive (AR) model fit. The idea is to compute two AR models; one for current data, and a delayed model recursively tracking older data. Each model has an associated model error $\mu(t)$ for given input data. Wherever the data has changed (in the sense that it is modeled differently), the "delayed" model applied to current data will have larger error than the current model applied to current data. Thus, one can detect changes in the data by forming a ratio of the two model errors. If the error ratio is plotted against time, there will be a "spike" marking each new event, because the model errors will suddenly differ substantially, thus the ratio will suddenly depart from unity.

The above method was tested on excerpts of flute, 'cello, and vibraphone music, with interesting results. The AR model error ratio will detect events effectively even when there is substantial overlap between events, and it works on instruments with inharmonic spectra as well as harmonic ones. Problems with this method are:

1. For all the computation, one only finds out that "something happened," but one has no idea *what* happened (though it is certainly important to know when the event occured even if it is as yet uncharacterized).

2. The method, although it is easy to describe, is difficult to intuitively tune or modify; the parameters do not correspond in a clear way to physical properties of the signal.

3. Since amplitude changes alone will not change model error, this method will miss repeated attacks on the same note. For this reason, amplitude thresholding methods might complement this method.

The main signal processing method used initially by Foster, et al. is similar to that used by other researchers, in that the method proceeds directly into the frequency domain. However, it differs from other similar routines, in that the data are processed in *reverse* short time segments, noting the location and amplitude of the spectral peaks. The rationale for tracking the spectrum *backwards* is that, typically, the "middle" of a note will have the clearest and most coherent spectrum. If we proceed backwards *toward* the attack, we can obtain the time-varying spectra of the note, just up to the attack, at which point the spectrum becomes incoherent. A pitch estimate is made for each time frame, and consecutive time frames are compared. In this way, moderate tremolo and vibrato can be characterized. The method is a trade-off with Moorer's, in which he is able to track two voices, but he restricts the behavior of the voices considerably. This pitch-tracking method is the source of the "event-list" that is analyzed in the companion article [Chafe, et al., 1982].

From this point, the higher-level part of the system aims at extracting tempo and meter from the raw timing data; this is not attempted in other studies. Since Chapter 3 of this thesis describes an elaboration of this process, it will not be discussed here.

Desiring better temporal accuracy in their pitch-tracking, Foster et al. also devised a pitch-synchronous spectral analysis routine. This method generates spectral plots over integral numbers of periods of the time-waveform, and for quasi-periodic input, the amplitude and phase can be accurately measured over each interval of time. Then an 'instantaneous frequency' is estimated for each partial. This fine-resolution method was tested on various musical examples, but has not yet been integrated into the analysis system, because it is very computationally intensive.

One interesting test case that was analyzed was a 1913 recording of Alma Gluck singing *Ave Maria.* The reason for choosing this example as a test case was to compare our results with those of Carl Seashore and his colleagues' studies of vibrato of the same recording. We found substantial agreement between our pitch plot and the plots done by R. Miller, a graduate student of Seashore [Miller, 1932], which is a tribute to the resourcefulness and meticulousness of these early researchers.*

Segmentation by amplitude was also tried, and was found to be very difficult, due to room reverberation and other factors. Trying to effect slope detection by differentiating (1- or *n*- point differencing, in fact), yielded very poor results due to inevitable noise, and large fluctations of amplitude. Low-pass filtering to reduce noise did not improve the situation, because it effectively smoothed the attacks so much that they were not detected. Finally, a robust amplitude segmenter was implemented, one that works quite well with percussion instruments. It is described in Chapter 3.

2.3. Timing Studies

No matter what definition one considers appropriate for rhythm, it seems impossible to deny that rhythm has to do with some kind of accurate perception of events in time. If we hope to make sense out of temporal information, it is important from the start to have some notion of the kind of accuracy or granularity that is required in order to capture what a human listener responds to. There must be some physiological basis for rhythm, which of course expands into higher level (central) processing of musical and perceptual dimensions. In this section, we address many interrelated questions, such as: What is the best we can do in terms of temporal discrimination? Is it roughly the same for everyone? Are musicians more skilled than others? When is the *perceptual* onset of a tone? Does our temporal discriminability vary over a large span of durations? Is rhythm based soley on physiological sources? How do we "measure" time? And at a higher level, how do performers deviate from "normal" (scored) rhythmic patterns?

---

* In fact, it is probable that the first concerted efforts to objectively study significant aspects of musical performance were done by Seashore and his students at the University of Iowa, published in several volumes under the title *University of Iowa Studies in the Psychology of Music.* In particular, Volume I: *The Vibrato,* Volume III: *Psychology of the Vibrato in Voice and Instrument,* and Volume IV: *Objective Analysis of Musical Perfomance,* [Seashore, 1932, 1936, 1936] are of considerable interest.

2.3.1. Lower Level

In trying to define limits on temporal discrimination, researchers have not always agreed on their results; one fact stands out, however—the "ear" is the most accurate sensory system in the domain of duration analysis. There are three kinds of temporal discrimination that might be tested: duration, intermittency, and regularity. Duration discrimination means estimation of a single time interval. Discrimination of intermittency refers to estimation of a beat rate, and regularity discrimination involves the evaluation of evenness of a repeated pulse. Then, in a musical context, one wants to evaluate precision of a beat pattern, or rhythm.

Naturally one is curious to see if the dimension of auditory temporal discrimination follows a Weber law.* It turns out that, for very long or very short durations, Weber's law fails, but in the area from 200 milliseconds to 2 seconds, a modified version of Weber's law seems to hold, according to Getty [Getty, 1975]. This range is in the range of typical musical notes (two seconds is equal to 𝅗𝅥 at ♩ = 60). Below 200 msec., the relationship is not clear, but is probably more like a "constant offset." This basic result is also reported by Michon [Michon, 1964], but with two regions: .01 for (100 msec. $< t <$300 msec.), and .02 for (300 msec. $< t <$ 1 second). For Michon, the law did not hold outside these regions.

In Michon's Ph.D. thesis [Michon, 1967], he extends the discrimination experiments to include other kinds of timing questions. All the experiments are based on key tapping by the subjects. This is a very accurate method, because there is no possibility of error in deciding when the key is tapped, but there is a limitation on the *musical* inferences that can be made. In a musical context, the performer is in a much more complicated feedback loop involving his instrument, the room he is playing in, and the auditory system. Thus the efforts in this thesis to go from the *acoustic waveform* are worth the trouble, because the feedback loop of the performer *precedes* the analysis, and is therefore implicitly being analyzed. See also the next section for experiments by Gabrielsson, et al. that attempt to deal with this problem.

Michon posits a signal detection theory for the perception of time, similar to

---

* Weber's law is a general psychophysical law that states that the perceptual discriminability of a subject with respect to a physical attribute is proportional to its magnitude, that is, that $\Delta x/x = k$ where x is the attribute being measured, and Δx is the smallest perceptual change that can be detected. k is called the *Weber ratio*, a dimensionless quantity.

Creelman's counter model [Creelman, 1962], in which a short interval "ticks" by and is incremented and compared with a standard; this is presumed to be characteristic of the auditory system, but is really a model, and not a description of actual events. Divenyi [Divenyi, 1971] also tries to apply a counter model to his data, in which he uses marking tones of different frequencies. There is implicitly an internal clock proposed in the counter model, but no description of its mechanism is given. In any case, Getty reports that for his data, Weber's law (slightly generalized) is more successful than the counter model.

Certainly long intervals (> 2 seconds) are processed in a different way than are what we might call the "musical range" of 100—2000 msec. Perception of long intervals is much less accurate, unless the subject actually counts and/or subdivides the interval, but this is based on the smaller intervals again. Lorraine Allan [Allan, 1979] specifies four ways of investigating time perception that clarify the problem for longer durations:

1. Verbal estimation: "The interval was 45 seconds."
2. Production: "Tell me when 45 seconds have passed."
3. Reproduction: "Play this."
4. Comparison: "Which interval is longer?"

Because 1 and 2 above allow no sense of continuity, they are not really musical situations. Cases 3 and 4 do not require "conversion" to non-musical time, so are closer to musical experience.

Lunney employed a different experimental paradigm. He had listeners controlling an electronic metronome in which every fourth beat was irregular, and the amount of irregularity could be controlled by the listener. The task was to cause the fourth beat to be *just* perceptually irregular [Lunney, 1974]. He found a Weber ratio of about .04 for durations up to 300 msec. With practice, he was able to become quite consistent, but not more accurate. For this reason, he says "the limits of discriminability are biologically imposed." In musical terms, we are not too concerned with durations longer than 2 seconds; these are not typically assigned musical note values. But, as we have seen, for durations less than about 200 msec., Weber's law fails. This is below the range of most note values, and falls into the range of deviations from expected values (see Section 2.3.3.).

To see just what the limit of discriminability might be in terms of two events close together in time, we look first at the classic study on temporal order perception done by Ira Hirsh [Hirsh, 1959]. Using synthetic stimuli (a turntable with adjustable activation switches) he found that it was possible to separate perceptually two brief sounds with as little as 2 msec. between them; but in order to determine the *order* of the stimulus pair, about 15—20 msec. was needed. When this experiment was repeated by Patterson and Green [Patterson and Green, 1970] they found that listeners could do much better, that their subjects could distinguish temporal order with much smaller intervals between the events. They explained this by noting that the overall durations of their stimuli were much shorter.

It seems that people typically have the most accurate sense of temporal acuity in the range between 500—800 msec., which, for example, at ♩ = 60, corresponds to the note values ♪ to ♩. In the normal musical range, it is likely that to be within 5 msec. in determining attack times is adequate to capture essential (intentional) timing information. It will be seen that tempo fluctuation and other inaccuracies dwarf this 5 msec. standard for accuracy in musical contexts, but one wants as much accuracy as is reasonably possible at the lowest level, so that the "normalization" (or fitting to closest metric values) that proceeds from the original values begins with the clearest data. Eventually, one can compare the original data with the "normalized" score to begin to establish a statistical evaluation of intentional or unintentional deviation from "canonical" note values. See the next section for more information on this subject.

One wonders whether the sense of rhythm is inherently physiological (in the sense of the body as opposed to the mind). Fraisse has written about this topic in several articles [Fraisse, 1978, 1982]. He describes what he calls "personal tempo" in which a person taps his forefinger at a spontaneous rate, which is measured. It is usually in the range from 380—880 msec., with 600 msec. perhaps the most representative. He claims that, though there is great interindividual variability, individual variability is slight. He also mentions that personal tempo correlates well with the swinging of the leg of a seated subject, or swinging of the arm when standing. The fact that the speed of walking, or of the heartbeat is also in this range may be misleading. For example, acceleration of heartbeat does *not* correspond to an acceleration of Fraisse's personal tempo.

It happens that the most accurate time discrimination is also in the neighborhood of 600 msec., but this too may be a coincidence, or non-causally related to gross physiology. In his Ph.D. thesis, *An Analysis of the "True Beat" in Music* [Lund, 1938], Max Lund makes a strong case that the sense of time in music is not a direct result of physiology, but rather a higher-level cognitive ability. This makes sense if we think about the cultural diversity and influence of environmental factors we see in rhythm, which would be less pronounced if rhythm were primarily physiologically based.

2.3.2. Perceptual Attack Time

To close this section, there is an issue that should be mentioned in the context of any careful study of timing in musical contexts—the problem of Perceptual Attack Time (PAT). The reason for mentioning PAT is that the first moment of a disturbance of air pressure is not the same instant as the first *percept* of the sound, which in turn may not necessarily coincide with the time the sound is perceived as a *rhythmic event*. There inevitably will be some delay before the sound is registered as a new event, after it is physically in evidence. This variable delay could cause error for an automatic transcription program, so it should be covered.*

Vos and Rasch investigated PAT using synthetic tones that were complex sawtooth-like waveforms, with rise-times varying from 5—80 msec. They played the tones in an identical A-B-A-B order; the subjects were instructed to adjust the physical onset until the repeated attacks sounded regular [Vos and Rasch, 1981]. They then tried to explain the perceptual onset by various models. Their principal hypothesis was that PAT occurred when the amplitude crossed some threshold relative to the local maximum amplitude, usually about 15 dB below maximum.

John Gordon, in his doctoral thesis [Gordon, 1984], extended the work of Vos and Rasch significantly by using real musical tones that were resynthesized and adjusted for equal loudness, and by also trying a larger number of mathematical

---

* A good example of this problem in a musical setting is provided by an attempt of John Grey to create a synthetic realization of a piece called "Loops," written by Robert Erickson, of UCSD. The piece is based on the concept of *klangfarbenmelodie*, which means that timbre is used as an conspicuous parameter, as pitch might be, and there is a sequence of tones that change not only in pitch but also in timbre (different instruments). The changes are so rapid that it was impractical to play with real instruments. When Grey reconstructed the piece by concatenating the specified tones at precise moments (to the nearest millisecond), it was evident that although the physical onset times of all the tones were correctly aligned, the sequence sounded uneven in rhythm.

models to see which best fit the data on PAT. Some of the models he tried were, for example, time of maximum amplitude, absolute amplitude threshold, relative amplitude threshold, integration threshold, and various slope threshold methods. Interestingly, he found that the relative amplitude threshold model of Vos and Rasch was not as successful for real data as a model based instead on *slope*; this is the approach used herein (see Chapter 3) to segment percussive music. In fact, since slope relates directly to the perception of attack transients, it should be especially appropriate for percussive instruments.

In order to find the slope, Gordon chose the same basic method employed in this thesis, which first finds the amplitude envelope, and then calculates the slope based on a least-squares fit to the envelope data. Gordon, in common with this thesis, abandoned the method of finding slope by using first order difference equations because of sensitivity to noise, and instead opted for a linear regressive fit to the points of the amplitude envelope, much like the method described in Section 3.2. It is interesting to see that Gordon's method of slope detection, chosen from among many possible methods, is similar to the segmentation method described in this thesis. (See Chapter 3.)*

It turns out that the actual "delay" caused by PAT in the case of drum sounds is quite small, because the slope is typically rather steep. Also, for consistent rise-times, the delay caused by PAT is consistent, and therefore constitutes a constant offset that will not affect the inter-attack durations; that is, $(t_{n+1}+\Delta t)-(t_n+\Delta t) = t_{n+1} - t_n$, or the duration between t_{n+1} and t_n. So, if the instruments involved have similar attacks, the problem is minimized.

2.3.3. Higher Level—Timing in Musical Contexts

In the previous section, we explored some of the basic questions of temporal discrimination that are quite general. Now we examine some research into questions of timing in musical contexts. Instead of experiments using clicks and isolated time intervals, there is an attempt to deal with specifically musical perception of time. We will see that, although the researchers mentioned in this section are addressing timing questions as they relate to music, many attempts still employ unnatural experimental materials because of the difficulty of extracting robust timing data from real performances.

---

* Actually Gordon decided to try this method after discussing it with the author.

Fifty years ago, Carl Seashore and colleages, ahead of their time in musical research, dealt with objective analysis of musical performances directly from musical data, not from artificial situations. For instance, in their extensive studies of the vibrato published in several volumes of the *University of Iowa Studies in the Psychology of Music* [Seashore, 1932, 1936, 1936], questions such as vibrato depth, rate, articulation, and intonation were explored. Of course, a large amount of their analysis included embellishments "by hand."

In addition, Seashore, et al. made substantial contributions to the issues of timing in piano music by the design and implementation of the "piano camera." The piano camera could be considered a first step towards automatic transcription. It provides a photographic record of begin-time, duration, and relative intensity of each note played on the piano [Henderson, Tiffin, Seashore, 1936]. The camera is truly a "Rube Goldberg" device in which strips of balsa wood are glued to the tail of each hammer on the piano, and via a complicated sequence of events, a continuous recording of duration and velocity of each keystroke is made on a moving film.

The resulting photographic record is transcribed to a "musical pattern score," which is a proportional notation bar graph superimposed on a grid representing the musical staff. M.T. Henderson then used this method to analyze the chorale section of Chopin's *Nocturne No. 6* (G minor) Opus 15, No. 3. Several findings were reported, for instance, a lack of correlation between perceived accent (of the first beat of a measure) and performed intensity—only .06 for one performer, and still only .19 for the other (though .19 is a much higher correlation, it is still surprisingly low). We can infer that duration, or delayed entrance, is more likely to result in a perceived accent than intensity cues alone.* Interestingly, Henderson found that when there is a pattern of duration (like the second quarter note of a pattern shortened with respect to the first), this relationship tends to be true even in the context of accelerando, ritardando, crescendo, or decrescendo [Henderson, 1936]. This lends credence to the concept of "local tempo," in the sense that, taking into account tempo fluctations, intended relations between adjacent notes are upheld. In Section 3.3, considerable effort is made to track tempo in order to distinguish intended musical relationships between durations.

Another area of interest is synchronization of notes in a chord. In fact, Henderson found that chords that are not notated as arpeggios are often played as such to varying degrees (.01 to .04 seconds for one pianist tested, .02 to .2 seconds for another). Further study of synchronization in piano music was done by Vernon, who did not use the piano camera, but rather, Duo-art rolls, that are recordings of world-famous pianists preserved in the form of piano rolls. He found great differences between pianists, but these general tendencies emerged [Vernon, 1936]:

---

* A short duration followed by a long one results in perceived emphasis on the longer tone, which is called an *agogic* accent.

- Frequency of deviation varies with average length of duration—this implies that most of the asynchrony is intentional.
- Slow or changing tempo correlates positively with asynchrony.
- Melody notes are often emphasized by being played early or late in a chord.
- Asynchrony is not related to beginnings or endings of phrases, or changes in tonality, contrary to expectation.

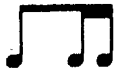
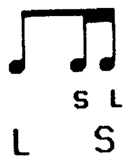
Of course, in musical contexts, there is an enormous difference between "random" deviation (such as Gaussian distribution around an intended duration), and *systematic deviation*, that can be shown to relate to musical context. Alf Gabrielsson and colleagues at the University of Uppsala in Sweden have done a considerable amount of research on systematic variation, which they call SYVAR. They published a series of articles in which they tried to characterize SYVAR in rhythmic musical examples.

In one study [Gabrielsson, 1974], two pianists and one percussionist performed notated rhythms that were recorded and subsequently analyzed by a device which they call the MONA analyzer, that gives fundamental frequency trace and amplitude trace through time. The MONA analyzer is an analog device that plots these parameters on millimeter paper, at a speed of 100 millimeters per second (see [Tove et al., 1966]). Gabrielsson reports that the onsets (rise-time) of the piano and drum are rapid—the peak amplitude is reached within 20—40 msec. Nevertheless, it is quite difficult to ascertain the exact location of the attack by looking at the amplitude trace on the paper, particularly because each millimeter = 10 msec. Therefore, ±5 msec. corresponds to a distance on the chart of only half a millimeter. If the paper were set to move more quickly to provide better time resolution, it would probably be too cumbersome, because fairly short musical examples would result in piles of paper chart to measure.

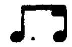
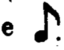
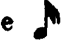
The task of segmenting the attacks from the chart analyzer is slow and prone to error, because one has to try to find (by eye) the point on the amplitude trace that corresponds to the desired attack, and do this for all notes in every musical example. In this thesis, a method for automating this task is described in Section 3.2. The automatic segmentation described herein makes possible a general-purpose and "tunable" segmenter that will eliminate subjective error in finding attacks, and would probably be a help to Gabrielsson, et al. in experiments of this sort.

In their examples, a metronome click is used by the performers to help steady the tempo. Thus, tempo tracking is not done; the determination of SYVAR is done from an assumed constant tempo. Some results reported are as follows:

1. There is a general tendency to insert a SHORT-LONG relationship to adjacent eighth-notes even though they are notated as equal.

2. Similarly, the figure  is often played: 

(L = long, S = short).

3. Many temporal relationships are exaggerated or "sharpened," e.g. for  the  is longer, while the  is shorter. In other words, dotted rhythms are typically "overdotted." The propensity to do this may relate to Fraisse's concept of "*temps longs*" or "*temps courts*" [Fraisse, 1978, 1982].

4. There are striking deviations from the "norm" in the case of syncopation, similar to (3)—the relationships have a tendency to be exaggerated.

5. The amplitude peak at the beginning of bars is a "weak" but noticeable characteristic of non-melodic examples, and is typically combined with prolongation of duration. However, for melodic examples, the highest peak amplitudes are irregularly spaced in the examples.

Later, Bengtsson and Gabrielsson [Bengtsson and Gabrielsson, 1980] extended the search for SYVAR by applying the MONA analysis to 28 melodies played on flute, clarinet and piano. The same problem of evaluation accuracy applies here as in the previous experiment, given that the attacks are found by hand. Once the attacks are found, a computer program (called RHYTHMSYVARD) was written that performs the following steps:

1. Calculate total duration.

2. Calculate tempo by dividing total duration by the number of beats and thus the metronome mark (MM). Note that this finds only an *average* tempo over the entire piece and has no provision for local tempo.

3. Normalize duration values by expressing them as percent values of total duration.

4. Calulate the deviation of each normalized inter-attack duration (they call this D_{ii}) from the corresponding normal value according to the notational/mechanical norm.

5. Calculate the *proportion* between neighboring D_{ii} values within a certain unit.

6. Perform factor analysis on the proportion data.

From this program, they graph the deviations from the mechanical values for the notation, and are able to make some general inferences. The deviation from the notational norm is found to be as high as 15—20% for half-notes and quarter-notes, or 20—40% for eighth and sixteenth-notes. These large deviations are obviously perceptible (we know the limits of discrimination as described in Section 2.3.1), but though quite large, these deviations do not apparently destroy the *structure* of the rhythm. Rather, they determine the character of the "flow" of the rhythm. In other words, these deviations, if they were random errors, would simply *destroy* the rhythmic intent, but due to the placement of the deviations, they are heard as a kind of "embellishment" of the rhythmic character. For example, it could be that "the perception of temporal/structural relations like 1:1, 2:1, 3:1 etc., that are so frequent in music, is an example of categorical perception. In other words, there is a rather wide tolerance zone around each such value, and as long as deviations stay within that zone the perceived temporal/structural relation does not change—but the deviations affect the perceived motion character, sometimes in very subtle ways." [Gabrielsson, et al., 1983]. In Chapter 3 of this thesis, we see that it is possible to "factor out" timing deviations, and find that there are several possible rational approximations to the actual normalized time values found.

Another path to understanding SYVAR is to try to *generate* SYVAR via synthesis. This is a powerful avenue that only in recent years has been possible with sufficient accuracy and control. In another paper, they begin with mechanical performances and introduce variations according to what their analysis showed. For example, they try synthesizing a Viennese waltz accompaniment with increasing amount of deviation of D_{ii}'s in terms of the typical pattern (first beat shortened, second beat lengthened, third beat left alone). The mechanical performance would have each beat equal at $33\frac{1}{3}$% of the measure duration. The best simulation is probably that with the first interval 25—27% and the second interval 40—42% of the measure [Bengtsson and Gabrielsson, 1983].

It is important to note at this point the distinction that Bengtsson and Gabrielsson make between the inter-attack time (D_{ii}) and the duration from the *end* of the previous note to the attack of the next (D_{io}). It is safe to say that, although D_{io} is quite important to detect (it usually corresponds to rests), D_{ii} characterizes the rhythm, whereas D_{io} the *articulation* of the rhythm. This finding is in agreement with the study called "Timing by Skilled Musicians" by Saul Sternberg, Ronald L. Knoll, and Paul Zukofsky at Bell Laboratories, who note "... dominance of the sequence of time intervals between the *onsets* of successive note (attacks) and the relative unimportance of offset time, which probably serve articulative rather than timing functions." [Sternberg, et al., 1982]. Another finding of this study was that in general, when judging small temporal intervals, musicians typically assigned values that are too large (overestimation), and in both the production and imitation of temporal intervals, they produced intervals that were too large (overproduction).

Gabrielsson used a subjective approach as well in two studies [Gabrielsson, 1973a, 1973b]. Here he played rhythms performed on drum and piano, and asked subjects to make *similarity ratings*, that were then analyzed by a multidimensional scaling (MDS) program called INDSCAL (see [Shepard, 1962] for a description of this method). In this experiment, four bars of each rhythm are presented to the listener, who tries to evaluate the rhythm from a purely subjective vantage point. This does not ignore the objective timing differences between the performances in terms of SYVAR, but rather attempts to identify the subjective dimensions that are invoked by the different styles of rhythm.

In seeking to interpret the "dimensions" found within the subjective space, he suggested the following:

1. "Meter"
2. "Rapidity"
3. "Tempo"
4. "Uniformity—variation or simplicity—complexity"
5. "Basic pattern"
6. "Movement character"

Another valuable contribution is work done by Sundberg and Lindblom, in which they try to *generate* timing fluctuations in a systematic way. They develop a rule-based program that takes standard notation as input, and automatically produces a "natural" performance as output, by introducing deviations from the notation at key places that are recognized by the program. The feedback one gets from this kind of work is very powerful; it is a natural way to test musical hypotheses [Sundberg and Lindblom, 1976].

2.4. Rhythmic Structure and Meter

In this section, we concentrate on theorists who are trying to characterize rhythm *as represented by a musical score, and not by a performance*. In this sense, there is no concern for automatic transcription *per se*, or for performance nuance; instead the focus is on inferences made directly from notation.

Ironically, this is equivalent to an analysis of a performance in which the music is played absolutely unexpressively or mechanically, which might seem uninteresting. On the other hand, when searching for a general theory of meter or rhythmic structure, this is probably a reasonable restriction. There is so much going on in a real performance that we might want to deal with one aspect at a time, and limit ourselves to the score, which is very well-defined as the written intention of the composer. Clearly, this approach is best suited to Western music.

In the next section (2.5), we will put forth a theory that is more global with respect to the music of the world.

2.4.1. Theoretical Studies

It is common to see, in many papers on rhythm and meter, an initial statement that laments the confusion and vagueness that surround the subject of rhythm in music. For example, in the preface of their book *The Rhythmic Structure of Music*, Cooper and Meyer state: "An understanding of rhythm is important for performer as well as composer, for historian as well as music theorist. Yet the study of this aspect of music has been almost totally neglected in the formal training of musicians since the Renaissance." [Cooper and Meyer, 1960]. It is safe to say at this time that there is still room for improvement—one wonders not *whether* the topics of rhythm and meter are more elusive than, for example, harmony and counterpoint, but *why*. It is not easy to establish, at the outset, what rhythm and meter are and how they are related.

Being elusive, the topic can quickly become philosophical. Indeed, two names that appear often in the literature on rhythm are the Greek philosopher Aristoxenus (350 BC), and Hegel, along with Hegelian dialectics. In Western music there are three theorists who stand out, and whose work has stood as a foundation for more recent theorists: Moritz Hauptmann (1792-1868), Hugo Riemann (1849-1918) and Rudolph Westphal (1826-1892). Hauptmann and Riemann (who was a direct theoretical descendent of Hauptmann) based their theories of rhythm on the Hegelian dialectic. In fact, the basic concept of thesis/antithesis/synthesis is elaborated upon as it relates to the realm of rhythm in the writings of both.

Hauptmann develops the idea of the basic metrical form as duple, which is defined simply by two equal units of time in succession. This is enough to imply a continuing beat. Ternary (triple) time is generated as the antithesis of duple time, specifically, two units of duple time that overlap and form an intersection of three. This seems somewhat artificial, and in fact, Hauptmann did not extend it to other prime numbered divisions, and tried to rule out more complex possibilities, like subdivisions of five and seven.

Riemann developed Hauptmann's work, and though he differed significantly in certain ways, he retained the idea of the dialectic as a general driving principle, in that he considered undifferentiated durations as the thesis, divided durations as the antithesis, and subdivisions, or internal groupings, as the synthesis. Riemann makes an interesting statement on the essence of rhythm: "As the essence of the harmonic-melodic element is change of pitch, so the essence of the metric-rhythmic element is change of living energy, of *tone-intensity* (dynamics) on the one hand, and *rapidity of tone succession* (agogics, tempo) on the other."*

Although Riemann had an *a priori* idea that large rhythmic structures are ideally based on some multiples of four, he was willing to see prime number divisions like five and seven as logical primitives rather than complex intersections of duple and triple divisions as Hauptmann suggested.

Rudolph Westphal based his theoretical work on the writings of Aristoxenus, rather than Hegel. Basically his attempt was to apply Greek metrics, or prosody, to the musical domain. The various rhythmic feet are derived from various combinations of the shortest indivisible unit, called the *chronos protos*. Westphal included

---

* This translation is from Yeston [Yeston, 1976] quoting Carl Alette [Alette, 1951] quoting Riemann's *Musikalische Dynamik un Agogic* [Riemann, 1884 p. 10].

the concept of hierarchical structure with respect to the foot: a given foot may be an accented or unaccented segment of a larger foot with longer total duration; this method of analysis is *architectonic*. Because rhythmic groups may exist on a lower level and be part of a higher level, this scheme could be considered as an analogue in the rhythmic domain to Heinrich Schenker's seminal writings in analysis, in which the foreground level is the actual music, and the background is the abstracted structure. (See [Salzer, 1952]). It should be mentioned here that an attempt to translate Schenkerian analysis into the domain of rhythm is not necessarily appropriate.

In the twentieth century, there are at least two important efforts to consider, which are Cooper and Meyer's *The Rhythmic Structure of Music*, and Maury Yeston's *The Stratification of Musical Rhythm*. Both theories are architectonic in principle, but differ significantly in detail.

Cooper and Meyer base their work on Aristoxenus and Westphal. They use poetic feet to indicate accent and grouping, but *not* duration; this seems to be a mistake, because there is no question that duration is a crucial parameter for rhythmic *performance*. They assume that there must be some differentiation between tones, to distinguish between accented and unaccented beats. The hierarchy proceeds from tones to motives to phrases to periods, all in the domain of rhythm. The prosody, or poetic feet, are as follows:

1. iamb: ˘ –
2. anapest: ˘ ˘ –
3. trochee: – ˘
4. dactyl: – ˘ ˘
5. amphibrach: ˘ – ˘

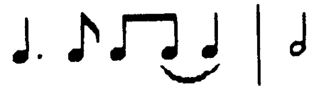
Again, these feet are defined, *not in terms of duration*, but rather *accent*. In fact, Cooper and Meyer basically define rhythm in terms of accent: "Rhythm may be defined as the way one or more unaccented beats are grouped in relation to an accented one." (page 6). Meter, on the other hand, is "the measurement of the number of pulses between more or less regularly recurring accents." (page 4). It is disturbing that they do not define accents, but leave them as undefined primitives.

Yeston presents a considerably more consistent theory, based on two possible methods of analysis: rhythm-to-pitch and pitch-to-rhythm, which means "to value a pitch in terms of its accentual placement (rhythm-to-pitch), while, at the same time, positing an accentual scheme on the basis of pitch value (pitch-to-rhythm)".

[Yeston, 1976 p. 33]. There is something dangerously circular, and yet appealing, about these two approaches; Yeston attempts to decouple them, or to separate rhythmic analysis from pitch analysis. He is critical of Cooper and Meyer's focus on prosody (poetic feet); the problem is that the way they use poetic feet to analyze rhythm is possibly only a relabeling process, and not a way of exposing underlying structure.

In analyzing a piece, Yeston describes five criteria by which one isolates rhythmic sub-patterns. These sub-patterns are crucial to finding the rhythmic strata in a piece. The criteria are as follows:

Attack-point. Derived from the score (as is the whole analysis), this refers to the distance (duration) between attacks, in terms of the smallest local unit. For example,



has the sequence 3 1 1 3 4.

Timbre. Rhythmic sub-patterns are here distinguished by changes in instrumentation or by other timbral shifts, as in an abrupt registral shift in a solo instrument. It is the clearest definer of rhythmic sub-patterns.

Dynamics. Either notated accents, or equivalent change in dynamic level, can determine an uninterpreted rhythmic sub-pattern. By uninterpreted, we mean that we as yet do not assign any strong or weak beats (accents) inside a given sub-pattern, but simply identify it as a single level of interpretation as implied by dynamics.

Density. Refers to changes in either the "quantity" of the sound or the number of simultaneous voices in the overall texture.

Pattern Recurrence. Here, one looks for patterns of repeating units, either in repeated duration figures, or pitch contours, or combinations thereof. This criterion is somewhat open-ended, in that recurrent patterns may be created by various interpretations.

Once the previous five criteria have been applied to a piece of music, one can follow Yeston's analysis. Initially, one views the music as an "uninterpreted structure," by not yet establishing any internal groupings. The rhythmic subpatterns identified by the five criteria are used to create the first stratum, from which the structural events are abstracted. At each level, the structural events are given the combined durations of the events from which they are abstracted. Again at each level, the new representation is considered as a new uninterpreted structure, and the process is repeated.

The levels found, from the foreground (the original piece), through several middleground levels, to the most abstracted background level, can be said to interact perceptually. Yeston sees the interaction of these levels as the determining factor in determining meter. In fact he says "a meter will never appear on any single stratum, but it will arise from the interaction of two strata, one of which must always be a middleground level." [Yeston, 1976 p. 67]. Yeston provides numerous examples of his analyses in the form of excerpts of scores with annotations.

The topics of rhythm and meter are obviously still open. One can muse over comments such as this one from the ethnomusicolgist Jaap Kunst:

> "Metre is the rational analysis of the living rhythm." [Kunst, 1950 p. 2].

This statement generates thought but not clarity!

2.4.2. Computer Simulation Studies

The previous subsection described entirely theoretical work. Here we examine some attempts by Longuet-Higgins et al. to automate an analysis from the score. They have tried to formalize their ideas as computer programs whose input is a score and whose output is an analysis, especially of rhythm.

In an article in *Nature* called "The Perception of Melodies," Longuet-Higgins describes an initial attempt to actually *create* the score, beginning from a performance on an organ keyboard that has been modified to record the history of each key [Longuet-Higgins, 1976]. In this article, he makes many assumptions, and requires the performer to establish the tempo and meter beforehand by playing one measure's worth of preliminary beats before the musical example begins. Longuet-Higgins describes the rhythmic structure as a binary (or sometimes ternary) tree, each terminal of which is either a note or a rest. Such a tree, in an obvious way, reflects the structure of Western musical notation.

Longuet-Higgins assumes that "the perception of rhythm and the perception of tonal relationships can be viewed as independent processes." This is in agreement with Yeston, although one would want to allow a level of interaction between these two factors at some point. Longuet-Higgins claims that 100 msec. accuracy in timing data is close enough to do the transcription. This is probably far too gross; in general one wants to be able to deal with very fine decisions in certain notational contexts that will be obscured by such a wide tolerance. He is possibly confusing "expressive" performance with the need to "factor out" expressiveness and local tempo variation (see Section 3.3).

Longuet-Higgins further assumes that the listener expects pure binary meter, which is absurd. He *does* allow for a change at any point, but there is no reason to assume a duple meter from the start, except for the preponderance of duple meter in Western music. It is not clear what the perceptual mechanism is for establishing meter in human listeners, most likely there is "suspended judgement" for a few beats during which the listener tries to parse the prosody and melodic and rhythmic patterns of the music before making a decision. Longuet-Higgins' program probably starts out assuming the music is duple until proven otherwise. Evidence that this is natural for listeners is lacking, though it may function satisfactorily in the program. It would not be worth mentioning this were it not for the emphasis Longuet-Higgins himself places on modeling human perception.

In the article, he also describes some algorithms used to produce the correct "spelling" of the notes in a given key or keys related by modulation. This process, like the rhythmic one, proceeds from left to right, forming an hypothesis about the key, and retaining it until it is violated by later chromatic pitches.

Mark Steedman, a student of Longuet-Higgins, tried to infer the meter of unaccompanied melodies automatically by using a computer program that is given as input the 48 fugue subjects of the *Well-Tempered Clavier* by J.S. Bach. He describes this effort as being complementary to the above article, in that Longuet-Higgins' program needed the absolute position of at least two principal beats at the beginning of a piece, whereas Steedman's program tries to deduce metric relationships "on the fly," i.e. without a hint about meter or tempo, but from the notation (score) rather than the performed music. This idea of the rhythm "unfolding," particularly from the standpoint of the listener, is appealing, though it may abandon an important possibility for analysis from the score itself—being able to make inferences from the entire piece at once, in a more *structural* way.

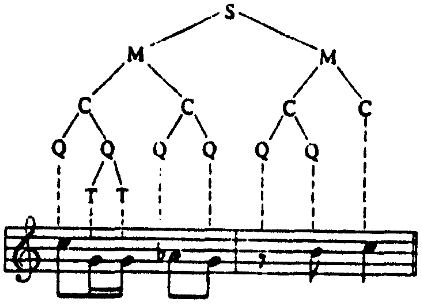
Steedman's program deals with syncopation, which he claims is possible to analyze by computer because his program relies on establishing enough metric context in the music *before* syncopation appears. Otherwise, he claims, it would be impossible to distinguish between a true syncopation and an entirely different metric or temporal context. This would, for example, be the way to tell the difference between an eighth-note pick-up and an identical meter an eighth-note out of phase with the original. He says: "No event inconsistent with either key or metre will occur in a piece until sufficient framework (of key or time signature) has been established for it to be obvious that it *is* inconsistent." [Steedman, 1977 p. 557].

The problem with such a view is that people seem to need very little framework to distinguish between consistency and inconsistency. Also, one has to allow for changes in key, meter, and tempo, and the possible variations can be quite subtle. In a transcription program, a change from eighth-notes to triplets in a sufficiently inaccurate performance could be misrepresented as an accelerando instead of a constant beat with different subdivisions.

The simplest way to create context the way a listener would is to work through a piece from left to right, as mentioned above. This is how Steedman's program proceeds. Because there are sometimes sections of a piece in which the rhythm is constant (a "pulse train" or "rhythmic tonicity"), the meter has to be implied in these areas by some other means, for example melodic patterns. (Actually, as seen in Yeston's theory, there are several other possible ways.) Steedman's program makes two passes through the data, the first searching for rhythmic patterns, and the second searching for melodic patterns, or repetitions of some fragment of the melody. He sets the metric grouping implied by a repetition to be equal to the duration between the figure and its repetition, and he assigns accents to their first notes.

The fact that the program cannot recognize rhythmic embellishments is a rather severe limitation. It is quite common to see a pattern with slight elaboration that is easy for the performer or experienced listener to recognize, but the Steedman program would see it as entirely new, and not be able to build a metric context from it. Still, there is enough consistency in his examples that he can say that his program is usually correct in its metric analysis (for the Bach), and when it is wrong, he reports that most but not all of the errors occur where there is sufficient ambiguity that human listeners might err also.

In an article called "The Perception of Musical Rhythms," [Longuet-Higgins and Lee, 1978], their research is continued, strictly in the domain of rhythm. They state that in many cases, the rhythm of an example is evident merely from the duration pattern of the notes (without any melodic information). This is certainly true of percussive music. They reiterate the basic idea of meter as a generative grammar, representing a simple example as a binary tree:



They claim that the listener creates this tree while experiencing the music: "The listener must identify the metre which generates the rhythm, and represent the latter as a tree structure which accommodates all the notes and rests as terminal symbols. The goal of a theory of rhythmic perception is then to explain how the listener accomplishes this task." (Page 3).

The principal focus of this attempt is to be able to deal with real performance data instead of mechanical (from the score) timing, by extending the program to make metrical inferences on *relative* note durations. Though this is a goal, at the time of writing of their article the program was still dependent on mechanical, or notation-derived durations.

2.5. Steps Toward a Global Theory of Rhythm

In the previous section, there was a natural bias toward Western music in the theoretical approach to rhythm. In this section, we make an attempt to widen the perspective considerably—we try to orient the reader to rhythm in music from a global perspective. We are trying to account for the diversity found in world music in terms of rhythmic paradigms. In the proposed scheme, any rhythmic music can be identified as belonging to one of four categories described below.

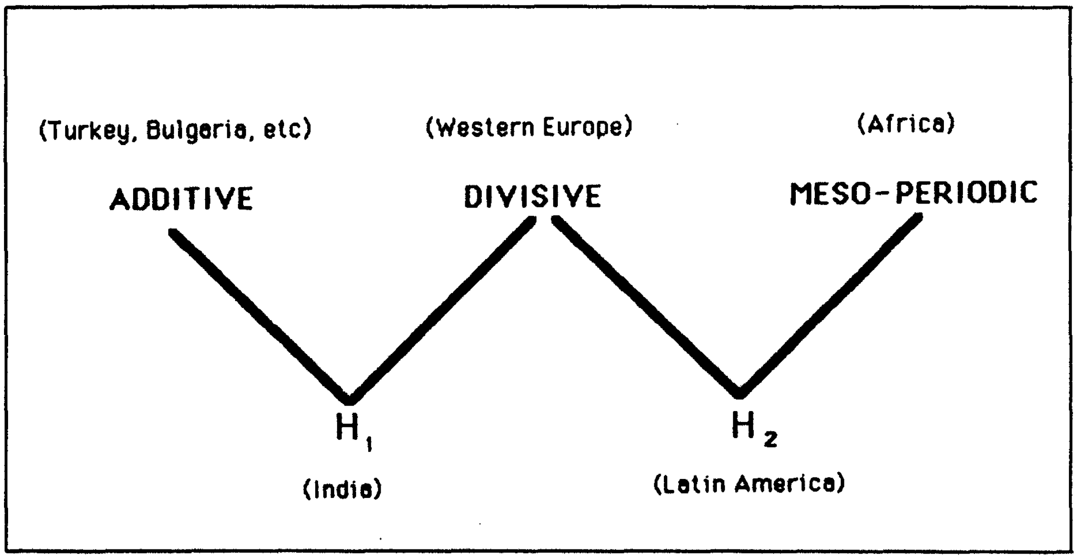


Figure 2.2. The Rhythmic Categories.

2.5.1. The Proposed Paradigm

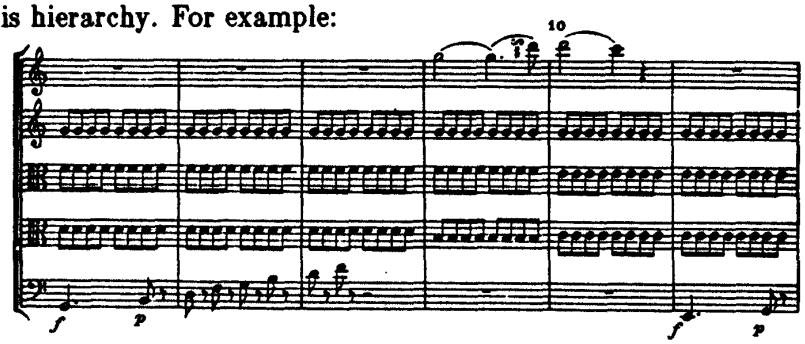
Interest in rhythm has a long history. Plato defined rhythm as "ordered movement." Obviously, in music there are ways of "marking time," of creating temporal structure by making durations evident. As with perception of pitch, it is relative durations that are the salient percept, rather than absolute durations. This will be true in any rhythmic form, but the *mechanism* by which events are parsed in the rhythmic domain can be considered in the following categories:

1. **Divisive**
2. **Additive**
3. **Meso-periodic**
4. **Hybrids of the above**

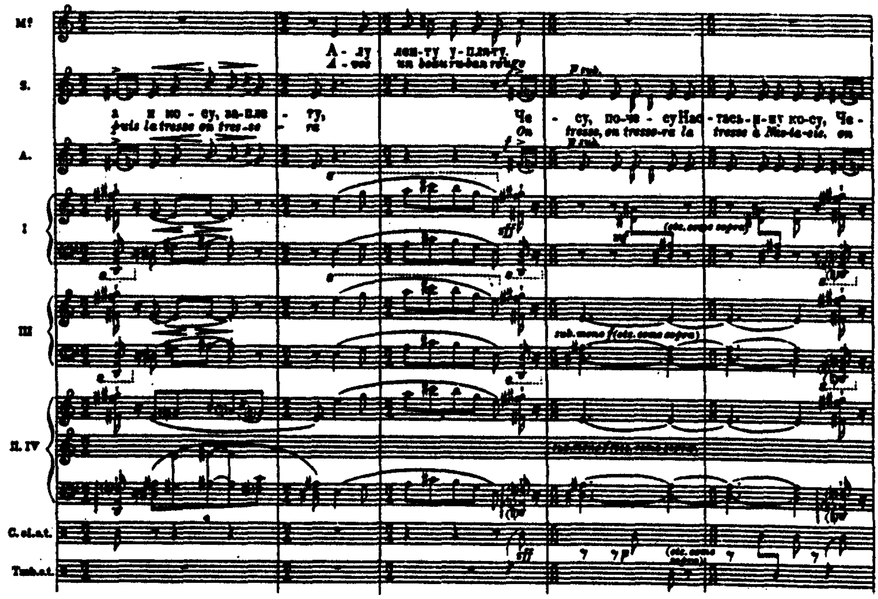
Figure 2.2 shows how these categories are related, and the following musical examples are intended to illustrate the nature of the categories. Notation is provided in lieu of aural examples. It would be preferable to listen to the examples to discern how they are constructed, especially the non-Western examples for which the notation is not necessarily an accurate representation. (See Appendix A for information on how to request taped examples from the author.)

1) **Divisive**—Western music: European Classical music, Pop music, etc.

The best example of this category is the march. Divisive music is constructed rhythmically such that beats are divided into parts, usually as a binary or ternary tree, that is, into halves or thirds, hierarchically. Typically, the beat is simply one of the nodes of the tree, with subdivisions going out to the leaves, and the meter is a higher node of the tree. (See previous section for an example of Longuet-Higgins.) Almost all Western music is constructed in this way. It is clear that western notation reflects this hierarchy. For example:



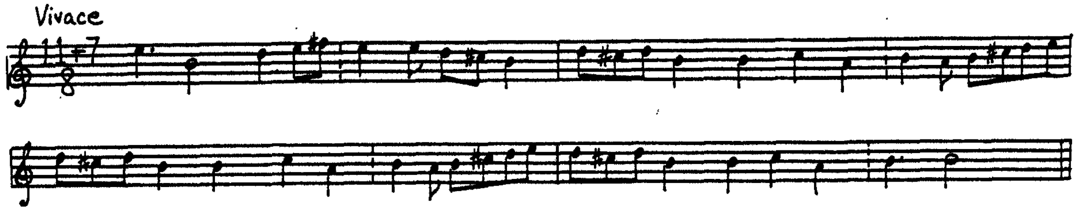
Mozart: *Quintet in C, k* 515. This example has an obvious rhythmic pulse—"tonicity" with some rubato, and some syncopation. It is very easy to tap your foot.



Stravinsky: *Les Noces*, 1917. This example has much more complex rhythm than the previous example of Mozart. Stravinsky based this piece in part on folk music. The sense of beat is still evident, but with unexpected accents and numerous abrupt changes of meter. The notation looks more complex than the music sounds.

2) **Additive**—Turkish, Greek, Bulgarian, Eastern European music.

This category of music is constructed of small "atoms" of duration. They are put together to form longer rhythms and melodies; the music is almost always a concatenation of two's and three's. The underlying pulse is usually fairly constant, and quite rapid (more than a pulse rate of MM = 300). If one taps one's foot, it is to the uneven size of the atoms; they are not subdivided. In fact, in Turkish, this rhythmic idea is called ***aksak***, which means 'to limp' (i.e. an irregular beat pattern). Kurt Sachs described the difference between divisive and additive rhythm in this way: "Divisive rhythm shows how the parts are meant to be disposed. It is regulative. Additive rhythm shows how the parts are actually disposed. It is configurative." [Sachs, 1953 p. 25].



Bulgaria: *Jove Malaj Mome*. This has 8 clear "beats" but they are of different lengths (not regular), so it is quite difficult for us to parse. This music is not subdivided! It is actually based on an additive scheme of:

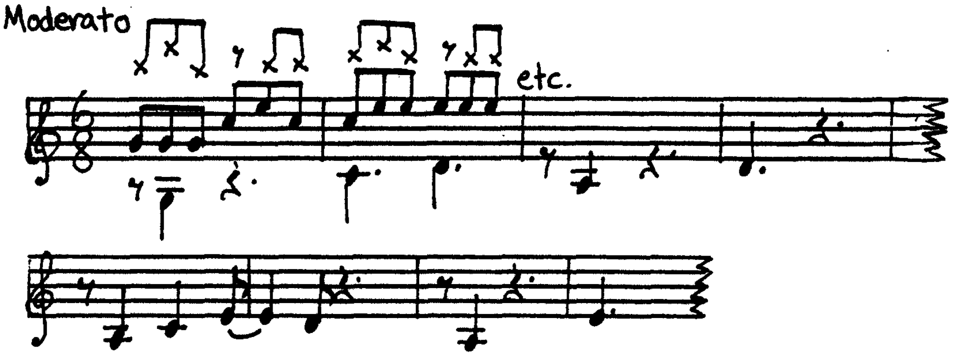
18 = 11 + 7 = / 3 2 2 2 2 / 3 2 2.

3) **Meso-periodic**—African music.

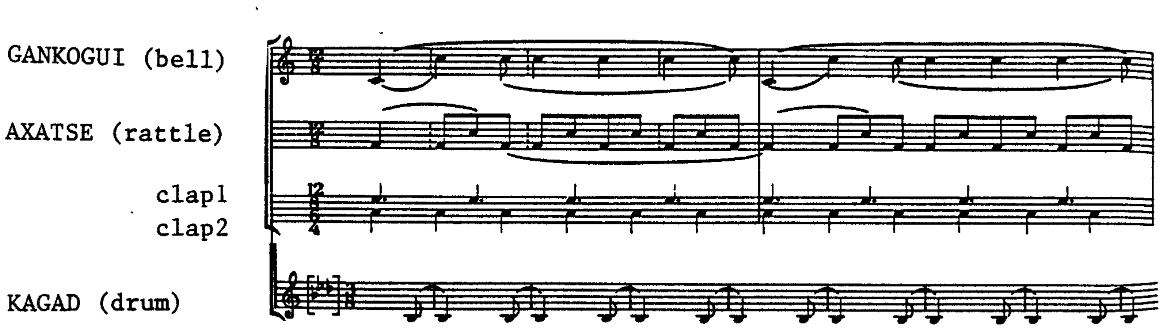
Music involves periodicity at many levels, from the signal itself to high-level structure. In the case of African music, there is a very fertile middle-ground temporal level that is based on what I call "meso-periodicity," typically a 1–4 second long pattern. The pattern is repeated thousands of times, with very small variations in two modes: 1. rational deviations from the pattern (embellishment), and 2. minute timing deviations from canonical pattern ("floating").

These varations can be introduced by a single player, or more typically, by several players who deviate in very small amounts from their given patterns, resulting in the bimodal deviations described above. This creates a succession (in varying temporal scope) of tension and release, which is what allows an endlessly repeated pattern to remain interesting. The repetition of this single period is fundamentally different from the other two forms, in which there is a metrical structure supporting the other aspects of the music. In African music, the meso-period *is* the focal point of the music.

This category is possibly the most subtle rhythmically. It is played in reference to movement, and not in reference to a pulse; that is, there is not a hierarchy of subdivided beats, but rather a parallel stream of voices (drum parts) that are "woven" together, in interlocking polyrhythm. More information about African percussion and how it is constructed is contained the next subsection and in Section 2.6.



Zimbabwe: Shona people. This is mbira music (African "thumb piano"). Obviously it is periodic, yet it is difficult to tell where "one" is (the beginning of the cycle). The low melody is offset from the downbeat, which makes the period seem to begin late, and creates ambiguity because of its relationship with the high part and the rattle. There is also a simple polyrhythm between the rattle and the other parts, in a basic two against three pattern. In contrast to the previous example of Stravinsky, the notation looks *less* complex than the music sounds. This is true for most meso-periodic music.

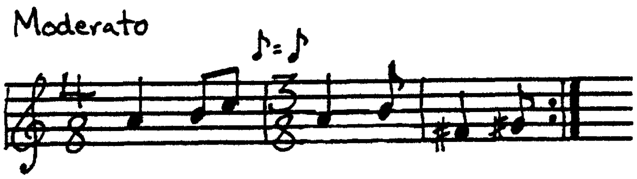


Ghana: *Sogo* Dance (*Ewe* people). Again, there is obvious meso-periodicity, and yet it is difficult to parse when one hears it. There are several possible ways of hearing the parts, and of the polyrhythms between the instruments.

There is another aspect of interest about this category of music, which is the existence of an underlying rapid pulse, (not a beat!), that I will call the "density referent" to which all parts relate. Though the interlocking parts may sound very slow in comparison to the speed of this implied pulse, the density referent is the common temporal denominator that keeps them together. There is much combinatoric complexity in the way the density referent is sampled. See the next subsection for an exposition on the possible relationship between the meso-period in African music, and the chromatic scale in Western music. It turns out that they are related in a surprisingly direct way.

4) Hybrids.

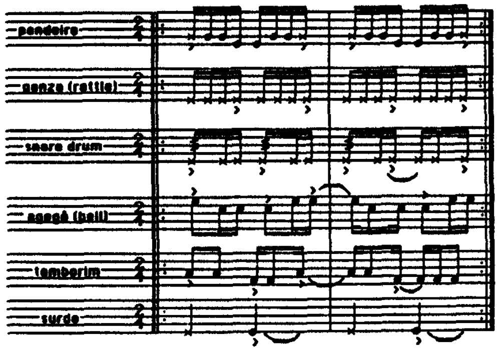
The most distinguishable hybrids are Indian music (hybrid of Additive and Divisive) and Latin music (hybrid of Divisive and Meso-periodic). In the case of Indian music, it is evident that there is a practice of subdividing a pulse, but also odd-numbered sequences are used as indivisible bases. In Latin music, there is a strong duple meter, with the meso-periodicity still evident.



Indian drumming: Alla Raka and Zakir Hussain (*tablá*). This is the harmonium accompaniment to the drum, which outlines the basic rhythmic cycle, called *tala*. Indian music is rhythmically a hybrid of 1 and 2. This example has 10 beats, parsed as 4 + 3 + 3. The first 4 is divided as a duple structure, but the following two groups of 3 are not subdivided; they are atoms as in the additive paradigm.

Some devices for creating tension in this music are:

1. Avoiding downbeat
2. Increasing density while narrowing tolerance to downbeat
3. Embedding patterns within patterns The peak of tension is reached just as the longest pattern finally ends on the "downbeat," or the first beat of the *tal* (rhythmic cycle), and the density is at a maximum. Just at this point of resolution, the density is greatly reduced and the process begins again.

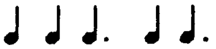


Brazil: *Batucada*. This is a hybrid of 1 and 3, even more clearly than in the previous example. It is obviously in a march-like a duple rhythm, but within that scheme, it still has retained many "Africanisms" in both its structure and methods of variation. Noise is everything in this music. There is no melody; instead, the extremely diverse percussion instruments cover an enormous bandwidth, from below 40 Hz to the limit of hearing. Each instrument itself has a wide spectrum, overlapping the others both in spectrum and in rhythmic patterns.

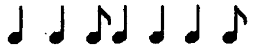
2.5.2. Anatomy of the Meso-period

We now focus on the third category—the meso-period. After some scrutiny, rather interesting structure appears, which is not surprising, given that the music has evolved over many generations with its form evolving and strengthening through oral tradition. This kind of music is based on repetition of patterns with very strict rules, and it would be impossible to have a vital form of music based on repetition at this temporal level (the meso-period of about 1–3 secs.) were it not for some rather deep cognitive foundations. That is, unlike Western music, which has what might be called "macro-periodicity," (A-B-A, sonata form, etc.) or ways to create larger melodic and harmonic structures, in African music the meso-period sustains the music through time both by variations in the patterns, and the way the meso-period is actually constructed.

In any survey of West African drumming, one finds that, embedded in the polyrhythmic structure, there is a rhythmic "skeleton" that is usually a bell-pattern* based on special subsampling of a 12-pulse meso-period. All the other rhythmic strata are conceived and played in relation to the bell pattern. There are many patterns used in African music, but by far the two most common are:



AND



A.M. Jones singles out these two patterns in his classic work *Studies in African Music*, as being the most important throughout Africa:

---

* The bell-pattern is a repeated figure played on a metal bell with a stick; it is the "glue" of the rhythmic ensemble.

> **"This pattern is sometimes made by hand-clapping, sometimes it occurs as a bell-rhythm, and it is even played on the drums. It occurs in various forms but always it is basically one and the same pattern. It is found widely in West Africa, in Central Africa, and in East Africa. In fact both its ubiquity and its typically African form qualifies it to be called the African 'Signature tune' . . . this particular pattern is very deep down in the African musical mind and is indeed part and parcel of their music."**

This pattern, which defines the meso-period, is not perceived (or generated) the way a Westerner typically deals with rhythm. As Jones later says:

> **"No one who has heard a party of villagers clap this pattern could possibly think that there was the slightest suggestion of syncopation in it, that is, the suggestion that it is 'out of step' with some primary background existing in the performer's mind. It simply sounds like a smooth irregular pattern existing as a complete conception in its own right."**

[Jones, 1959 pp. 210, 212, 224].

If we represent the above two patterns in terms of relative durations summing to 12, we have 2 2 3 2 3 and 2 2 1 2 2 2 1; it turns out that these two patterns are in fact complements, in the sense that if we play one and tap all the "missing" beats, we will get the other (figure-ground reversal). It seems rather extraordinary that if we think of these sequences as semitones of the chromatic scale, they exactly represent the pentatonic and diatonic scales, respectively. For example, the bell pattern (2 2 1 2 2 2 1) corresponds exactly to W W H W W W H, which is none other than the diatonic scale (W = whole step, H = half step). If one looks at this as a period, and changes the phase (this is equivalent to starting the pattern in another place), the rhythmic patterns still exist, with a different reference point. Similarly, in the domain of scales, this difference in "phase" corresponds to the Greek modes, for example, Dorian, Phrygian, Lydian, etc. Interestingly, if either the scales or the rhythms are played by themselves, there is no frame of reference, and no way of distinguishing the modes, rhythmically or tonally. But if there is a frame of reference (usually another part, or some reference to the "tonic"), then the phase can be critical to a correct performance or interpretation of the rhythm. See Fig. 2.3 for a graphical representation of these patterns, and how they relate to one another.

One wonders if this correspondence is sheer coincidence, or whether in fact

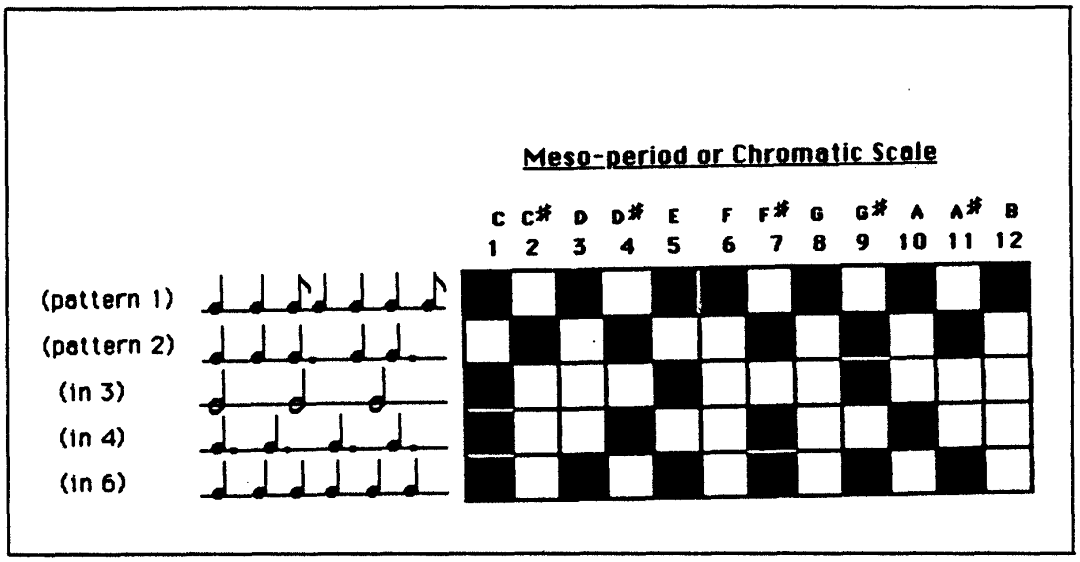


Figure 2.8. A graphical representation of the basic bell-patterns of the meso-period. Note that the patterns are embedded in a basic grid of 12. Patterns (1) and (2) are subsamplings of the period. They are complements of each other, and as they fit in the period of 12, they can be perceived in groups of 3 or 4. Thus there is a basic "instability" in terms of the way one parses the patterns. Note also that if the grid were viewed as the twelve steps of the chromatic scale, then pattern (1) is the standard diatonic scale, and pattern (2) is the pentatonic scale. Other rhythmic or melodic modes are created by starting in different positions in the period. This representation is in some ways more useful than conventional notation, because it shows the relationships between the patterns clearly. Each box represents the density referent, i.e. the fastest regularly occuring pulse. James Koetting calls this the TUBS notation (Time Unit Box System), and uses it to represent various kinds of percussive music [Koetting, 1970].

there is some natural combinatoric/musical link between these two systems of music. It turns out that there is an approach to the meso-periodic structure of African music which both explains its powerful musical basis, and links it with the Western notion of scales. Gerry Balzano wrote an article in the Computer Music Journal in which he investigated the implications of treating the chromatic scale as a cyclic group of order 12 and its subgroups [Balzano, 1980]. When interpreted in this way, the cyclic structures common in African music can be seen in many cases as isomorphic to the pitched cyclic structure of chromatic, diatonic and pentatonic scales [Pressing, 1979]. The argument for this particular choice of subgroups of the

12-grain cycle is comparable to the theory put forth by Balzano in his investigations into scales in Western music. In fact, it seems more appropriate because Balzano is forced to ignore the *actual tuning* of intervals for his theory; he only discusses the number and placement of discrete pitches per octave. By contrast, in reference to the meso-period, we are actually representing relative durations, so we are not throwing away information when we represent the sub-sampling of the meso-period as a pattern.

Another of the elements that make the meso-period come to life is ambiguity of interpretation, in the existence of several possible meters that one might perceive. Since there are several divisors of 12, there can be several possible interpretations of the period. It is not clear which divisor will be heard as dominant, or put another way, how does one hear (parse) simple polyrhythms? In a very much distilled experimental context, Oshinsky and Handel tried to see which pattern will be heard as dominant when listeners are presented with clicks that are in the relationship of 3 against 4. They found that in the case of 3 against 4, the choice was tempo-dependent, that is, whether listeners tended to hear the 3 as the dominant pattern, or the 4 as dominant depended on the rate of the clicks (tempo) [Oshinsky and Handel, 1978]. Their experiment was not in a musical context, however, in that they used undifferentiated clicks that were all identical in amplitude and envelope.

2.6. Issues Peculiar to Percussive Music

Because this thesis is concerned with transcribing percussive music, it is worth examining two areas that are specifically related to percussion: the physical/acoustical, and the ensemble issues that are peculiar to African or African-derived music.

2.6.1. Acoustical Considerations

The term "percussive music" is perhaps a misnomer, as it includes instruments that should perhaps not be classified together. For instance, the piano is a percussion instrument, and so is the "slit drum" (which is not a drum at all, but rather a gong, in the sense that it is an *idiophone**).

In a study of percussive music, one might begin with the Western orchestral percussion instruments. Of these, the most prominent is the timpani, or kettledrums. These are large drums with plastic or animal (usually calf) skins stretched over a metal or fiberglass kettle, or shell. The acoustics of the timpani have been described briefly by Thomas Rossing, including the reasons these drums elicit a reasonably clear pitch, when in fact they should be quite inharmonic [Rossing, 1982].

There is a problem with the reasons however, because Rossing's arguments could apply to many different drums, but there is a tremendous diversity of pitch clarity, from extremely vague to extremely sharp in various different drums. In fact, the complex mathematical treatment of vibrating membranes [Morse and Ingard, 1968], although elegant and true in the abstract, makes so many assumptions about the "ideal membrane" that it reduces most drums to the same imaginary membrane. The fact is that drums differ enormously in timbre and pitch clarity. Minute differences in shell size and shape, and the thickness and type of membrane, contribute a great deal to the unique sound of different drums. The particular sound of a drum is crucial to its rôle in the associated musical tradition.

The mallet instruments (marimba, xylophone, vibraphone, glockenspiel), are more like the piano in their repertoire, and will not be dealt with here. Because they are acoustically vibrating bars (they are idiophones with individual resonators for each bar), instead of strings, they will have different acoustical properties, but they have the same physical layout as the piano.

Orchestral percussion instruments, because of their position in Western music, are rarely played as solo instruments. If we choose to look at musics in which percussion plays a more central rôle, we will quickly gravitate towards certain non-Western cultures, especially African and African-derived musics, in which percussion and rhythm are the key to the entire musical structure and vocabulary, and in which aspects of rhythm have reached their highest level.

---

* The standard classification of instruments follows the Sachs-Hornbostel scheme of four major categories: 1. *Idiophones* (the vibrating material is the same object that is played (free of any applied tension), e.g. woodblocks, gongs, etc.) 2. *Membranophones* (the vibrating material is a stretched membrane, e.g. drums) 3. *Chordophones* (the vibrating material is one or more stretched strings e.g. lutes, zithers, etc.) 4. *Aerophones* (the vibrating material is a column of air, e.g. flutes, oboes, etc.).

One important difference between African and Western percussion is the notion of a "stroke-space," which can be defined as the universe of possiblities of ways of striking the drum; it is the vocabulary of strokes. Although in Western percussion the emphasis is usually on uniformity of tone, in African percussion, *how* the drum is struck is almost as important as *when*. For transcription, it is important to be able to distinguish between different strokes automatically, as we show in the next chapter. The importance of notating different strokes is stressed by A.M. Jones, who says:

> "African drummers vary not only in pitch but also in quality. If the wrong quality of note is played in any particular African drum-pattern, that pattern is no longer what it is intended to be and becomes another pattern. So it is much more important in the case of drumming, to know which hand a performer used for any given note, and how he played that note, than is the case with Western music. The score should ideally show not only the rhythm and the pitch of the notes but also *how* each note was made."

[Jones, 1959 p. 11].

One interpretation of this is that the idea of "tonality" in Western music can be modified to apply to completely different musical contexts. Instead of referring to scale patterns and harmonic structure, one can think of any organization of the sonic material that allows recognizable structure to appear; that is, permutations or reorderings of a basic set that allows "features" to be perceived. For example, for inharmonic instruments, where standard pitch perceptions do not apply, this could be a sequence of different textures, or different amounts of damping, which is recognizable as different decay rates, or overall amplitude contours.

In the case of drum ensembles, the tonality can be thought of as the different strokes that imply different rates of damping and different spectra. The musical material is the ordering and overall periodicity of patterns made up of these elements instead of notes in a scale. There may be a "tonal center" in this set of objects, but perhaps not a hierarchy as we think of with regard to a musical scale. Still, similar pattern-seeking algorithms may be applied to this material, looking for the equivalent building-blocks and repeated figures, structures one would find in tonal music. Expectation is created by repetition as well as by form. It is interesting to note that the patterns created in drumming music are in some sense more abstract than melodic music is. That is, there is ultimately a somewhat different mechanism for perceiving the drum patterns than for melody.

In the example that is analyzed in Chapter 3, there are eight basic strokes. We now describe these strokes that the program determines to be distinct sources. These stroke-types could be considered canonical in that almost all drumming will have at least a subset of strokes that resemble these basic strokes. They are as follows:

OPEN - high or low drum. (Hand snaps away from drumhead, allowing maximum ringing of drumhead in normal mode.)

MUFF - high or low drum. (Hand "sticks" to drumhead, damping and also raising pitch of tone by about a minor third.)

BASS - high or low drum. (Palm of hand hits center of drumhead, causing lowest perceived pitch.)*

SLAP - high or low drum. (Hand hits center of drumhead while damping edge, causing sharp attack and higher perceived pitch.)*

It is safe to say that one must at least be able to distinguish these strokes to make an adequate transcription. For plots of typical examples of the time waveforms of these strokes on a conga drum, see figures Fig. 2.4 — Fig. 2.7. The associated spectra are shown in Section 3.2.7.

2.6.2. Ensemble Considerations

The communal nature of African or African-derived music is well-known. This refers to the tendency to play the music using several musicians (sharing the parts that add up to the piece) when it could physically be played by one person. Instead, the parts are "woven" together, that is, they are interlocked to form one overall melody or rhythm. An anthropologist would probably see this as an example of music imitating life, that is, in a communal society many tasks will be naturally put together in this way.

This socio-musicological approach is probably valid; however, this way of constructing the music also has an interesting, nontrivial, and strictly *musical* effect:

---

* These strokes elicit a very unclear pitch, but there is a definite relative percept of "highest" and "lowest".

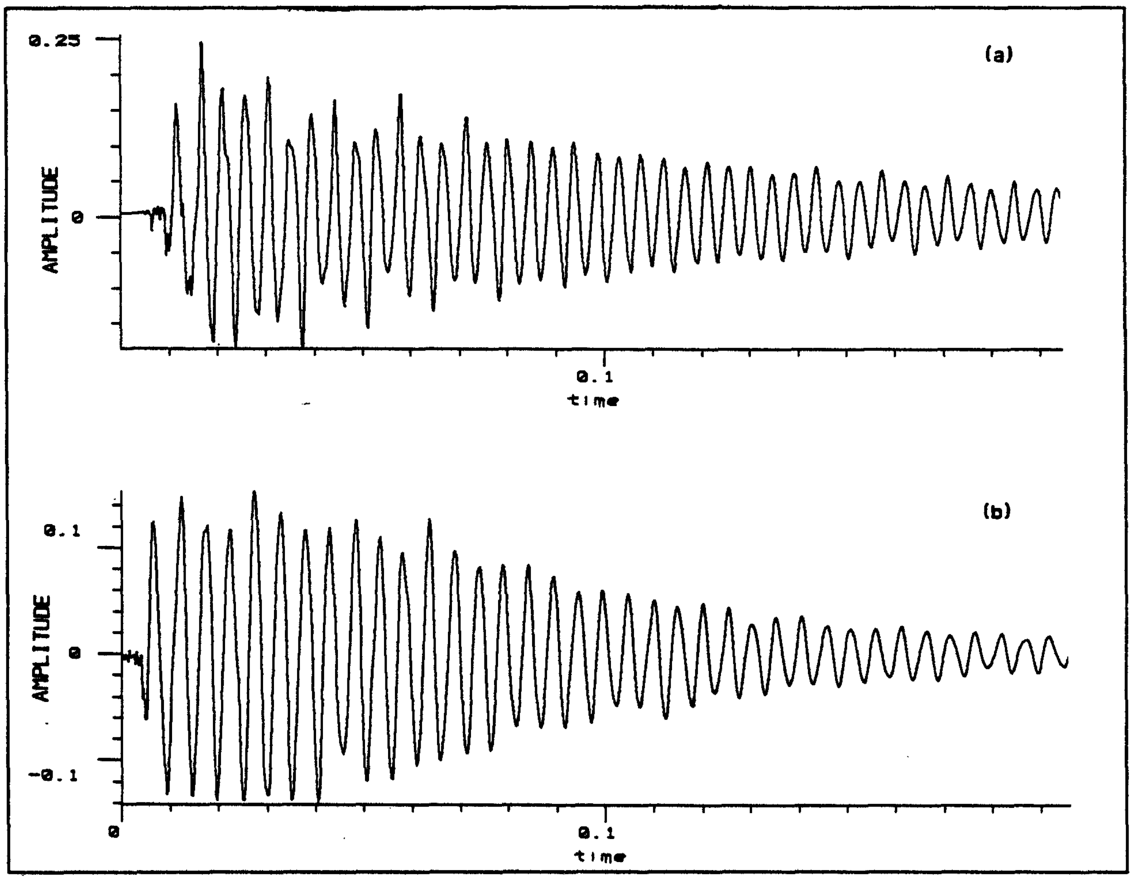


Figure 2.4. Time waveform of open tone on
a. high drum
b. low drum.

it creates tension. This is because the separate parts that fit together to make a whole are each subject to the same kind of intentional and unintentional slight variations in timing. Because each player represents an independent source that is nevertheless in a feedback loop with all the other players (via the players' ears), the overall pattern moves in a rather complex way around the intended music. This allows the music to "breathe," and imparts more life to the tone and rhythm. The liveliness is thus probably a result of:

1. Timing complexity—the interlocked parts are floating in independent ways, but are synchronized by the players.
2. Spatial and tonal complexity added by the actual distance between players.

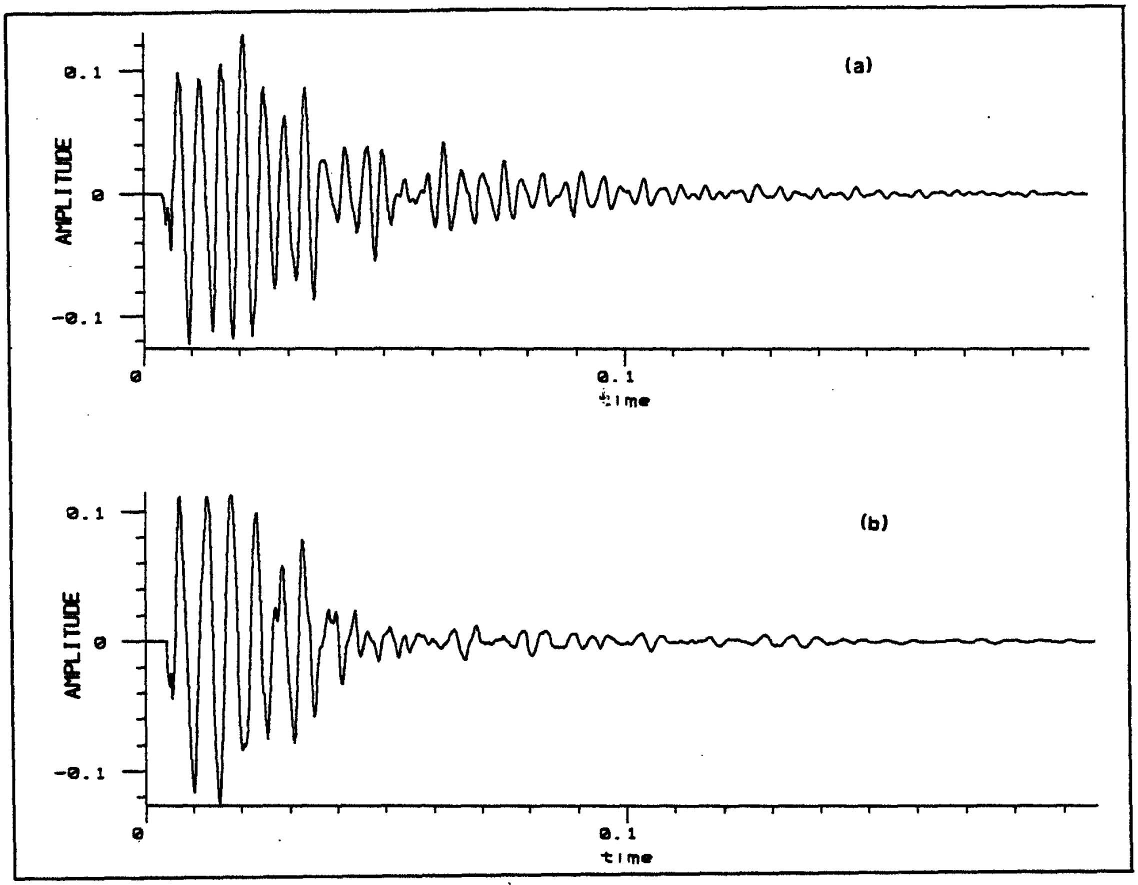


Figure 2.5. Time waveform of **muff** tone on

a. **high** drum

b. **low** drum.

This is one reason why recorded music is not as interesting as live music (the speakers cannot adequately represent the space that the players occupy).

Each player needs to deviate only *slightly* from his or her pattern in order to create an overall impression of fluidity and constant flux. This is very important—only the soloist (if there is one) will deviate significantly from a given pattern, and yet the cumulative effect of all the tiny variations is quite striking.

As mentioned earlier, this kind of music depends heavily on repetition. This is the meaning of the term meso-periodic. Aside from the timing variations just mentioned, there are polyrhythms, or cross-rhythms that are embedded in the

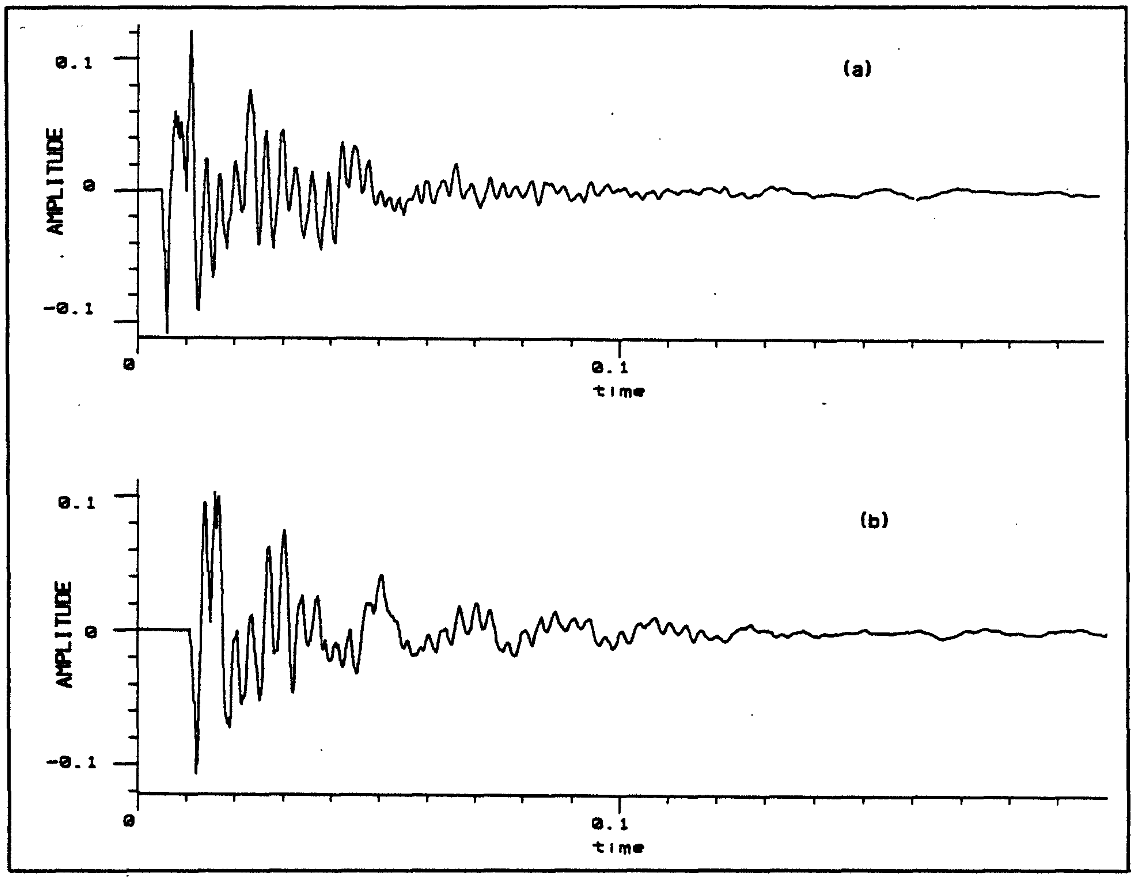


Figure 2.8. Time waveform of bass tone on
a. high drum
b. low drum.

music. Typically in this music, there will be at least two ways to parse the meso-period, and they compete for the listener's attention. It is the abstract quality of percussive music (it is not melodic), that promotes the percept of polyrhythm as a salient feature. As mentioned earlier, this is another way to create tension.

A.M. Jones, who lived as a missionary for twenty-one years in Northern Rhodesia, wrote *Studies in African Music* based on both his observations in Africa and his sessions with Mr. Desmond K. Tay, who was a master drummer from the *Ewe* tribe in Ghana. The experiments were carried out at the School of African and Oriental Studies in London, where Mr. Tay was requested to try to play traditional drum ensemble music in separate parts so they could be transcribed.

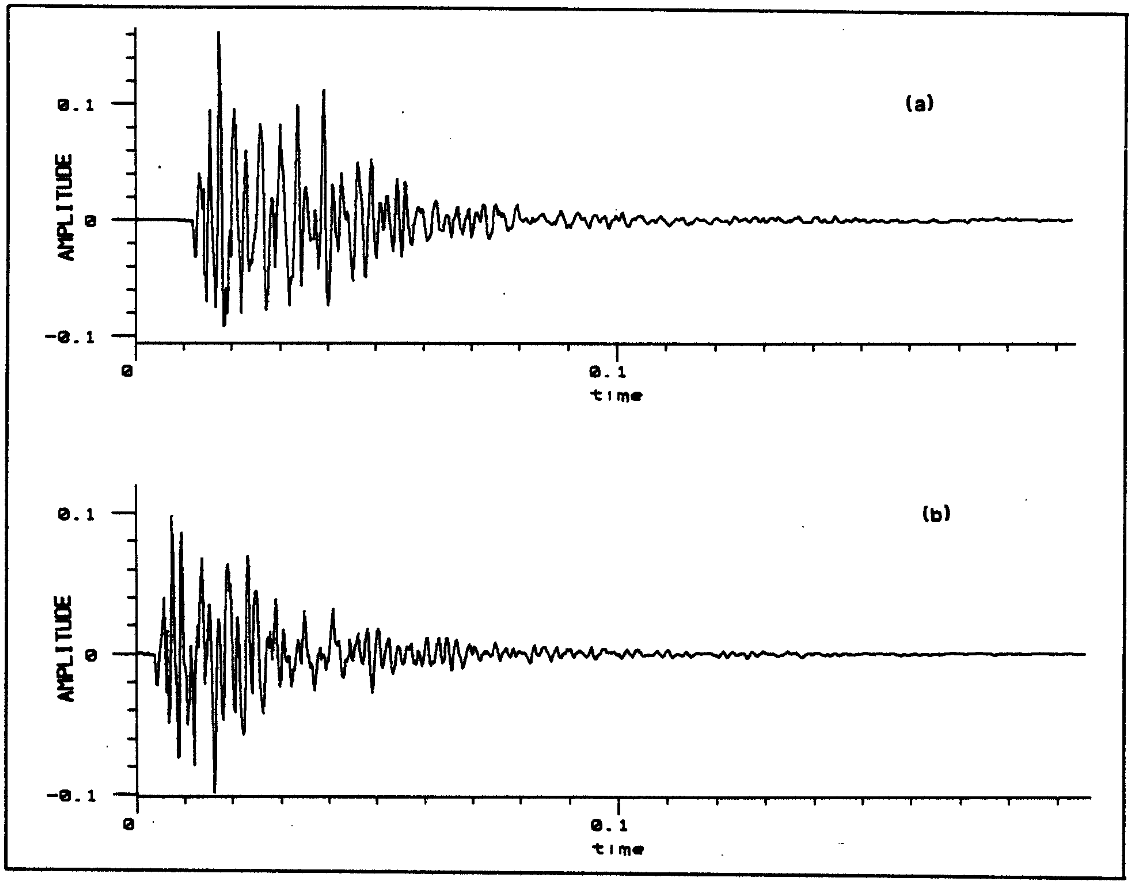


Figure 2.7. Time waveform of slap tone on
a. high drum
b. low drum.

To do the transcriptions, A.M. Jones designed and implemented an apparatus for analyzing African drum patterns. The machine was quite simple; the drummer taps on a metal plate with a metal rod, and each tap of the plate completes an electrical connection that then makes marks on a paper roll moving at a constant rate of speed. The time patterns correspond to the intended rhythms. It can be seen that this method is similar in principle to Seashore and his colleagues' attempts to build the "piano camera," in that the machine, if is working correctly, will mark distances that correspond to exact timings. No attempt is made to evaluate the timing information automatically; that is done by hand at a later date, but at least the raw temporal data should be dependable.

Jones was able to transcribe separate parts by having the drummer play them one at a time, with a common bell-pattern being played for a reference beat. Since the device gives no aural feedback, the efficacy of the experiment depended heavily on the musicianship and tremendous aural imagination of the master drummer. This is because great accuracy is required, and the player does not have the other parts to refer to while playing each part; he is translating and condensing a full performance into this restricted laboratory situation. Also Mr. Tay was asked to report how the drum is struck for each note, as the machine is incapable of making any distictions between strokes. Finally Jones was able, using the machine, to transcribe a number of African drum rhythms that stand today as being accurate transcriptions.

In closing this section, it should be mentioned that the effect of percussive music (especially meso-periodic music), can go beyond just being interesting. That is, percussive music in many parts of the world can exert considerable power over listeners, and is often used to induce trance states. This is such a universal phenomenon that it prompted the anthropologist Rodney Needham to discuss it in an article entitled "Percussion and Transition" [Needham, 1967] in which he observes that in almost all rites of passage and other important ceremonies around the world, especially those in which the participants reach a trance state, percussion is the primary source of the music. Numerous other writers have come to the same conclusion; see for example [Diószegi, 1962; Neher, 1962].

Needham and other researchers have no exact explanation, but we might suggest that beyond the specific social/cultural setting, it has to do with:

1. The multiplicity or ambiguity of possible "meters."
2. The physical correlates of the source, e.g. typical broad bandwidth of drums and rattles, sharp attacks, high volume associated with percussion. The broad bandwidth is important for two reasons: physiologically, in that it may result in a wider breadth of neural excitation patterns, and cognitively, in that it obscures pitch, which results in the abstraction of melody and the resultant strengthening of the pure rhythmic impact.

Chapter 3

Methods

> *"Without exception, the best and most practical solutions were provided not by estimation algorithms that have proved their usefulness elsewhere, but by methods evolved by imaginative researchers familiar with speech problems. Thus, in spite of the evolution of more and more powerful statistical estimation algorithms, intimate knowledge of the signal and the idiosyncrasies of its source have been paramount and will continue to be so."*
>
> — from Schroeder, 1970

3.1. Introduction

In Chapter 2, we described various characterizations and theories of rhythm at several levels. In particular, in Section 2.2, an historical review of major attempts at automatic transcription of music was presented. Here we focus on *percussive* music, which has some different problems. For instance, we are very concerned with accurate onset detection, and not as concerned with pitch detection. Because many of the signals are not periodic, we follow a different approach towards segmentation.*

In Section 3.2, the operations on the signal are described. Here the analysis proceeds from segmentation to source detection, which in this case reduces to trying to identify the various strokes used on the drum. This is done first by determining whether a stroke is damped or undamped; an estimation of *tau*, the exponential decay constant, facilitates this decision. Next, the program automatically analyzes the spectrum of the waveform at a location that depends on the damped/undamped decision. By comparing the spectral distribution of the given tone with a data base, the type of stroke can be identified.

---

* Segmentation may be defined as any method of breaking the sample into pieces, which correspond to musical (or in the case of speech, phonetic) events.

At this point, we have the "event-list." From here, it is possible to go in several directions, either relating to resynthesis or to higher analysis. In Section 3.3, we describe the higher-level methods by which we can track tempo, derive meter, and reach a level of representation resembling Western notation.

3.2. Approach to the Low-Level Analysis

The most direct response to percussive music is nominally in terms of energy, or rise-time; therefore we want to detect a sudden increase in amplitude of the signal. This might not seem to be the right approach at first, and indeed there are many difficulties with trying to segment by amplitude. In a sense, amplitude is a "weak" characteristic of musical material, but it is certainly a crucial parameter to consider for any careful study of timing in music. Perceptual onset is probably most closely correlated with amplitude slope [Gordon, 1984].

So, in order to successfully parse percussive music from the acoustic waveform, we need automatic slope detection algorithms. As mentioned in Section 2.2.5, some of the simpler approaches to slope detection, such as 1-point differencing, do not work, because there can be a considerable amount of noise, ringing, reverberation, and substantial overlap between events. For this reason, methods that work well on synthetic data will often fail when applied to real data. In automatic transcription efforts, as in psychoacoustics experiments, synthetic data does not approach the complexity of real data, and therefore can lead to misleading results.

After much experimentation, a slope-detector based on amplitude was implemented that worked in a wide variety of situations. It is based on the *envelope* of the signal, derived as follows.

3.2.1. Envelope Derivation

The derivation of envelope is done by operating in a very intuitive way on the amplitude envelope. The envelope is created by finding the maxima and minima (peaks) of the waveform in each cycle; that is, the max and min in a window that moves through the data. The only difficulty is in setting the size of the window, which depends of the lowest frequency expected in the data. For periodic signals, the best results are obtained when the window is precisely one period long. Since percussive sounds are not usually periodic, it is prudent to pick a window that is sufficiently large; if the lowest frequency is f_0 Hz, then the window must be no smaller than $T_0 = 1/f_0$ seconds. This is because one wants to trace only the peaks, and if the window is smaller than the period of the waveform, one begins to track the intra-period excursions instead of the peaks. If, on the other hand, the window is too large, time resolution is lost. For example, if the lowest frequency is 100 Hz, it is safe to pick a window of at least 10 ms. This envelope represents a data reduction of about $200:1$ from the original sound file.

Figure 3.1 shows what this process looks like when applied to data that corresponds to a single note. If the calculation is carried out over an entire excerpt, it looks like Fig. 3.2, where it is apparent that there is some kind of periodicity, or pattern, but it is not obvious exactly how large it is, or how it is constructed. The answer to this question will unfold as the process of transcription is described.

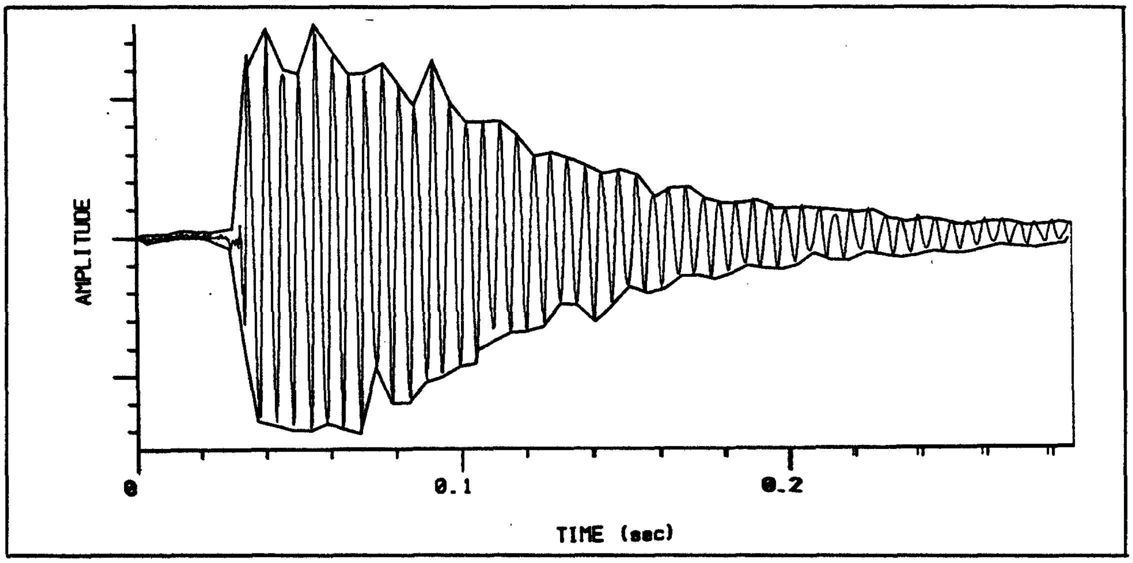


Figure 3.1. Finding the envelope by "connecting the dots" — find the max and min in a moving window, and store as the amplitude envelope. This is a single note, with its envelope superimposed.

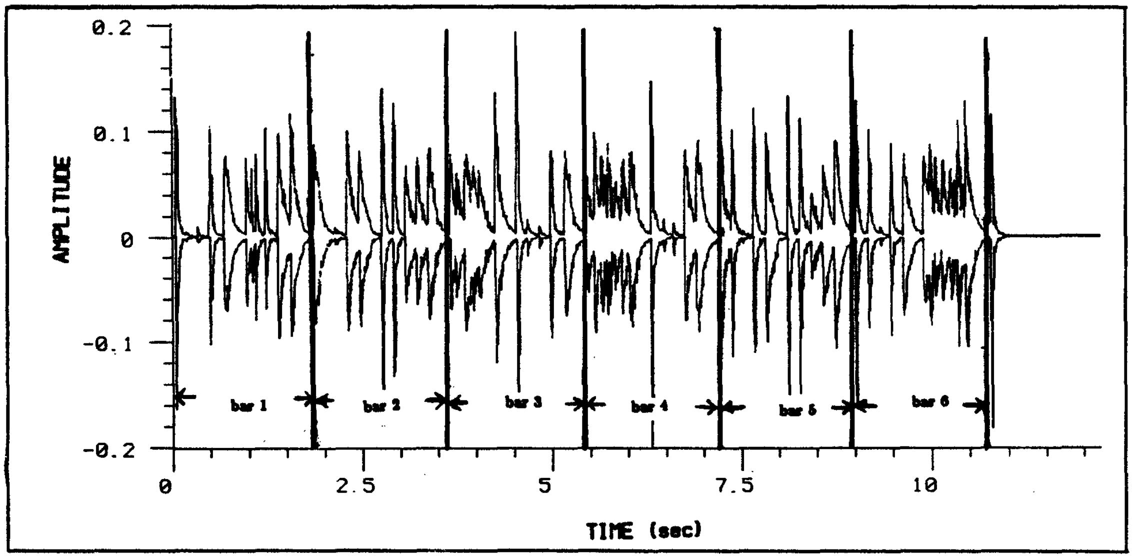


Figure 3.2. The result of finding the envelope as in Fig. 3.1, but done for the entire excerpt. The overall rhythmic periodicity in this example is suggested visually, but it is difficult to describe as yet. The measures are marked here by hand, and a goal is to find them automatically.

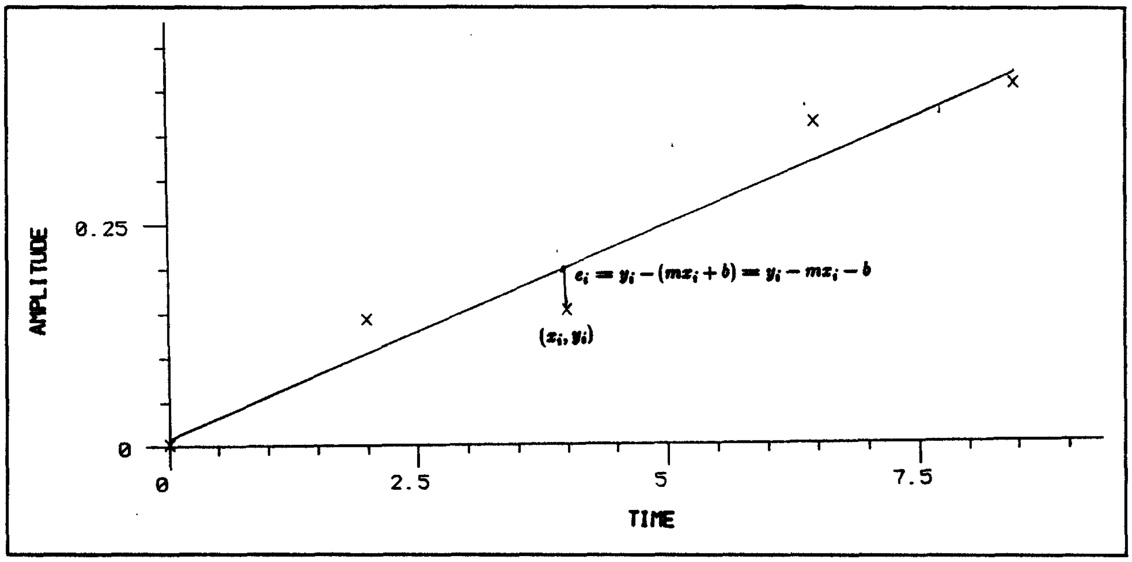


Figure 3.3. This figure and Fig. 3.4 show how the "surfboard" method is derived. A linear regression is performed on a few points of the amplitude envelope at a time, by minimizing the least squares error criterion: $\sum_{i=0}^{n} e_i^2 = \sum_{i=0}^{n} (y_i - mx_i - b)^2$.

3.2.2. Slope Detection

Once the envelope has been created, a novel slope detector is applied to it. The slope detector works by calculating a linear regression over several points of the amplitude envelope using least squares error, as depicted in Fig. 3.3. The regression moves one point at a time through the envelope, approximating n points at a time (n is usually four to eight). This does not create a piecewise linear approximation to the amplitude envelope, but rather, a *sequence* of overlapping line segments that "float" over the data, and are not greatly affected by noise. (See Fig. 3.4). The slope of each line segment is recorded, and segmentation of the material proceeds based the slope array $\{S_n\}$ (defined for each point S_n) and the rules described in the next section.

3.2.3. Segmentation Rules

The following rules are used to segment musical material, based on the slope array $\{S_n\}$. The program proceeds automatically, and has several run-time parameters

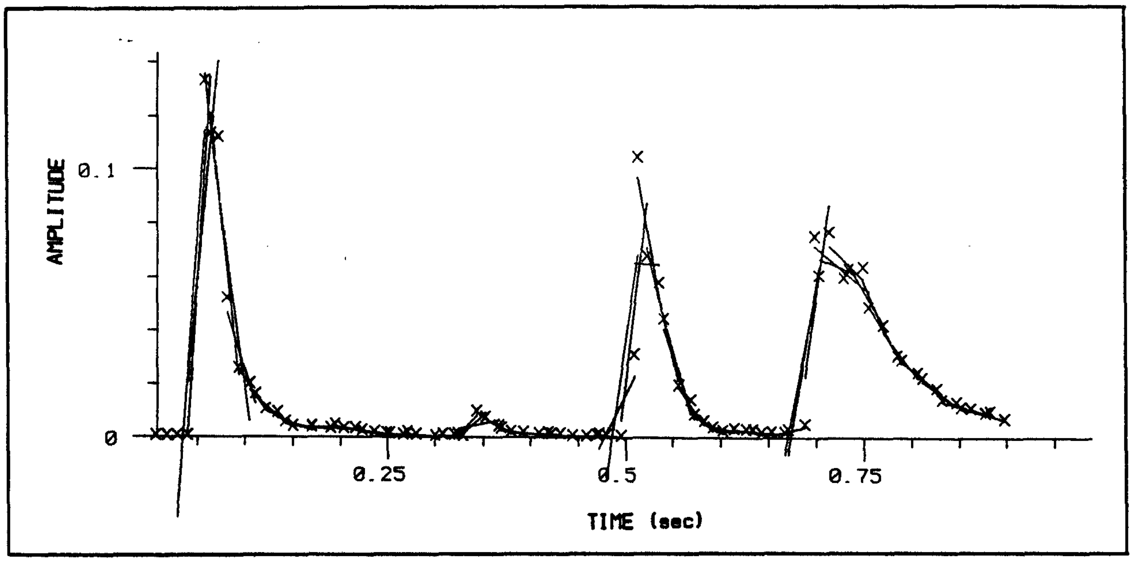


Figure 3.4. The "surfboard" method applied to a real envelope extracted from Fig. 3.2. x's are data points of envelope; line segments are approximating 4 points at a time. Note that the lines are *not* tangent to the curve, but rather approximate groups of overlapping points.

that can be changed; this process is repeated as many times as the user wishes. Typically, the parameters are set on a small segment (excerpt) of the desired musical example, and when the user is satisfied with the results, this procedure is applied to the whole piece. The parameters are described in the next section.

1. Search through slope array $\{S_n\}$ looking for an abrupt increase in slope over a given duration. This is where the first attack is, where the music starts. This is the same method used to measure perceptual attack time, as it seems to yield the best correlation with experimental data [Gordon, 1984].

2. After the attack, assume a "forbidden attack region" for n msec; skip ahead by n msec. There is no new attack allowed for this duration; it is used to avoid detecting spurious attacks that are within n msec. of the previous attack. This parameter is adjustable, as are all the parameters, over each segment.

3. Search for a local maximum in the data after the forbidden region. This local maximum must occur where the slope changes sign, i.e. $S_{n-1} > \epsilon$ and $S_n < -\epsilon$ (mean value theorem). Save the location and the value of the local maximum.

4. After the local maximum, progress through the data, checking to see if the moving average of the power of the signal (over some given duration) obeys the

inequality

$$\text{Moving Average} < .01 \times (\text{Local Maximum})^2.$$

These samples are thus below the noise floor, and we call this a rest, waiting for the next attack. The rest detector is not triggered very often, because in reality, reverberation tends to sustain events at a level above the noise floor. Actually, many rests are not detected that should be, and this issue needs more work. Luckily, what is most important is inter-attack time (see Section 2.3.3.).

5. Search for a local minimum, that is, a point where $S_{n-1} < -\epsilon$ and $S_n > \epsilon$. Record value and location.

6. If (4) or (5) occurs, begin searching for the next attack.

7. For each attack found, send window of data between local max and local min to "Source-detector" that classifies type of stroke, amount of damping, and which drum. (See Section 3.2.7).

3.2.4. Setting Segmentation Parameters

As mentioned above, in order to "tune" the system to a particular set of examples, it is necessary to run it on a segment of data repeatedly, checking intermediate results. These parameters, set during runtime, will affect the segmentation results. The values given are default values (shown below) that have been found to be fairly robust, in that 90% accuracy was achieved in most test examples. The first three depend on the size of a point of the amplitude envelope, usually 2 msec. per point.

Span(4) Onset(2) LengthofDecay(3) Threshold(1) Epsilon(.001) Display(F) or ?:

The interpretation of these parameters is as follows:

- Span — the length of the "surfboard," in this case 4 × 2 = 8 msec. (each point is 2 ms).
- Onset — refers to the "forbidden region" described above; is set to 4 ms.
- LengthofDecay — refers to the size of the window over which the moving average of the power of the signal is taken for rest detection.
- Threshold — The threshold for slope to trigger an attack. A dimensionless quantity.

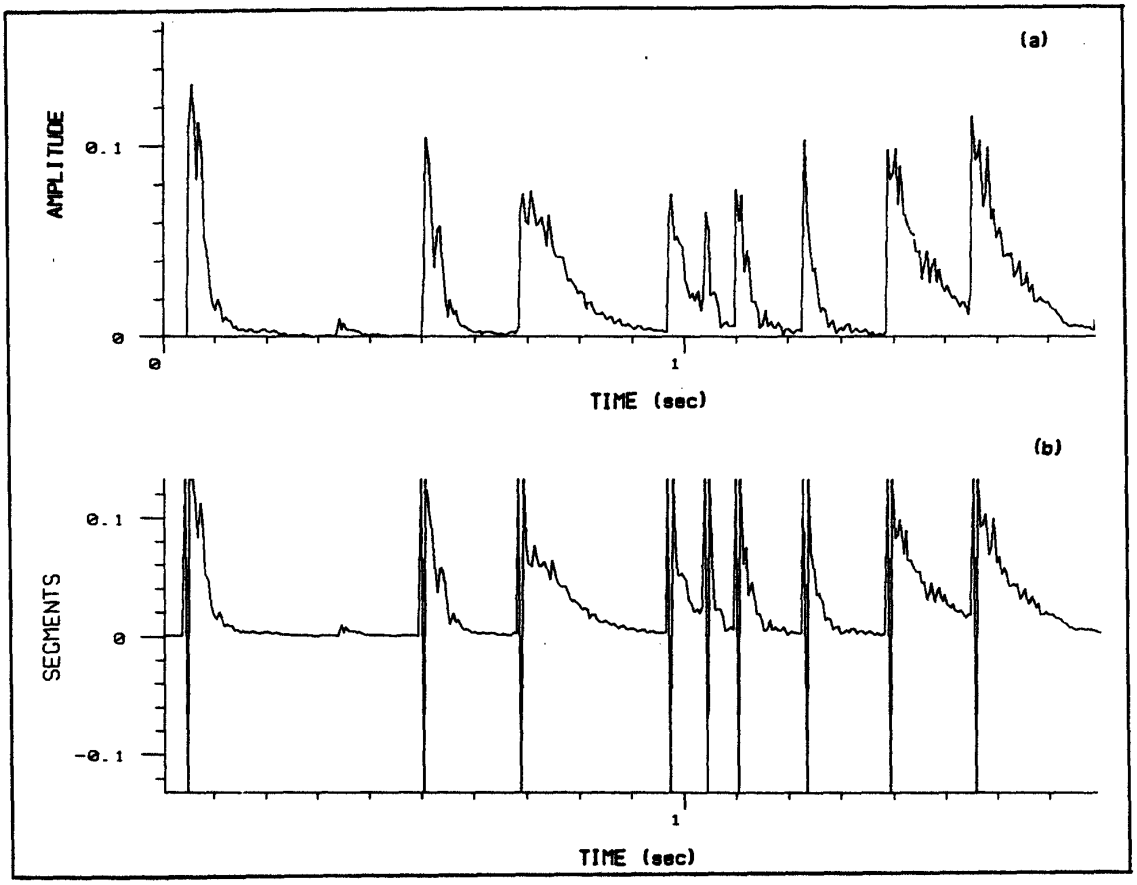


Figure 3.5. An enlargement of "bar 1" from Fig. 3.2, positive values only.

a) The envelope as given.

b) The program's attempt to mark detected attacks.

- Epsilon — a number near 0. Its value is not critical, but it is kept as a parameter for generality.
- Display — Boolean used to display extra information for debugging.

Fig. 3.5 shows a typical example of the interactive segmentation display. It is also interesting to look at the slope array, $\{S_n\}$, which is shown in Fig. 3.6. One can see, in Fig. 3.6, the contours of the different decay types.

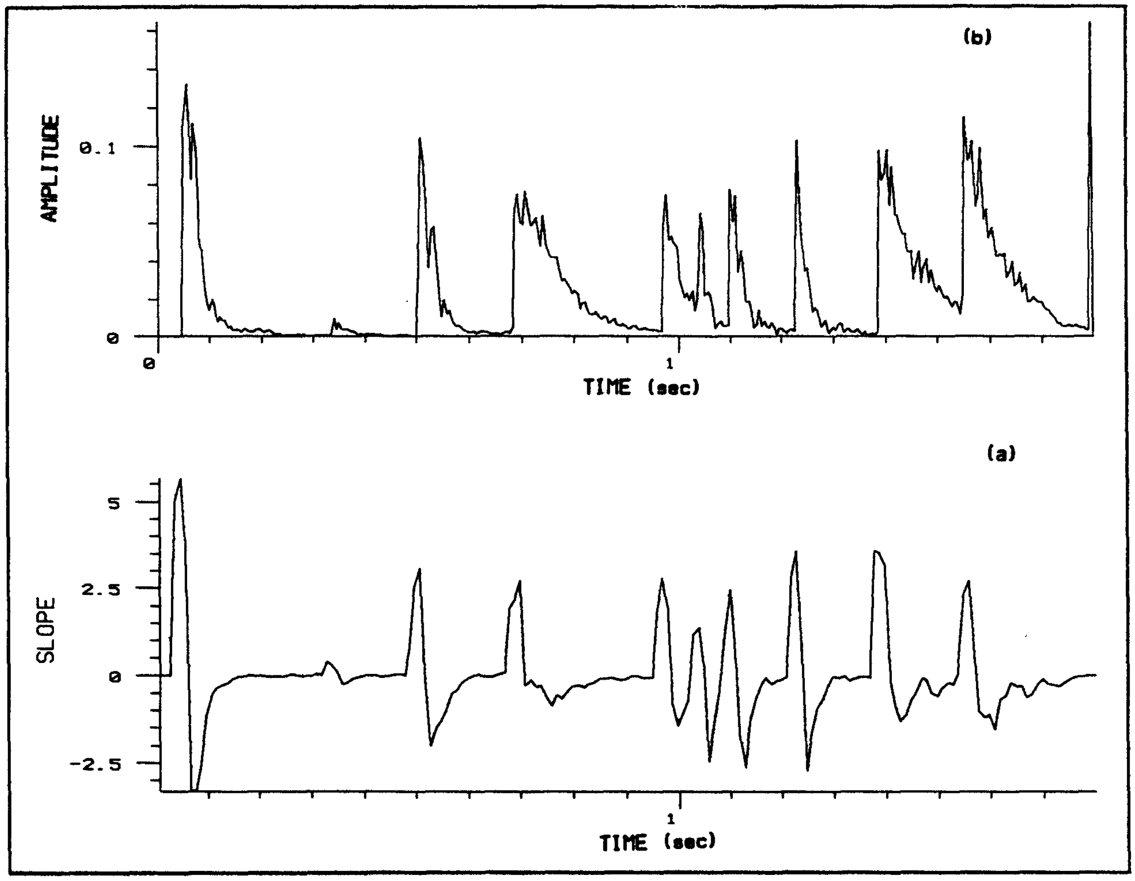


Figure 3.6. The slope array, $\{S_n\}$, for the previous example (Fig. 3.5). Note that the shape of $\{S_n\}$ reflects the behavior of the attack characteristics, and is much less noisy.

a) Original envelope.

b) Slope array $\{S_n\}$.

3.2.5. Perceptual Attack Time

There is a discrepancy between the physical attack time, which is being measured here, and the perceptual attack time (PAT), the instant a listener perceives the attack. It is necessary to correct for this discrepancy, because the system cannot have *a priori* knowledge of the delay. In fact, it is not a serious problem in this analysis for two reasons:

1. The physical attack times are very short (steep), assuring little delay between the physical attack and the percept of the attack. (See Fig. 3.7).

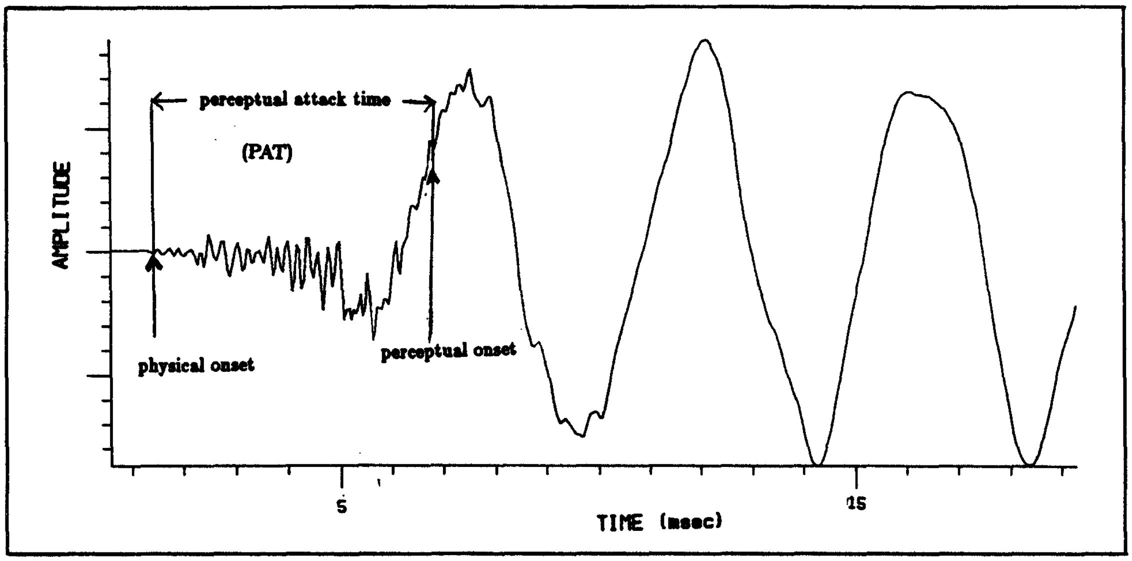


Figure 3.7. Perceptual onset vs. physical onset. Close-up of a typical percussive attack. There is always some delay between the physical onset and perceptual onset. In instruments with very sharp attacks like this one, it will be about 5 msec. The time scale here is 1 ms per tick.

2. Whatever error *is* made tends to be equally made for all attacks from a given source, so the *interattack time* remains fairly stable. In any case, an attempt has been made to deal with this problem as described in Section 2.3.2.

3.2.6. High-Pass Filtering to Facilitate Segmentation

It turns out that the above procedure will fail to detect notes during passages in which adjacent notes are ringing substantially. In Fig. 3.8 (a) we see a fragment of music in which it is impossible to tell by the amplitude how many notes there are. Nor will the previously described methods work; certainly a pitch detector will not detect the seven attacks, nor will the segmenter based on autoregressive modeling (AR), because the AR segmenter cannot detect repeated notes in general, as mentioned in Section 2.2.5.

If we look closely at this excerpt, we see that what is happening is that at each attack there is bit of high frequency noise, but more importantly, there is a *phase discontinuity* at the moment the drum is struck. Figure 3.8 (b) and (c) show this clearly. Intuitively, this corresponds to the reinitialization of the membrane at a random moment when the hand strikes the oscillating drumhead. The cusp in the waveform (sharp corner) causes a wide spread in the frequency domain. Thus, if we high-pass the signal, the attacks become quite obvious, as shown in Fig. 3.9 (b). If we change the cut-off of the filter, the results look as in Fig. 3.10.

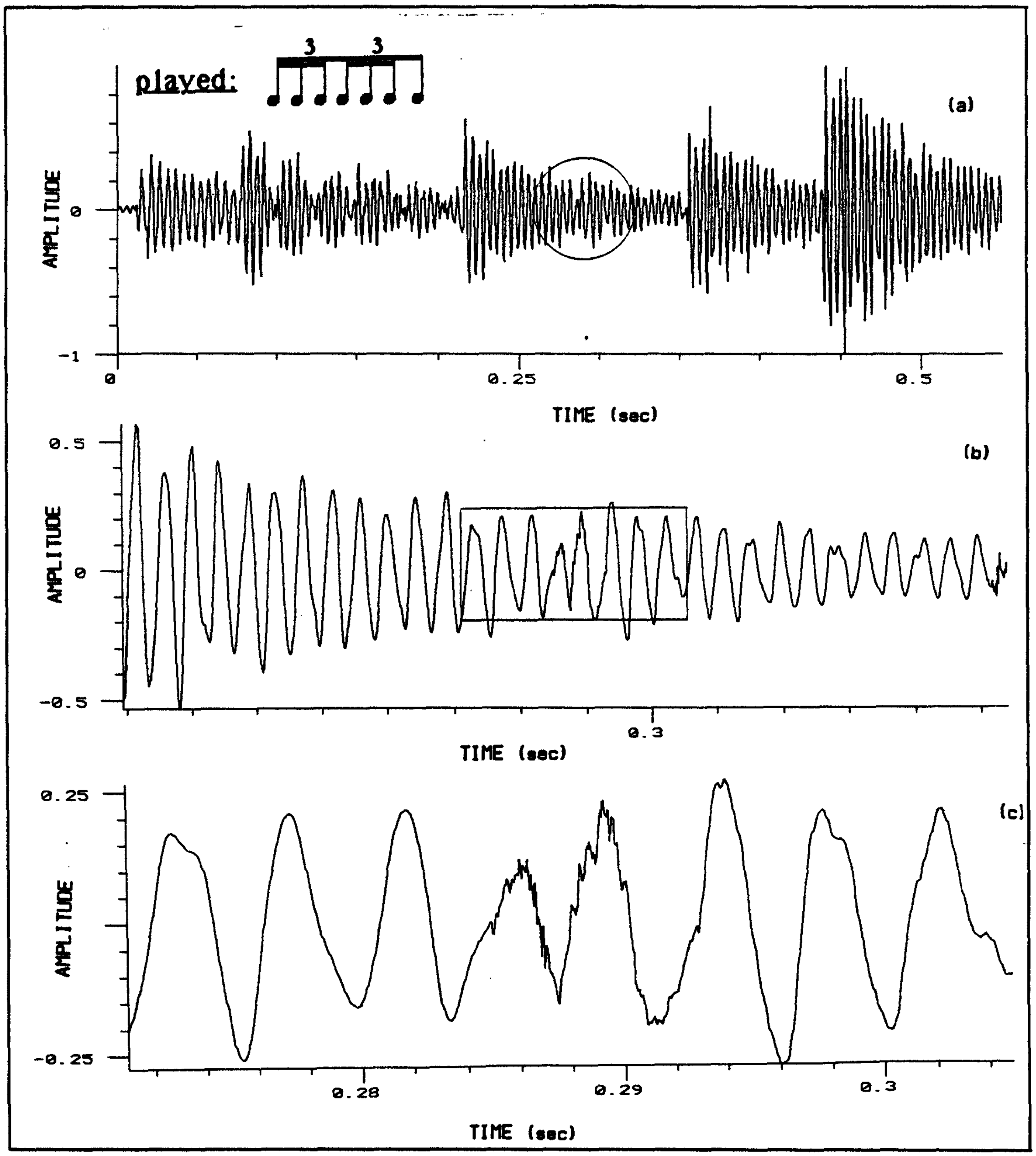


Figure 3.8. A segment that is impossible to parse from the original signal.

a) The original waveform

b) Closeup

c) The key spot. There is high frequency noise at this spot, and also a *phase discontinuity*. The phase discontinuity seems to account for the perception of a new attack more than the high frequency noise.

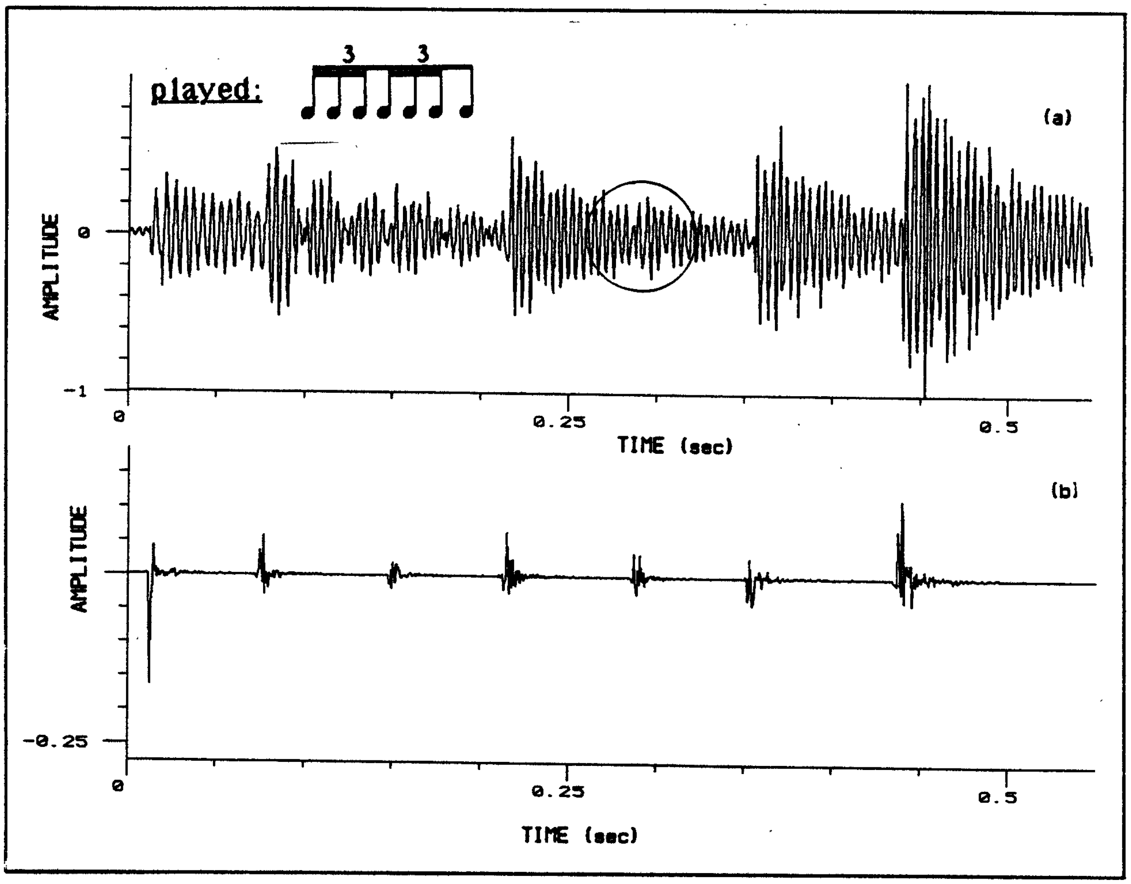


Figure 3.9.

a) The original waveform.

b) High-passed with elliptic filter (6 poles, 6 zeros) at 2.2 kHz. The attacks are easy to detect.

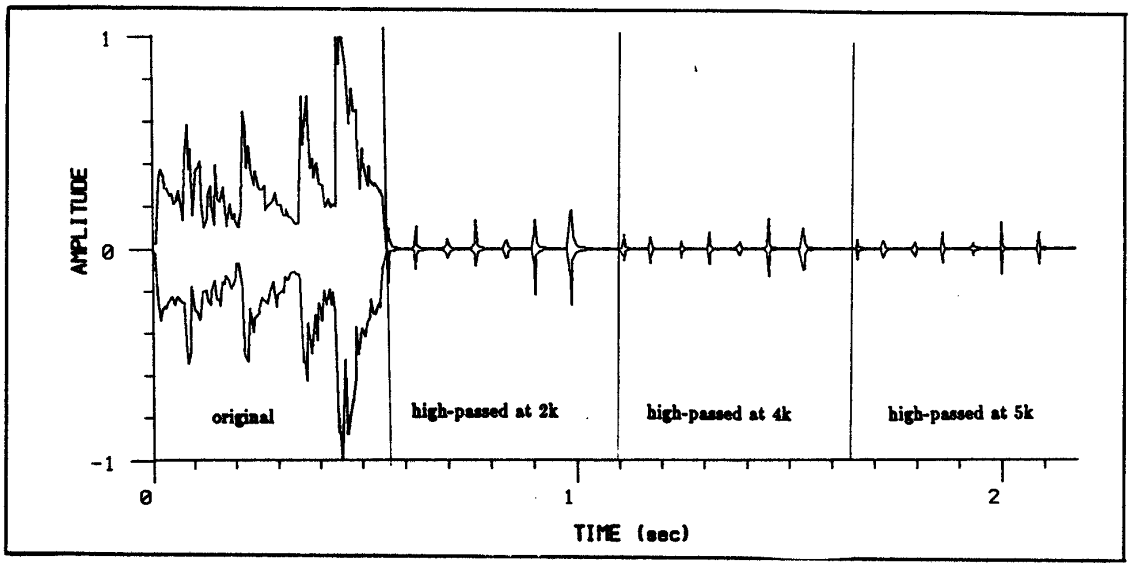


Figure 3.10. The envelope of Fig. 3.8 (a), with three high-pass filters with increasing cut-off frequencies, first 2.2 kHz, then 4 kHz, and then 5 kHz. The sound of this example is surprising; after the original figure played on the drum, the sequence sounds as if it is repeated on a set of three increasingly tiny "soprano maracas," while the rhythm remains absolutely clear.

An interesting perceptual sidelight to this phenomenon is the fact that, if we add only low amplitude noise to the signal at this point, but retain the initial phase relationship, we hear a click at this point, but *not* a new attack.* If, conversely, we add no noise, but instead artificially create a phase discontinuity (for example, 180°) at the same moment, we *do* hear a new attack. Sensitivity to phase in this setting could imply that the auditory system is modeling or tracking the temporal aspects of the waveform, or alternatively that the ear is acting as a finely tuned resonator, responding to the phase shift. Von Békésy thought the latter: "If a resonator is exposed to a tone whose phase is suddenly altered, say by 180°, the amplitude of the resonator falls to 0 and then builds up again." [von Békésy, 1960, p. 412]. He tried to test this hypothesis with a listening experiment, but his experiments with phase reversals were unsuccessful. He was unable to observe the phenomenon, probably because of the imprecise analog equipment available at the time. See Fig. 3.11 for examples of phase reversal vs. added noise, and their perceptual effects in the case of a synthesized sinusiod.

---

* Based on an informal listening experiment with several subjects. See Fig. 3.11.

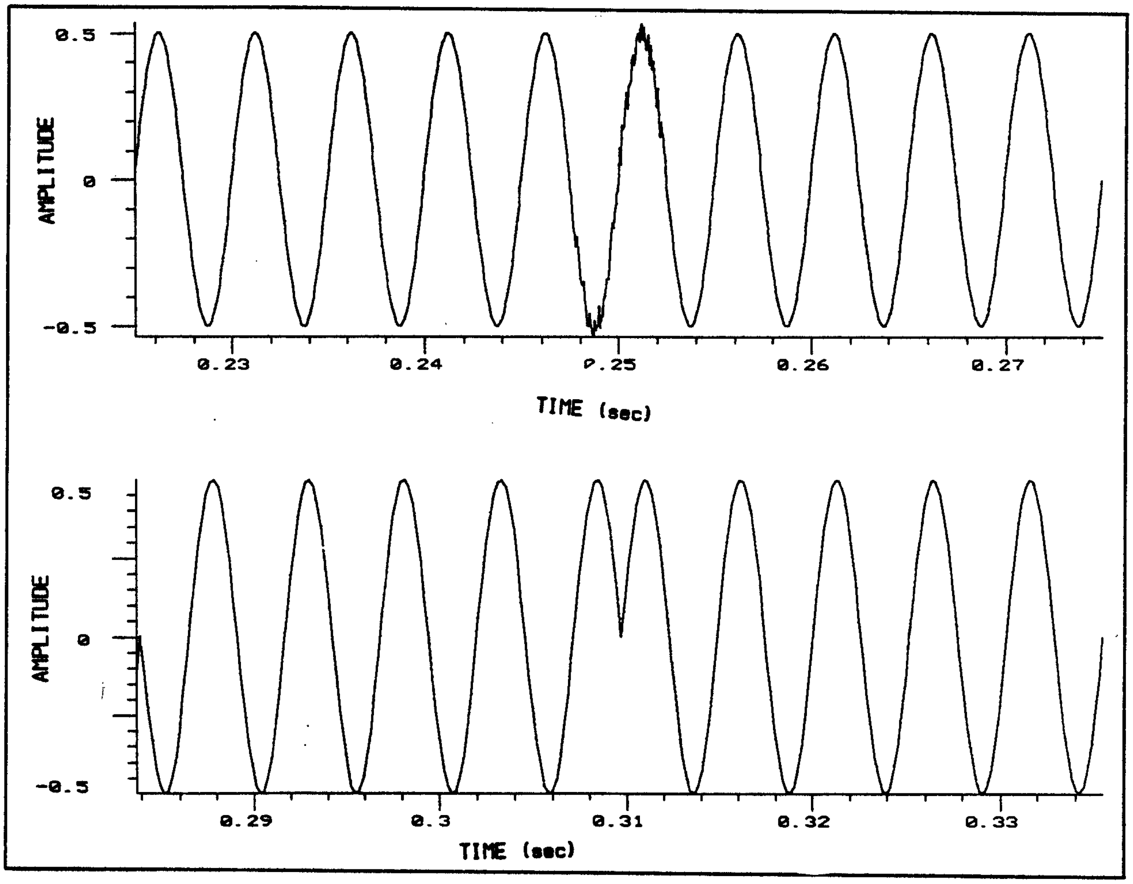


Figure 3.11. Experiment done to test the independent effect of high frequency noise vs. phase discontinuity on percept of a new attack.

a) A 200 Hz sine tone with high frequency, low amplitude noise added for the duration of one period (5 msec). One simply hears a click at the moment of the noise.

b) A 200 Hz sine tone with no added noise, but instead a sudden 180° phase reversal (exactly 1/2 period removed). One hears a definite new attack at the point of the phase reversal.

3.2.7. Source-identification

Now that the data have been segmented, the system tries to discriminate among possible ways the drum was struck. The following methods are generalizable to some extent, at least for other kinds of percussive instruments.

Finding τ

The first objective is to decide whether the stroke is damped or undamped. This decision is critical to the structure of the rhythm. Because of the expected approximate exponential decay for a vibrating membrane, it would be convenient to assign a single value to the decay rate. This is done by determining τ (tau), the exponential decay constant, for each stroke. If we express the envelope y_n as a sampled exponential function with sampling period T,

$$y_n = e^{-nT/\tau},$$

then

$$y_{n+1} = e^{-(n+1)T/\tau} = \left(e^{-nT/\tau}\right)\left(e^{-T/\tau}\right) = y_n e^{-T/\tau}.$$

If we let

$$a = e^{-T/\tau},$$

then

$$y_{n+1} = ay_n = a^2 y_{n-1} = a^3 y_{n-2} = \cdots = a^{n+1} y_0.$$

The coefficient a yields τ in the following way:

$$\tau = \frac{-T}{\ln(a)}, \qquad (T = \textit{sample period}). \tag{3.1}$$

We must find a in order to compute τ. To find a, form the sum

$$\sum_{n=0}^{N-2} y_n y_{n+1} = \sum_{n=0}^{N-2} y_n a y_n = a \sum_{n=0}^{N-2} {y_n}^2,$$

where N is the total number of points in the amplitude envelope.

Solving for a, we have

$$a = \frac{\sum_{n=0}^{N-2} y_n y_{n+1}}{\sum_{n=0}^{N-2} {y_n}^2}.$$

In this derivation for a, what we have really done is to express the autocorrelation method for Linear Prediction as applied to the first-order case, for which it reduces to the above calculation [Markel and Grey, 1976].

Thus, we find tau by first calculating a from the one-pole fit to the envelope, and setting τ from (3.1) above. The rate of decay is thus characterized for the amplitude envelope passed to this routine by the segmenter, from the maximum to the minimum of a given note. A one-pole fit is particularly effective here, because it fits an exponential decay to the curve with minimum error, no matter what portion of the envelope is passed to it—it is not important if the piece of envelope it is given is missing points before, during or after the segment given.* Figure 3.12 shows a typical example.

Once tau is found, a heuristically derived threshold is used to label the event as a damped or undamped stroke. (The difference in decay rate between damped and undamped strokes is usually quite pronounced, so the threshold value is not difficult to set.) This decision is reported to the main process, and the next step of analysis is performed.

Identifying Strokes

Next, a portion of the stroke, (depending on the damped/undamped decision) is sent to the stroke-detector. If the note was determined to be undamped, a window in the middle of the time-waveform (100 – 200 msec.) is analyzed by a pitch detector. The undamped stroke, though its spectrum is changing rapidly, has a nominal "steady state." That is, unlike the various damped strokes, one can identify a pitch for about 500 ms. subsequent to the attack.

---

* Note that τ cannot be easily computed by trying to minimize $\|A(n) - e^{-nT/\tau}\|$ over τ by least squares.

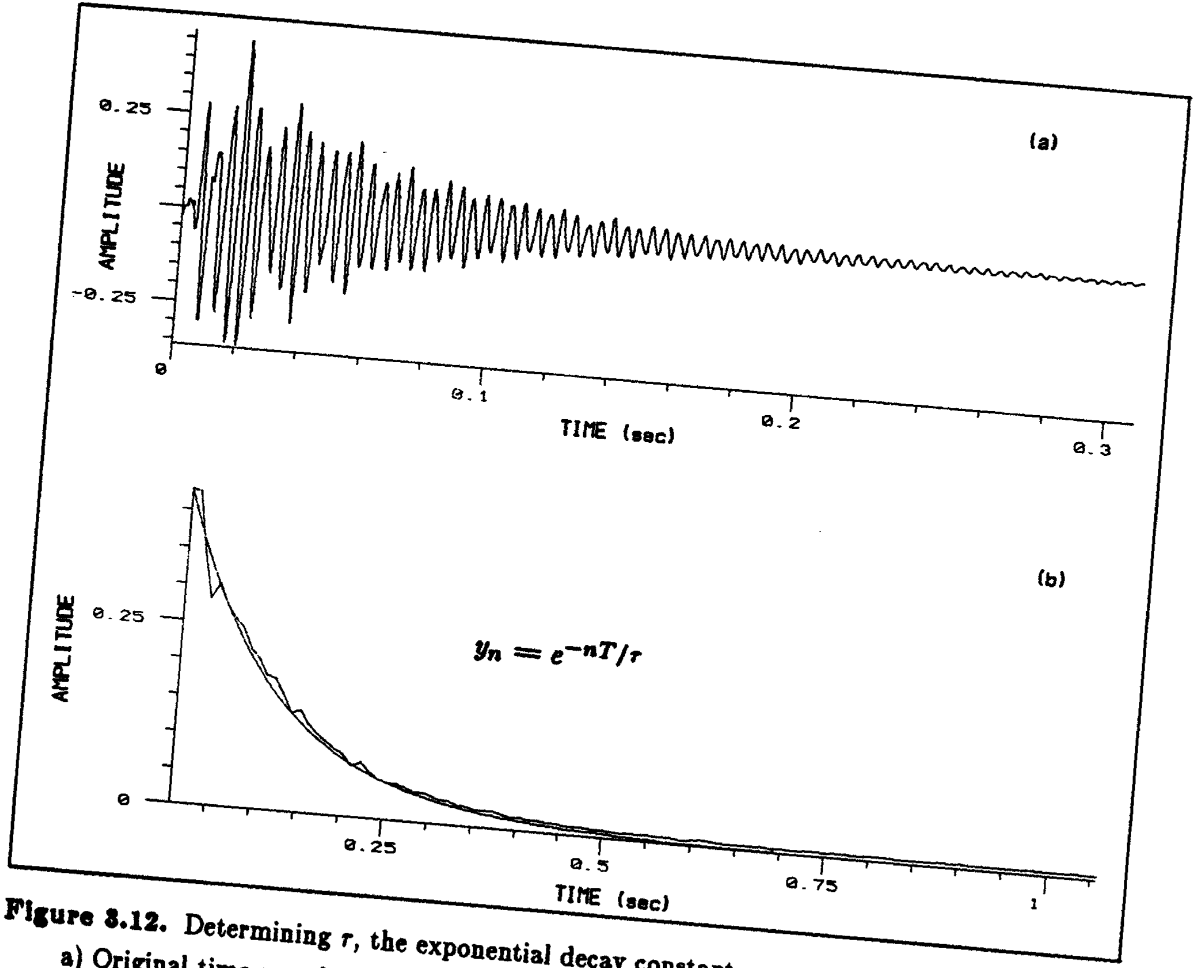


Figure 3.12. Determining τ, the exponential decay constant.

a) Original time waveform

b) Envelope of above waveform, (beginning at the maximum amplitude), with overlay of 1-pole fit to the envelope.

Otherwise, for damped strokes, the stroke-detector is sent the entire duration of the attack ($0 - 100$ ms), and a k-cell partition of the spectrum is created ($k = 3$ in this case), finding the energy in each partition. The number of partitions, k, is programmable, and can be set to any arbitrary number, corresponding to a set of filter banks. The smallest value of k that successfully distinguishes between strokes in the "stroke-space" is used.

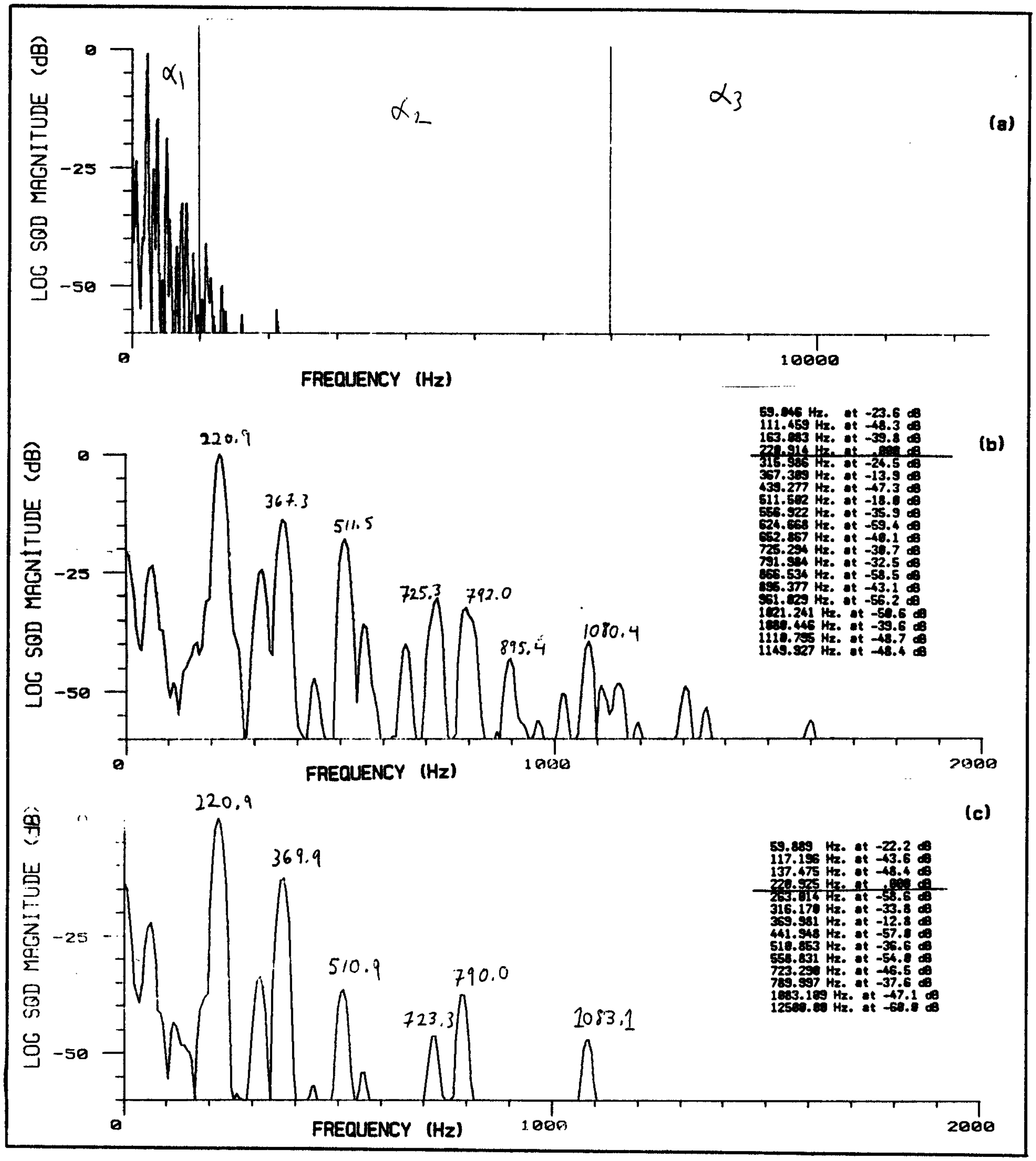
It is important to note that one can actually succeed with a smaller value for k if the partitions are allowed to be asymmetrical, that is, if one picks different sized bins to emphasize the difference between strokes. This can be automated, but at the moment it is set by hand to roughly capture the spectral variety in these particular data. The best method would be to implement a general search that proposes the optimal number of bins (called α_n), and their relative width over a given set of spectra that maximizes distinctions between sources. In Fig. 3.13 — Fig. 3.17 we see the particular spectra of interest in this study. (The time waveforms from which these FFT's were made are shown in Section 2.6.2).

The partitions were selected to facilitate the distinction between the possible strokes, and are set to the following ranges of the spectrum:

$\alpha_1 : [0 - 1\ kHz]$.
$\alpha_2 : [1 - 7\ kHz]$.
$\alpha_3 : [7\ kHz - \text{Nyquist}]$.*

---

* 1/2 the sampling rate, in this case, 12 kHz.



```
59.046 Hz.  at -23.6 dB
111.459 Hz. at -48.3 dB
163.083 Hz. at -39.8 dB
220.914 Hz. at  .000 dB
316.986 Hz. at -24.5 dB
367.389 Hz. at -13.9 dB
439.277 Hz. at -47.3 dB
511.502 Hz. at -18.0 dB
556.922 Hz. at -35.9 dB
624.068 Hz. at -59.4 dB
652.867 Hz. at -40.1 dB
725.294 Hz. at -30.7 dB
791.984 Hz. at -32.5 dB
866.534 Hz. at -58.5 dB
896.377 Hz. at -43.1 dB
961.029 Hz. at -56.2 dB
1021.241 Hz. at -50.6 dB
1080.446 Hz. at -39.6 dB
1110.795 Hz. at -48.7 dB
1149.927 Hz. at -48.4 dB
```

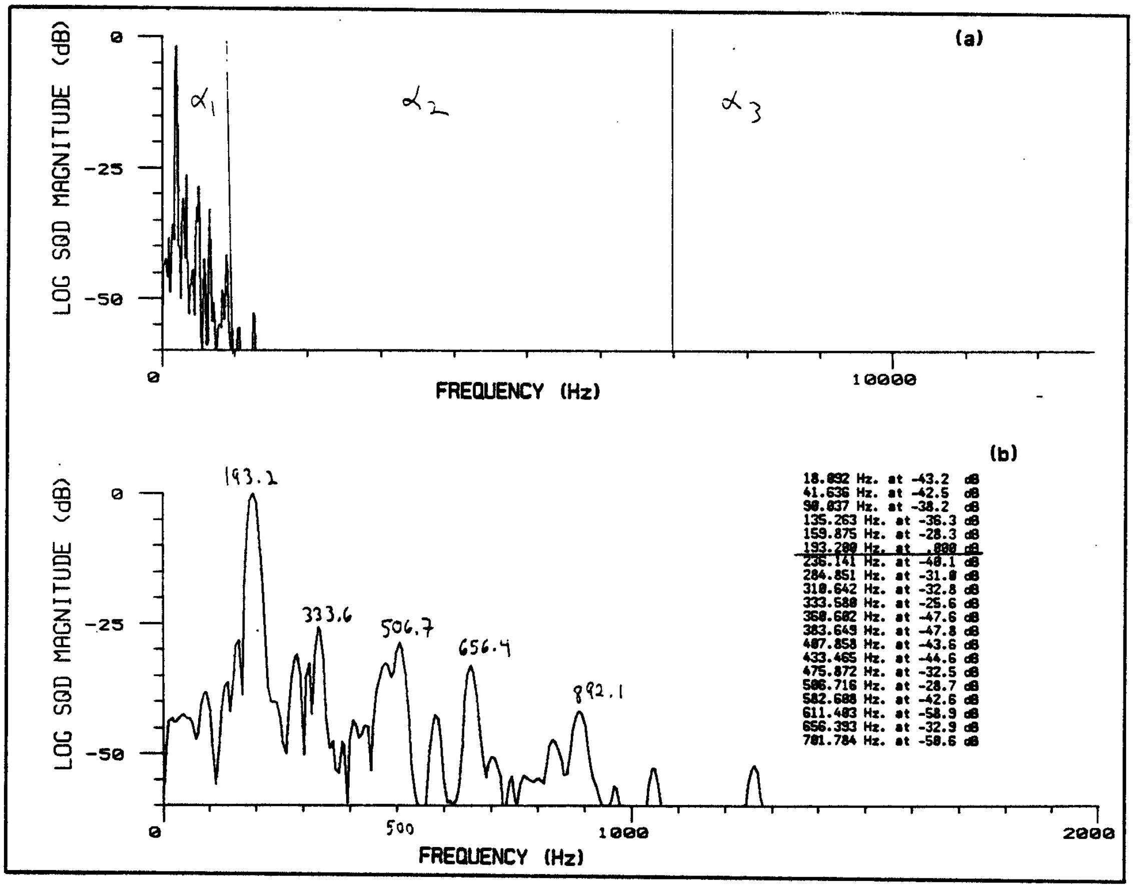
```
59.889  Hz. at -22.2 dB
117.196 Hz. at -43.6 dB
137.475 Hz. at -48.4 dB
220.925 Hz. at  .000 dB
263.014 Hz. at -58.6 dB
316.170 Hz. at -33.8 dB
369.981 Hz. at -12.8 dB
441.948 Hz. at -57.0 dB
510.853 Hz. at -36.6 dB
558.831 Hz. at -54.0 dB
723.290 Hz. at -46.5 dB
789.997 Hz. at -37.6 dB
1083.109 Hz. at -47.1 dB
12500.00 Hz. at -60.0 dB
```

Figure 3.13. FFT of open tone on high drum, (100–200 ms). Normalized to peak; Blackman window (defined as $W[i] = 1 + .19685cos(2\pi i/N) - 1.19685cos(4\pi i/N)$).

a) $0-12$ kHz displayed, showing α regions.

b) Close-up: $0-2$ kHz. Note inharmonic peaks in spectrum.

c) Another example of the same kind of stroke, automatically extracted at random from a larger excerpt. Note that the major peaks are close to those in (b). This is included to verify that indeed the peaks do not vary greatly in different instances, and it is therefore valid to look at their ratios. 100 msec. corresponds to about 20 cycles of the waveform.



```
18.092 Hz. at -43.2  dB
41.636 Hz. at -42.5  dB
90.037 Hz. at -38.2  dB
135.263 Hz. at -36.3 dB
159.875 Hz. at -28.3 dB
193.200 Hz. at  .000 dB
236.141 Hz. at -40.1 dB
284.851 Hz. at -31.0 dB
310.642 Hz. at -32.8 dB
333.580 Hz. at -25.6 dB
360.602 Hz. at -47.6 dB
383.649 Hz. at -47.8 dB
407.858 Hz. at -43.6 dB
433.465 Hz. at -44.6 dB
475.872 Hz. at -32.5 dB
506.716 Hz. at -28.7 dB
582.600 Hz. at -42.6 dB
611.403 Hz. at -58.9 dB
656.393 Hz. at -32.9 dB
701.784 Hz. at -50.6 dB
```

Figure 3.14. FFT of open tone on low drum, (100 – 200 ms). Normalized to peak; Blackman window.

a) 0 – 12 kHz displayed, showing α regions.

b) Close-up: 0 – 2 kHz. Note inharmonic peaks in spectrum.

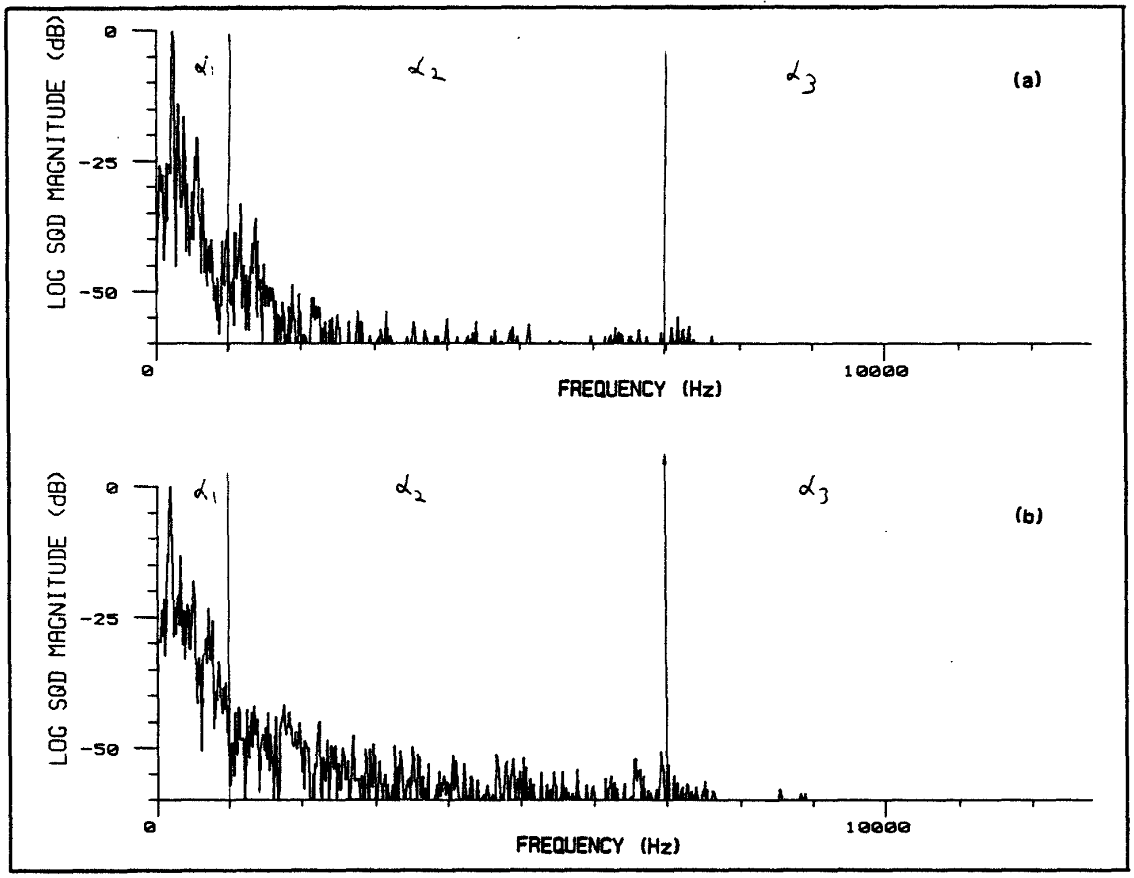


Figure 3.15. FFT of muff tone, (0 – 100 ms). 0 – 12 kHz displayed, showing α regions. Normalized to peak; Blackman window.

a) high drum.

b) low drum.

Given the partition (any partition can be used for any data, but some partitions will work much better than others), the program has a learning capability for dealing with new input. Currently, it expects a set of "canonical" samples of strokes, that are played or otherwise provided by the user as good examples of each kind of stroke. The user provides several examples of each stroke type, along with a name for each.

The program then automatically isolates each stroke based on the segmentation methods described, calculates the energy in each bin, finds the mean and standard deviation of the points, and the spectral peak, and updates its data base with these

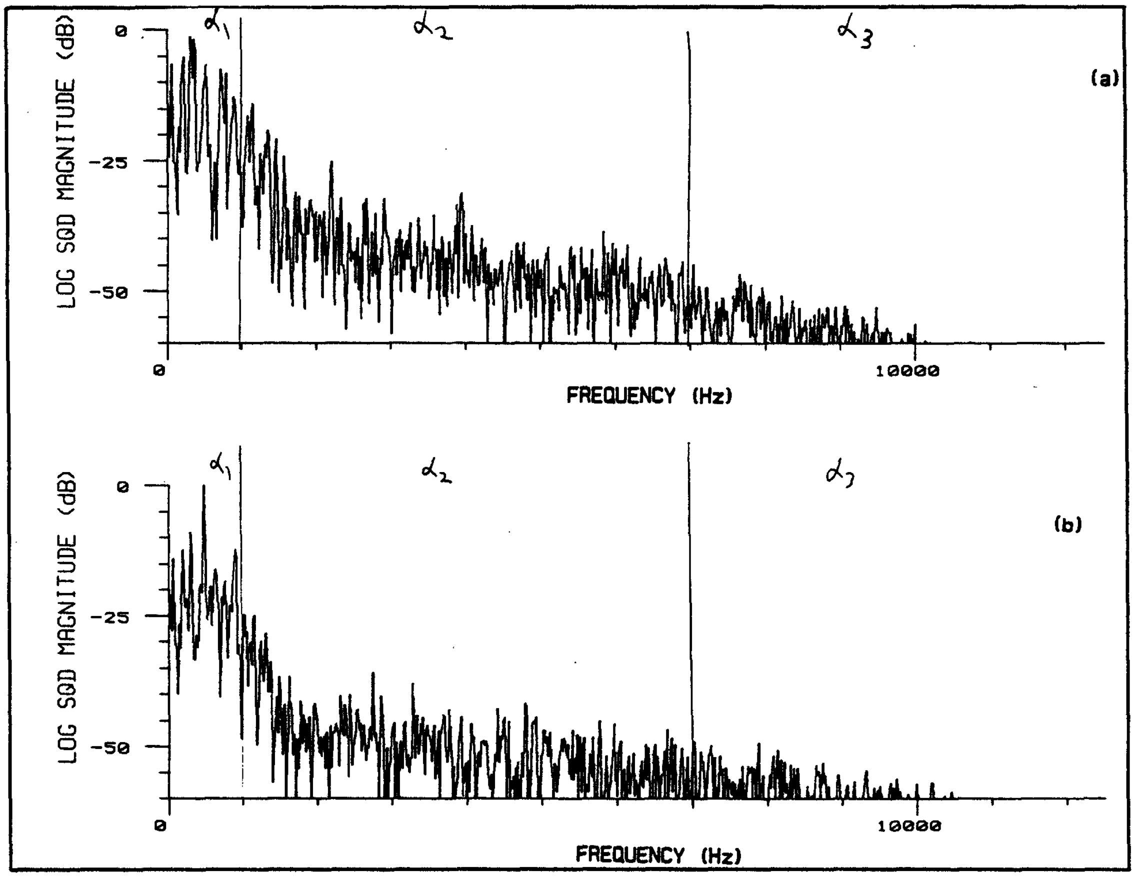


Figure 3.16. FFT of **slap** tone, (0 – 100 ms). 0 – 12 kHz displayed, showing α regions. Normalized to peak; Blackman window.

a) **high** drum.

b) **low** drum.

values, as shown in Table 3.1. In this way, it is possible to "train" the program to deal with new data played on different drums or played by a different performer, because the database will be updated to reflect the new types of strokes.

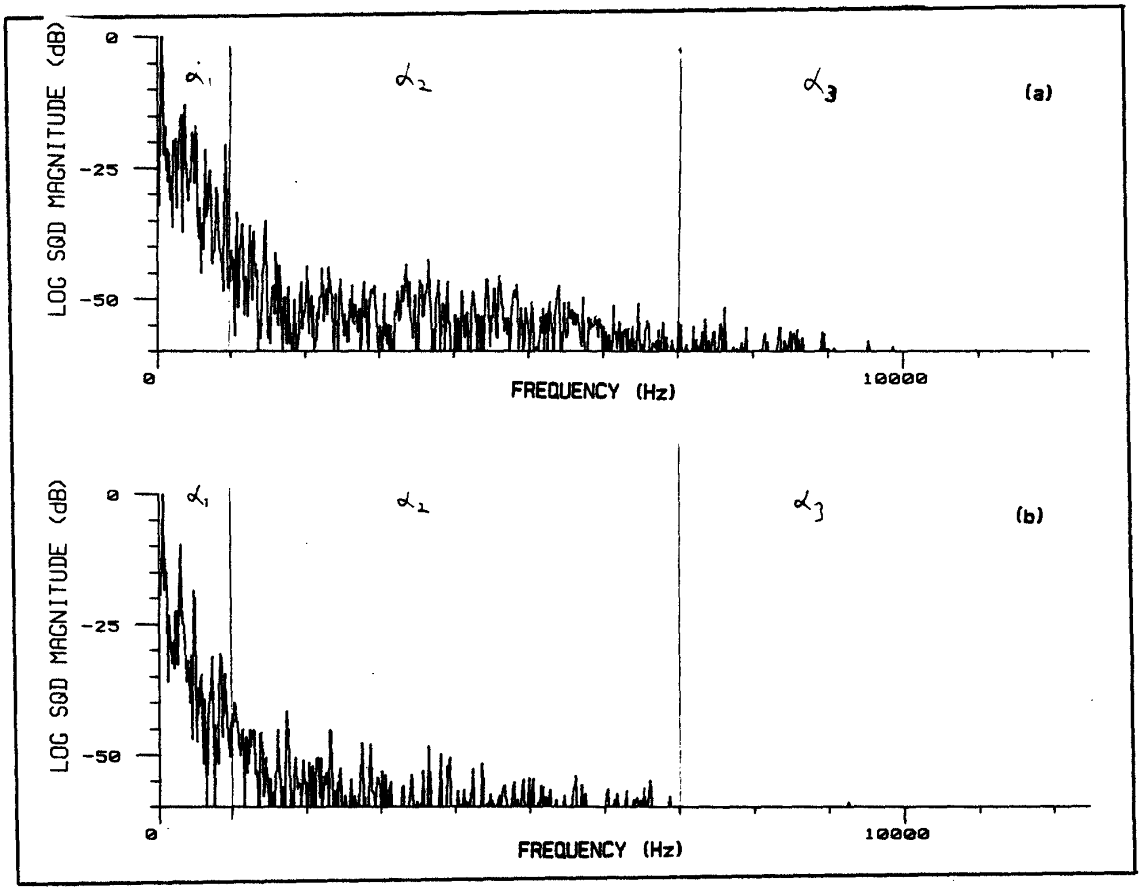


Figure 3.17. FFT of **bass** tone, (0 – 100 ms). 0 – 12 kHz displayed, showing α regions. Normalized to peak; Blackman window.

a) **high drum.**

b) **low drum.**

Data Base for Stroke Search						
STROKE	α_1	α_2	α_3	SD	PEAK	No. Ex.
$H-OPEN$	1.0000	0.0001	0.0000	0.0541	220.0	2.0
$L-OPEN$	1.0000	0.0001	0.0000	0.0541	185.0	2.0
$H-MUFF$	0.6560	0.3400	0.0077	0.0784	240.0	10.0
$L-MUFF$	0.5390	0.4370	0.0251	0.1110	195.0	9.0
$H-SLAP$	0.3630	0.6100	0.0287	0.0377	415.0	9.0
$L-SLAP$	0.3269	0.6217	0.0536	0.4365	479.8	8.0
$H-BASS$	0.3706	0.5839	0.0471	0.0276	52.26	7.0
$L-BASS$	0.4666	0.5237	0.0122	0.0458	53.21	6.0

Table 3.1. Data that are compiled automatically by the program when the user provides sample "canonical" input. These values are compared with unknown values as explained in Fig. 3.18. Note that, for example, for the open strokes on both high and low drums, virtually all the energy is in the first frequency bin (α_1). α's are the energy in each frequency bin, SD is the standard deviation of the α's, PEAK is the average peak of the spectrum for each stroke, and No. Ex. is the number of sample strokes of each type. The open tones tend to be fairly consistent, so only two examples were used to set data points for these.

When then confronted with a real example, the system computes the *normalized* distance from a given 3-tuple to canonical values in the data base, and picks the minimum value that corresponds to one of the strokes. Normalized in this context means that the distance to the various stored values is divided by the SD (standard deviation) for each value. This has the effect of "partitioning" the spectral space in the following way: if a particular stroke is not played uniformly, it will occupy a large region of the spectral space, and therefore the probability of an unknown stroke being in that category is higher than it would be in the case of another stroke that is the same distance away, but with a smaller SD. The small value for the SD would imply that the latter stroke is more constrained, and is played more uniformly (see Fig. 3.18).

At this point, a confidence measure ($0 < m \leq 1$) is reported for each stroke, according to how well the various aspects of the process "agree" by a simple weighting scheme, based on:

1. Closeness of fit of spectrum to one of the values in the database.
2. Closeness of fit of spectral peak to corresponding peak in database.
3. Agreement between damped/undamped decision based on tau vs. that

implied by spectral shape.

The stroke names originally given by the user are retained in the note field as tokens, which are used by the higher level for scoring and pattern detection, and also directly for resynthesis of various kinds.

An example of the auxillary output for debugging purposes is shown in Fig. 3.19. Note that this whole page can be condensed to a single line of Fig. 3.20, which is the output of this part of the program, and the input to the next section (§ 3.3). It can be called a "notelist," as it represents the actual performance; it is an unabstracted representation of the music.

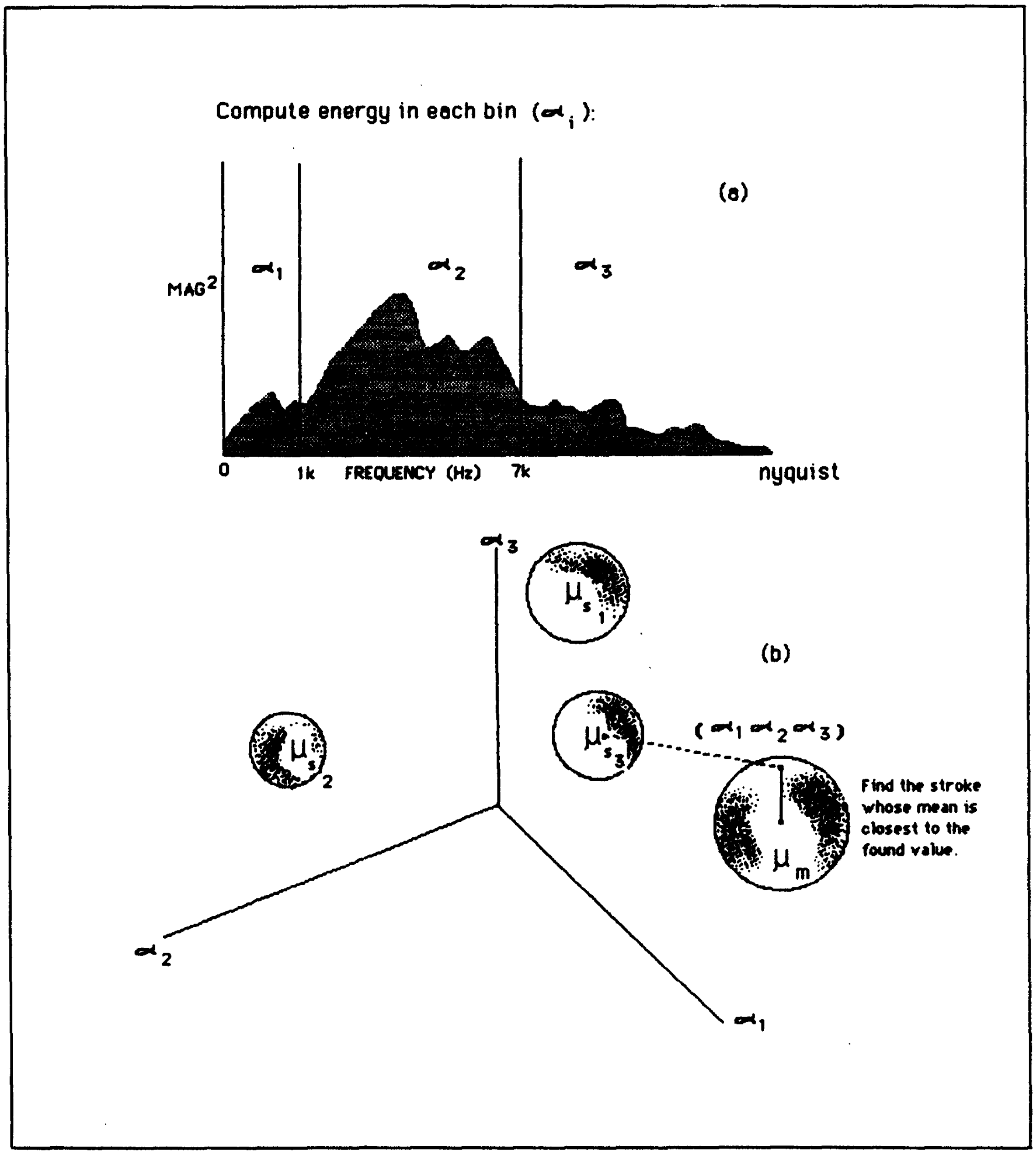
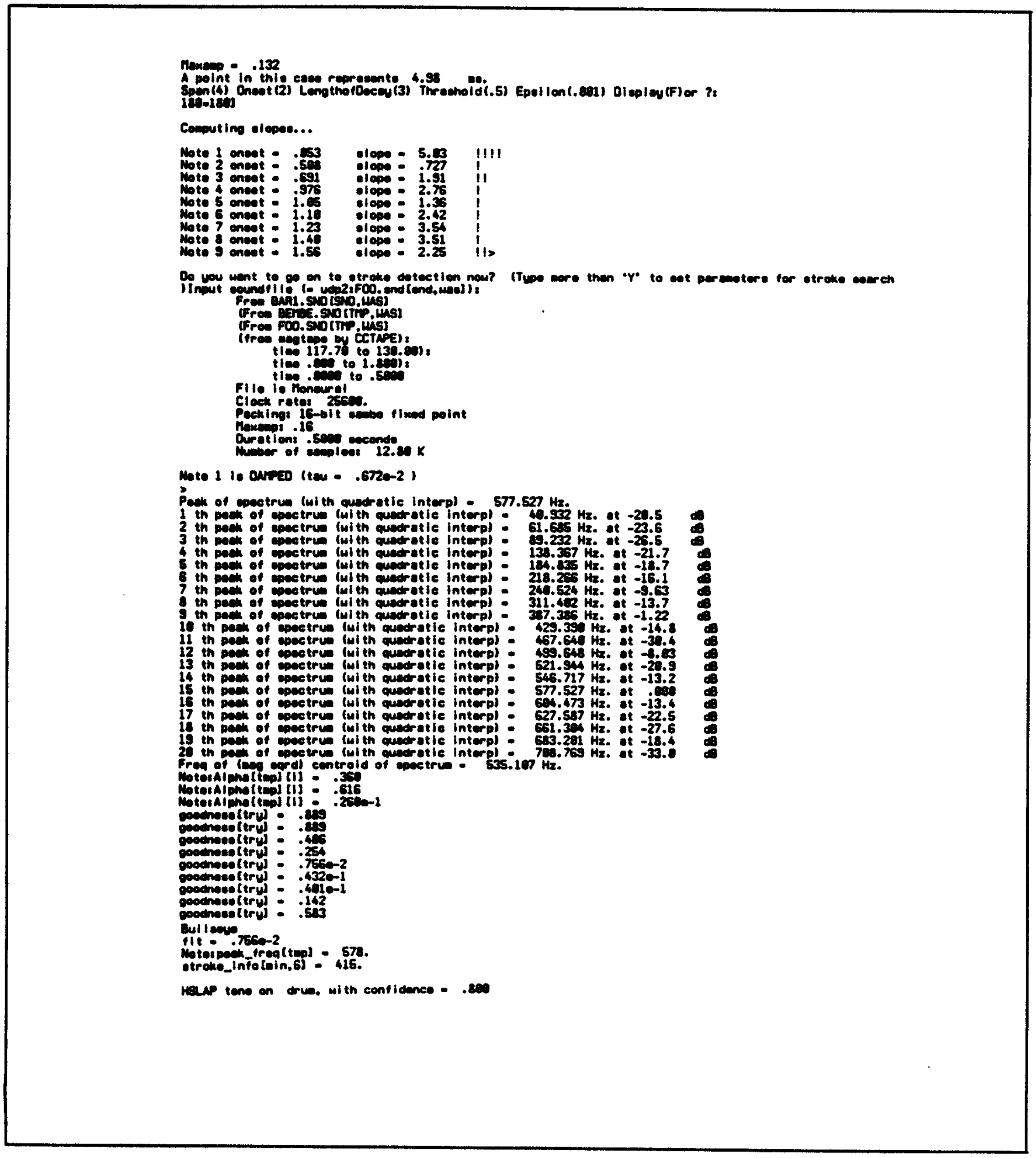

Compute energy in each bin (α_i):
(a)
MAG^2
α_1
α_2
α_3
0
1k
FREQUENCY (Hz)
7k
nyquist
α_3
μ_{s_1}
(b)
μ_{s_2}
μ_{s_3}
(α_1 α_2 α_3)
μ_m
Find the stroke whose mean is closest to the found value.
α_2
α_1

Figure 3.18. The scheme for source detection.

a) The spectra of all strokes is partitioned as shown, to maximize the difference between strokes.

b) The energy in each bin is computed, and compared with values that are stored in Table 3.1. For each unknown stroke, the token, or stroke, that corresponds to the minimum value of $\frac{\|\vec{a}-\vec{\mu}_k\|}{SD_k}$, is retained.



```
Maxamp =  .132
A point in this case represents  4.98    ms.
Span(4) Onset(2) LengthofDecay(3) Threshold(.5) Epsilon(.001) Display(F)or ?:
100-1001

Computing slopes...

Note 1 onset =  .053     slope =  5.83    !!!!
Note 2 onset =  .508     slope =  .727    !
Note 3 onset =  .691     slope =  1.91    !!
Note 4 onset =  .976     slope =  2.76    !
Note 5 onset =  1.05     slope =  1.36    !
Note 6 onset =  1.10     slope =  2.42    !
Note 7 onset =  1.23     slope =  3.54    !
Note 8 onset =  1.40     slope =  3.51    !
Note 9 onset =  1.56     slope =  2.25    !!>

Do you want to go on to stroke detection now?  (Type more than 'Y' to set parameters for stroke search
)Input soundfile (= udp2:FOO.snd[snd,was]):
        From BAR1.SND[SND,WAS]
         (From BEMBE.SND[TMP,WAS]
         (From FOO.SND[TMP,WAS]
         (from magtape by CCTAPE):
              time 117.70 to 130.00):
              time .000 to 1.000):
              time .0000 to .5000
        File is Monaural
        Clock rate:  25600.
        Packing: 16-bit sambo fixed point
        Maxamp: .16
        Duration: .5000 seconds
        Number of samples:  12.80 K

Note 1 is DAMPED (tau =  .672e-2 )
>
Peak of spectrum (with quadratic interp) =   577.527 Hz.
1 th peak of spectrum (with quadratic interp) =   40.932 Hz. at -20.5    dB
2 th peak of spectrum (with quadratic interp) =   61.685 Hz. at -23.6    dB
3 th peak of spectrum (with quadratic interp) =   89.232 Hz. at -26.5    dB
4 th peak of spectrum (with quadratic interp) =   138.367 Hz. at -21.7    dB
5 th peak of spectrum (with quadratic interp) =   184.835 Hz. at -18.7    dB
6 th peak of spectrum (with quadratic interp) =   218.266 Hz. at -16.1    dB
7 th peak of spectrum (with quadratic interp) =   240.524 Hz. at -9.63    dB
8 th peak of spectrum (with quadratic interp) =   311.402 Hz. at -13.7    dB
9 th peak of spectrum (with quadratic interp) =   307.386 Hz. at -1.22    dB
10 th peak of spectrum (with quadratic interp) =   429.390 Hz. at -14.8    dB
11 th peak of spectrum (with quadratic interp) =   467.640 Hz. at -30.4    dB
12 th peak of spectrum (with quadratic interp) =   499.648 Hz. at -8.03    dB
13 th peak of spectrum (with quadratic interp) =   521.944 Hz. at -20.9    dB
14 th peak of spectrum (with quadratic interp) =   546.717 Hz. at -13.2    dB
15 th peak of spectrum (with quadratic interp) =   577.527 Hz. at  .000    dB
16 th peak of spectrum (with quadratic interp) =   604.473 Hz. at -13.4    dB
17 th peak of spectrum (with quadratic interp) =   627.587 Hz. at -22.5    dB
18 th peak of spectrum (with quadratic interp) =   661.304 Hz. at -27.6    dB
19 th peak of spectrum (with quadratic interp) =   683.201 Hz. at -18.4    dB
20 th peak of spectrum (with quadratic interp) =   708.769 Hz. at -33.0    dB
Freq of (mag sqrd) centroid of spectrum =   535.107 Hz.
Note:Alpha[tmp][i] =  .360
Note:Alpha[tmp][i] =  .616
Note:Alpha[tmp][i] =  .260e-1
goodness[try] =  .889
goodness[try] =  .889
goodness[try] =  .406
goodness[try] =  .254
goodness[try] =  .756e-2
goodness[try] =  .432e-1
goodness[try] =  .401e-1
goodness[try] =  .142
goodness[try] =  .583
Bullseye
fit =  .756e-2
Note:peak_freq[tmp] =  578.
stroke_info[min,6] =  416.

HSLAP tone on  drum, with confidence =  .800
```

Figure 3.19. This is a log of a session, that includes all changes, or repeated runs with different parameters. Note the confidence measure, which is a weighting value for how well various aspects of the analysis "add up." This whole page boils down to one line (the first note) in the notelist, shown in Fig. 3.20.

```
PLAY;
PARS IGN BEG DUR FRQ AMP INS;
∂ bar1;
pluck, .053, .455, 577.908, .1, HSLAP;
pluck, .508, .183, 577.969, .1, HBASS;
pluck, .691, .285, 194.726, .1, LOPEN;

pluck, .976, .071, 227.456, .1, HBASS;
pluck, 1.047, .057, 241.303, .1, LSLAP;
pluck, 1.104, .129, 241.936, .1, LMUFF;
pluck, 1.234, .163, 309.188, .1, LBASS;
pluck, 1.396, .166, 228.183, .1, HOPEN;
pluck, 1.562, .308, 228.214, .1, HOPEN;

∂ bar2;
pluck, 1.870, .434, 194.666, .1, LOPEN;
pluck, 2.305, .165, 227.969, .1, HOPEN;
pluck, 2.469, .299, 227.725, .1, HOPEN;

pluck, 2.769, .148, 388.498, .1, LSLAP;
pluck, 2.917, .165, 578.725, .1, LSLAP;
pluck, 3.082, .147, 228.013, .1, HOPEN;
pluck, 3.229, .155, 227.823, .1, HOPEN;
pluck, 3.384, .293, 194.735, .1, LOPEN;

∂ bar3;
pluck, 3.678, .080, 228.127, .1, HMUFF;
pluck, 3.757, .107, 227.851, .1, HMUFF;
pluck, 3.864, .186, 194.798, .1, LOPEN;
∂ missing note;
pluck, 4.050, .213, 194.716, .1, LOPEN;
∂ missing note;
pluck, 4.263, .283, 227.992, .1, HOPEN;

pluck, 4.546, .444, 576.958, .1, LSLAP;
pluck, 4.990, .183, 576.634, .1, HSLAP;
pluck, 5.174, .287, 194.615, .1, LOPEN;

∂ bar4;
pluck, 5.461, .097, 227.994, .1, HMUFF;
pluck, 5.558, .186, 227.934, .1, HOPEN;
∂ missing note;
pluck, 5.743, .102, 228.217, .1, HOPEN;
pluck, 5.845, .086, 227.949, .1, HOPEN;
pluck, 5.931, .098, 227.951, .1, HMUFF;
pluck, 6.028, .281, 194.591, .1, LOPEN;

pluck, 6.310, .450, 386.001, .1, LSLAP;
pluck, 6.759, .157, 227.857, .1, HOPEN;
pluck, 6.917, .310, 134.725, .1, LOPEN;

∂ bar5;
pluck, 7.227, .151, 228.248, .1, HMUFF;
pluck, 7.378, .288, 390.810, .1, HSLAP;
pluck, 7.665, .166, 536.222, .1, LSLAP;
pluck, 7.832, .292, 227.655, .1, HOPEN;

pluck, 8.124, .155, 386.074, .1, LSLAP;
pluck, 8.279, .155, 391.038, .1, HSLAP;
pluck, 8.434, .151, 194.188, .1, HMUFF;
pluck, 8.584, .146, 194.568, .1, LOPEN;
pluck, 8.730, .289, 195.002, .1, LOPEN;

∂ bar6;
pluck, 9.019, .170, 386.168, .1, HSLAP;
pluck, 9.189, .290, 576.928, .1, LSLAP;
pluck, 9.479, .161, 581.006, .1, LSLAP;
pluck, 9.640, .263, 227.682, .1, HOPEN;

pluck, 9.903, .073, 195.731, .1, HMUFF;
pluck, 9.976, .078, 227.601, .1, HMUFF;
pluck, 10.053, .089, 227.855, .1, HOPEN;
pluck, 10.142, .102, 227.742, .1, HOPEN;
pluck, 10.244, .095, 194.519, .1, LOPEN;
pluck, 10.339, .116, 227.837, .1, HOPEN;
pluck, 10.455, .332, 227.733, .1, HOPEN;

∂ bar7;
pluck, 10.787, .620, 578.062, .1, LSLAP;
FINISH;
```

Figure 3.20. This level of analysis, which is the output of all analysis carried on so far, is the input to resynthesis routines and high-level analysis described in the next section (§ 3.3). Its fields include BEG (begin time) DUR (duration—inter-attack time), AMP (amplitude, currently ignored), and STROKE (the token that describes the stroke-type). In this example, there are three missing (undetected) notes in a total of fifty-five, a success rate of about 95%. (Refer to this figure, called the *notelist*, throughout Section 3.3.)

3.2.8. On Pitch Detection

Drums present an interesting problem in pitch perception: as vibrating membranes, they have intrinsically inharmonic spectra, yet many drums elicit a relatively clear sense of pitch. An ideal circular membrane will have modes that are in the relation 1 : 1.59 : 2.14 : 2.30 : 2.65 : 2.92, which is anything but harmonic [Morse and Ingard, 1968]. As any percussionist knows, membranophones span a large space of "pitch clarity," ranging from large bass drums that elicit almost no sense of pitch, to the North Indian *tablá*, with a very clear pitch. Two-headed drums generally have the weakest pitch clarity, due to interference between the two membranes, and the coupled air mass joining them. Single-headed drums are in the middle; for instance, orchestral timpani and the conga drum under study here are fairly well-behaved in terms of perceived pitch. (There is a second-order effect in these drums in which modes are slightly altered due to acoustic coupling with the shell; both of these drums are topologically hemispheres rather than cylinders.) The clearest sense of pitch is elicited by those drums, like the *tablá*, Burmese *pa waing*, and Cuban *batá* that utilize a tuning paste in the center of the membrane whose mass tends to damp out most of the inharmonic modes.

Of the numerous mechanisms by which the modes of a vibrating membrane are brought closer to integral relationships, the most important is probably the effect of air-mass loading; the mass of air that oscillates with the membrane causes the frequencies of the modes to be lowered significantly. In the case of the timpani, this causes the modes to be close to the relation 2 : 3 : 4 : 5 : 6 [Rossing, 1982]. In the case of the conga drum, the relationships are instead (assuming the highest peak to be the fundamental): 1 : 1.66 : 2.32 : 3.28 : 4.90 for the high drum ($f_1 = 220.9$ Hz), and 1 : 1.73 : 2.62 : 3.40 : 4.61 for the low drum ($f_1 = 193.2$ Hz).*

It seems, from listening to the drum and identifying the pitch informally, that the sense of pitch for the conga drum (and probably other drums) is overwhelmingly due to the "fundamental" in the undamped strokes; the higher partials are considerably lower in amplitude, as shown in Fig. 3.13 and Fig. 3.14. For the purposes of transcription, it suffices to note the peak of the spectrum in the case of undamped strokes. Damped strokes do not elicit a sense of pitch; it is better described as a relative sense of "higher" or "lower."

---

* Identified from spectral analysis. See Fig. 3.13 and Fig. 3.14.

For damped strokes, the source detector attempts to distinguish high drum from low drum on the basis of spectral "profile,", but makes some mistakes in this process. However, it is not always possible for the listener to distinguish between these either, and this decision is most crucial (and obvious to the listener) for the undamped strokes, which are easy for the program to distinguish. The sense of the pattern can be adequately represented by a notation that simply represents high and low drums; the exact pitch is interesting to note but not crucial to the structure of the music. It is clear that the question of pitch perception of drums as a theoretical issue should be pursued further, but for the purposes of transcription, the method of finding the peak of the spectrum of the undamped strokes will suffice.

On Accent Detection

Detection of accents could be considered a subtask of detection of articulation, which is a very difficult task [Strawn, 1982]. In the case of percussive music, it seems that the key to an *intrinsic* accent (as opposed to an *implied* accent, which is induced by musical context) is steepness of slope of envelope coupled with spectral width. For this reason, the slap is heard as an accent. Surprisingly, amplitude *per se* is not as important a clue. At the moment, the program identifies the slap as an accent. This is indeed what listeners report when listening to examples.

On Rest Detection

As mentioned in Section 2.3.3, what really matters in describing a given rhythm is *inter-attack time*. The duration of a certain note, if it ends before the next attack, defines the remaining time as a rest. This partial duration, and the rest following it, are not as perceptually salient as the full duration implied by the time interval between attacks. In fact, one might say that the latter defines the rhythm, whereas the former defines the *articulation* of the rhythm. For this reason, more emphasis is put on detecting attacks than detecting rests.

3.3. Approach to the High-level Analysis

Now that the analysis has proceeded to the level of the notelist we have an accurate description of the musical events. This is a significant point of arrival, but it is clearly not the end result; in fact, if we were to proceed from keyboard data instead of from the acoustic signal, this would be the point of departure. There are many possible directions leading from this level of analysis, toward an abstraction of the numbers (a score), or in the direction of performance/synthesis; one can use the notelist to drive other instruments in arbitrarily complex relations to the original music. These directions are interrelated.

To proceed towards a musical transcription from this notelist, we need "hooks" into the data, or ways of extracting what is musically meaningful from a list of numbers. Since we are dealing mostly with rhythm here, we will be concerned first of all with timing information and tempo. As mentioned earlier, tempo and performance fluctuations can result in an eighth note in one part of a piece being longer in duration than a quarter note not far away. It is clear that the numbers cannot just be plotted as they are. Although this kind of map has some useful features, it is not musically enlightening.

The next level of analysis, intended to proceed from the notelist to the score, is based on work by Bernard Mont-Reynaud that was not originally designed to deal with percussive music. (See [Mont-Reynaud, 1984] for a more detailed description of this part of the system.) It is interesting to note that this higher-level program, though it was originally intended for application to tonal music, is sufficiently general (not attached to a particular style) that it can be applied to these data with considerable success. Its concentration on temporal analysis is an interesting feature that will be described in this section.

It turns out that the basic problems of tempo and meter can be approached via simple primitives in the music. The method is sufficiently robust that it not only handles diverse styles, but also performances that are far from mechanical. (Of course, mechanical performances are difficult to listen to but easy to analyze.) The foremost goal is to "factor out" the tempo variations.

In this section, we will continue with the analysis of the excerpt of drumming whose notelist is shown in Fig. 3.20.

3.3.1. Important Durations

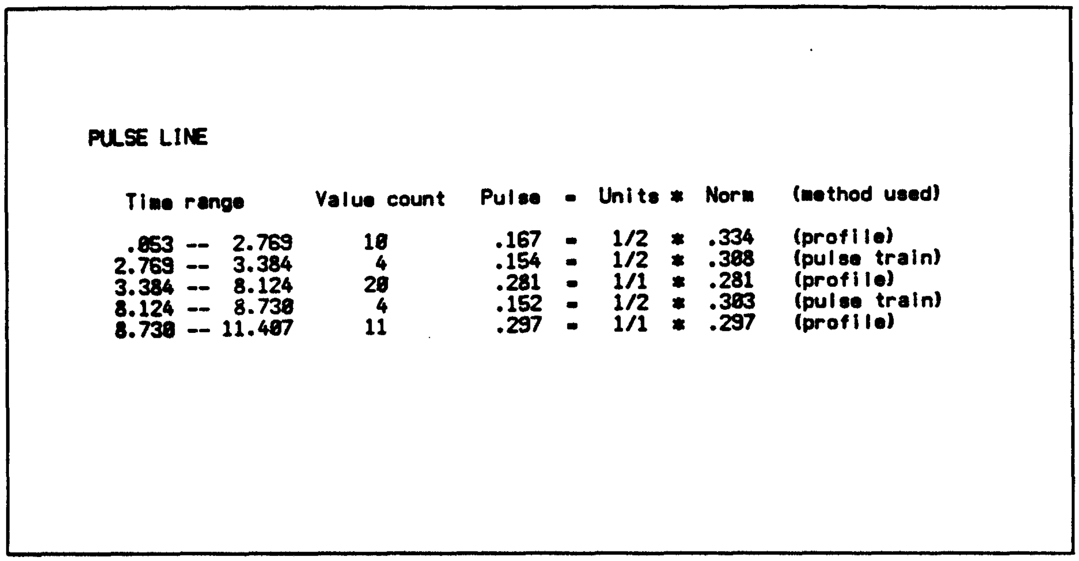
The first step is to scan the durations of the notes in the notelist, and look for sequences of repeated durations (within a certain tolerance for accepted fluctuation and minimal number in sequence). The average duration of each sequence found is noted, and the span during which the durations were repeated is marked. This group of repeated durations is called a "pulse train." For example, in the notelist (Fig. 3.20), we can see that from time 2.769 to 3.384, there are four successive notes of nearly equal duration, whose average value (.154) is saved as a pulse train in Fig. 3.22.

Spans in which there are no such pulse trains are also marked, and broken into smaller pieces by the program until they are short enough so that duration relations will not be obscured by tempo changes, but long enough to have statistical significance. Within this span, we look for frequently-occurring durations, (not contiguous, as in the pulse train). To find this most frequent duration, the program creates smoothed histograms of durations during the non-pulse-train spans, and looks for predominant peaks.

In Fig. 3.21, we see a histogram from the last section of the example. It is designed to show important durations and their relationships in several ways. First, a logarithmic scale for durations is used, because ratios are more important than differences in the rhythmic (macro-temporal) domain, just as in the domain of pitch (micro-temporal) at the lower level. Second, each duration value in the histogram is weighted by its square root, in order to emphasize longer durations. We do not want the short notes to dominate the histogram, as they are much more common. Thirdly, the histograms are smoothed; each duration contributes to neighboring cells, in a sort of convolution of the data with a Gaussian curve. The highest peak in the described histogram during its span of time is chosen as the *important duration.*

```
SMOOTHED HISTOGRAM OF LOG(DUR), time range [  8.730 -- 11.407]

DURATION | DENSITY  (pulse is  .297)
    .088 |-------------------
    .094 |-----------------------
    .099 |-----------------------
    .105 |---------------------
    .111 |-----------------
    .118 |--------------
    .125 |--------
    .132 |-------
    .140 |--------
    .149 |------------
    .157 |------------------
    .167 |---------------------
    .177 |----------------
    .187 |----------
    .198 |-----
    .210 |---
    .223 |-----
    .236 |-----------
    .250 |--------------------
    .265 |--------------------------------
    .281 |--------------------------------------
    .297 |****************************************
    .315 |-------------------------------
    .334 |---------------------------
    .354 |---------------
    .375 |--------
    .397 |---
    .420 |-
    .445 |
    .472 |
    .500 |--
    .530 |------
    .561 |-----------
    .595 |----------------
    .630 |-------------------
    .667 |-------------
    .707 |--------
    .749 |---
    .794 |-
    .841 |
    .891 |
    .944 |
   1.000 |
   1.059 |
   1.122 |
   1.189 |
   1.260 |
   1.335 |
   1.414 |
   1.498 |
   1.587 |
```

Figure 3.21. Example of smoothed histogram of durations where no pulse is found. The vertical scale (duration) is logarithmic, and the histogram is weighted to avoid overemphasis of the more common shorter durations.



```
PULSE LINE

  Time range       Value count   Pulse  =  Units *  Norm    (method used)

  .053 --  2.769       10         .167  =   1/2  *  .334    (profile)
 2.769 --  3.384        4         .154  =   1/2  *  .308    (pulse train)
 3.384 --  8.124       20         .281  =   1/1  *  .281    (profile)
 8.124 --  8.730        4         .152  =   1/2  *  .303    (pulse train)
 8.730 -- 11.407       11         .297  =   1/1  *  .297    (profile)
```

Figure 3.22. Summary of pulse train and histograms.

Now that we have two complementary notions of important durations, they can be summarized for the whole piece, as shown in Fig. 3.22. Included in this table are the important durations (Pulse), their derivation (method used— average of pulse train or peak in histogram, as described above), the time span over which it was computed (Time range), the number of events in the associated span (Value count), and a first approximation of ratios between successive values (Units), with an associated "warping" value that makes the ratio exact (Norm).

3.3.2. Accents and Anchor Points

At this point, the *reference unit* begins to be established; this is a canonical unit in terms of which all durations are represented. Context plays a large part in driving this decision, though it may be difficult to determine better than a factor of two above or below the actual unit of the time signature (if indeed there is a time signature). For example, in 4/4 time, the reference unit may be found to be the eighth note instead of the quarter note. There is not a great loss of generality here. The decision is based on finding a pulse that is most commensurate with other values found.

Some definition of "commensurate" is needed here. The simplest ratio is of course 1:1, after that, 2:1, 1:2, 1:3 etc. The goal is to combine accuracy of the ratios with goodness of fit. An approach to this problem is described in the next section.

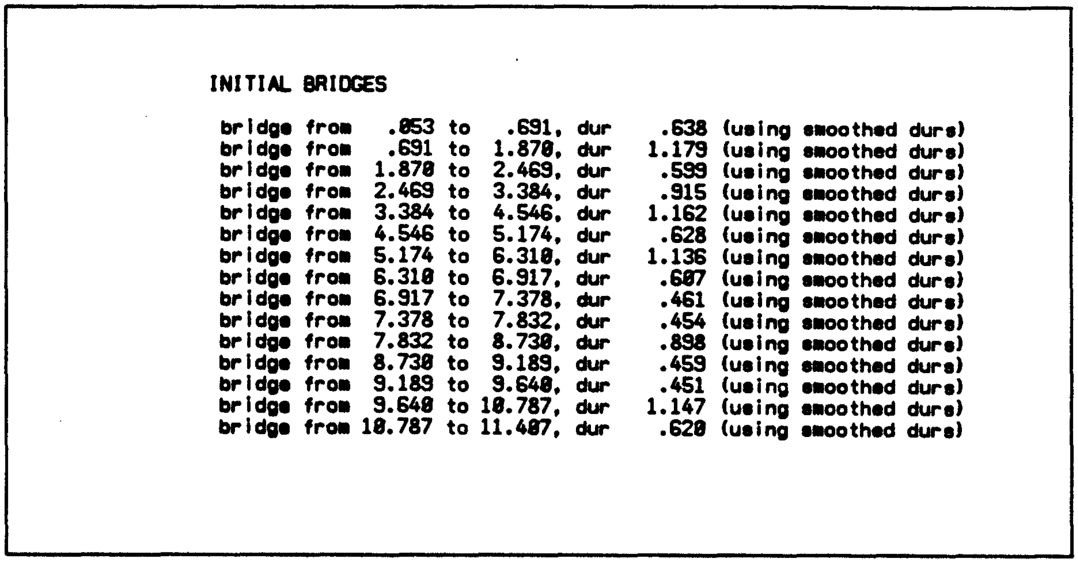
The first approximation to *important durations* will enable further analysis, when combined with another fundamental feature in the data, that of *important events*. A simple, but powerful insight into this problem stems from defining important events (indeed, *anchor points* in the music), as agogic accents. That is, we search in the data for a "short-long" pattern, marking the long as a rhythmic accent if it is *significantly longer* than the previous duration, and is not followed by a longer duration by the same criterion. The details of this heuristic are described in [Chafe, et al., 1982]. There is a large body of work that points to this duration rule as a fundamental one. See Section 2.3.2 for references regarding agogic accents in general.

This identification of anchor points is carried throughout the analysis. The rhythmic structure of a large body of music typically bears an identifiable relation to repetitions at regular intervals of these accents. This allows the system to try to divide the span between accents into metrical pieces, or multiples of a basic pulse. In fact, the establishment of rhythmic accents allows us to take the first step in establishing a tempo line. Figure 3.23 shows how the spans between rhythmic accents break this example into fifteen segments.

So far, the rhythmic accents, as shown in Fig. 3.23, and the pulse line, as shown in Fig. 3.22, have not been integrated. At this point, the system has enough information to create a line-segment approximation that maps physical time to metronomic time. This map, called a *tempo line*, has powerful implications. Once determined, it allows us to "factor out" tempo variations, and see what stylistic variations are accounted for by alterations in the regular metronomic structure. This is tantamount to Gabrielsson's SYVAR (Systematic Variations—see [Bengtsson and Gabrielsson, 1980]).

The endpoints of the line segments will correspond to the rhythmic accents. We try to adjust the local tempo segment so that:

- Successive bridges are commensurate.
- Bridges are commensurate with associated pulse values.



```
INITIAL BRIDGES

 bridge from   .053 to   .691, dur   .638 (using smoothed durs)
 bridge from   .691 to  1.870, dur  1.179 (using smoothed durs)
 bridge from  1.870 to  2.469, dur   .599 (using smoothed durs)
 bridge from  2.469 to  3.384, dur   .915 (using smoothed durs)
 bridge from  3.384 to  4.546, dur  1.162 (using smoothed durs)
 bridge from  4.546 to  5.174, dur   .628 (using smoothed durs)
 bridge from  5.174 to  6.310, dur  1.136 (using smoothed durs)
 bridge from  6.310 to  6.917, dur   .607 (using smoothed durs)
 bridge from  6.917 to  7.378, dur   .461 (using smoothed durs)
 bridge from  7.378 to  7.832, dur   .454 (using smoothed durs)
 bridge from  7.832 to  8.730, dur   .898 (using smoothed durs)
 bridge from  8.730 to  9.189, dur   .459 (using smoothed durs)
 bridge from  9.189 to  9.640, dur   .451 (using smoothed durs)
 bridge from  9.640 to 10.787, dur  1.147 (using smoothed durs)
 bridge from 10.787 to 11.407, dur   .620 (using smoothed durs)
```

Figure 3.23. Summary of intervals defined by rhythmic accents, and their durations.

3.3.3. Rational Approximation to Metric Unit

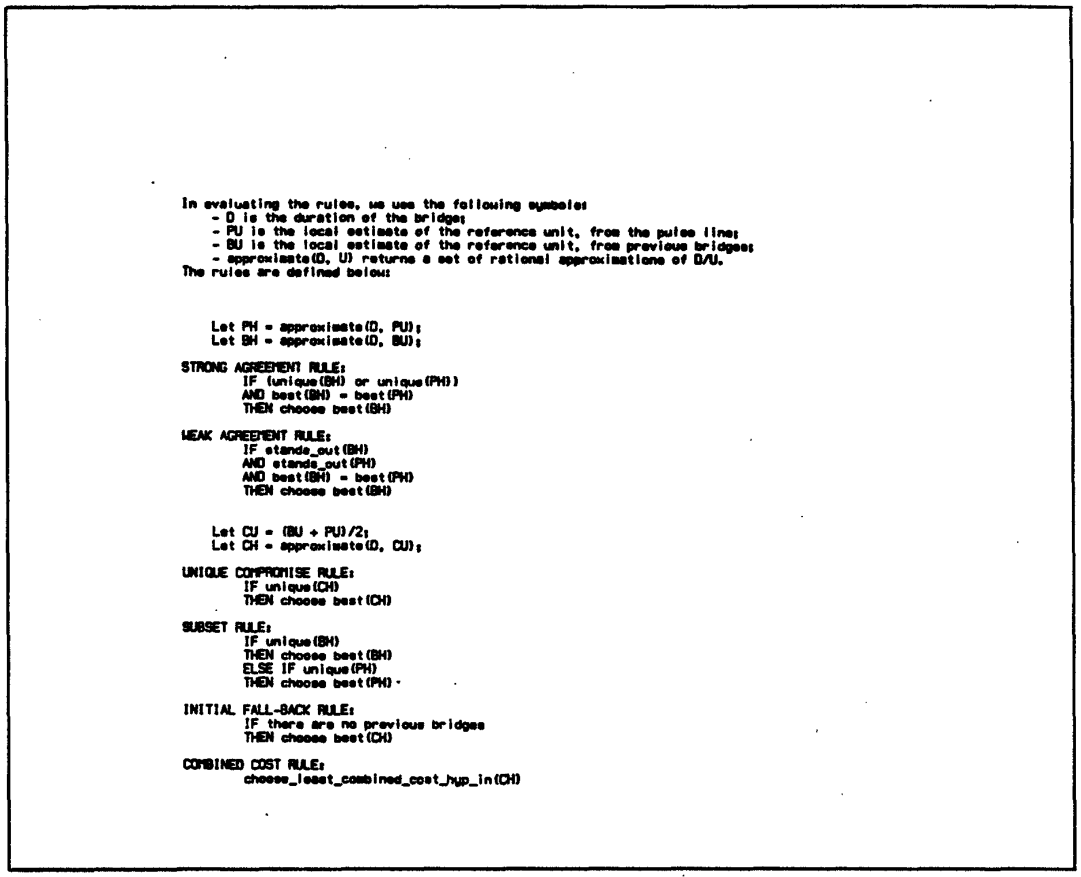
What does commensurate mean and how do we quantify it? This question goes to the heart of the issue of rhythm. As we saw in Chapter 2, a global theory of rhythm, though approached through very different paradigms, still is couched in some level of categorical perception of duration. That is, although the notion of *subdivision* is mainly a Western idea, it is still necessary to put different durations in different categories in order to recognize rhythmic patterns. The "trickier" a rhythmic pattern is, the more crititcal it is to play it accurately. This implies that the timing tolerance within a given musical context goes down with syncopated passages, lending credence to the possibility of distinguishing tempo fluctuation from syncopation at the level of the analysis system: "nasty" syncopation *has* to be played accurately (close to rational subdivisions); otherwise the effect of the syncopation is lost. Conversely, simple, regular, easily parsed figures can deviate much more from the normative time values and still be interpreted correctly. There is a tendency to "normalize" durations; to find simple rational approximations to a given time span. This means that though there will always be intentional deviations from "ideal" patterns, we must make an attempt to rationalize durations if we are to create a coherent representation of what we hear.

The implication here is that we need a robust, tunable rational approximation method that finds simple metric proportions between performed durations. There is obviously a trade-off here between closeness of fit to the data, and simplicity of the fraction, in the analysis method as well as in the performance. The dichotomy exists in the music; the analysis is merely recognizing it.

There is not necessarily a single best answer to this problem. The rational approximation generator will rate possible fractions, and retain a small set of possible solutions. It is given a set of acceptable fraction denominators, constraints on numerators, a criterion to determine relative simplicity of fractions, and a measure of fit between a given number and a rational approximation. There is also a partial ordering established in the two dimensions of simplicity and fit, defined as follows: (x_1, y_1) dominates (x_2, y_2) if $x_1 \leq x_2$ and $y_1 \leq y_2$ and $(x_1, y_1) \neq (x_2, y_2)$. Here, x could be a measure of simplicity of a fraction, and y could be its fit to a rational number. (For further detail on the rational approximation methods, see [Mont-Reynaud, 1984, pp. 20-22].)

Since context can be an important factor in making decisions as to the rational approximations, the program will redefine some fractions as being simpler if they fit an hypothesis for meter, for example in ternary meter, 2/3 might dominate 1/4 in terms of simplicity.

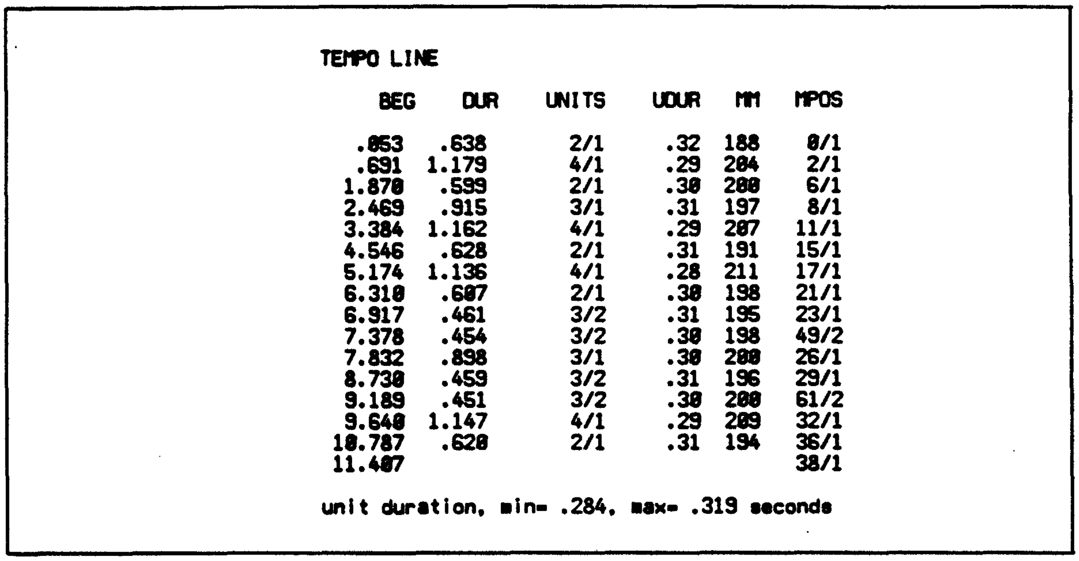
At this point, an attempt is made to integrate the bridge lengths with the pulse train and profile statistics. They have been determined independently. The intention is to account for the relation between bridge length and local pulse/profile, that is, to use the rational approximation generator as applied in the following way: using the set of rules given in Fig. 3.24, estimate the local unit during each bridge by comparing rational approximations to the bridge unit vs. the pulse unit, and decide on the best number of reference units for each bridge. In this way, tempo variations are determined piece-wise linearly with each segment equal to the length of the respective bridge.



```
In evaluating the rules, we use the following symbols:
    - D is the duration of the bridge;
    - PU is the local estimate of the reference unit, from the pulse line;
    - BU is the local estimate of the reference unit, from previous bridges;
    - approximate(D, U) returns a set of rational approximations of D/U.
The rules are defined below:


    Let PH = approximate(D, PU);
    Let BH = approximate(D, BU);

STRONG AGREEMENT RULE:
        IF (unique(BH) or unique(PH))
        AND best(BH) = best(PH)
        THEN choose best(BH)

WEAK AGREEMENT RULE:
        IF stands_out(BH)
        AND stands_out(PH)
        AND best(BH) = best(PH)
        THEN choose best(BH)


    Let CU = (BU + PU)/2;
    Let CH = approximate(D, CU);

UNIQUE COMPROMISE RULE:
        IF unique(CH)
        THEN choose best(CH)

SUBSET RULE:
        IF unique(BH)
        THEN choose best(BH)
        ELSE IF unique(PH)
        THEN choose best(PH)

INITIAL FALL-BACK RULE:
        IF there are no previous bridges
        THEN choose best(CH)

COMBINED COST RULE:
        choose_least_combined_cost_hyp_in(CH)
```

Figure 3.24. Rule system for tempo line decisions. In this figure, the predicate unique(H) is true if the set H of hypotheses contains a single member. The function best(H) returns the hypothesis with minimum cost in H. Cost means the linear combination of complexity and fit computed by the approximation number generator.

The combined cost rule selects an hypothesis with minimum combined cost in CH, the set obtained from the compromise estimate. The combined cost of hypothesis H for duration D is the sum of separate costs, expressing the fact that one wants both a good relationship to CU, the duration of the compromise reference unit, and to the duration and metric value of the previous bridge. Note that the hypothesis selected is not always the one preferred by BU, PU or even CU ratings, and that it does not always have the simplest relationship to the previous bridge either. Thus all these criteria carry some weight in the decision, but one resorts to numerical combinations of ratings only when stronger forms of selection have all failed. (Above figure and caption from [Mont-Reynaud, 1984])



```
TEMPO LINE

     BEG    DUR      UNITS    UDUR   MM   MPOS

    .053   .638       2/1     .32  188    0/1
    .691  1.179       4/1     .29  204    2/1
   1.870   .599       2/1     .30  200    6/1
   2.469   .915       3/1     .31  197    8/1
   3.384  1.162       4/1     .29  207   11/1
   4.546   .628       2/1     .31  191   15/1
   5.174  1.136       4/1     .28  211   17/1
   6.310   .607       2/1     .30  198   21/1
   6.917   .461       3/2     .31  195   23/1
   7.378   .454       3/2     .30  198   49/2
   7.832   .898       3/1     .30  200   26/1
   8.730   .459       3/2     .31  196   29/1
   9.189   .451       3/2     .30  200   61/2
   9.640  1.147       4/1     .29  209   32/1
  10.787   .620       2/1     .31  194   36/1
  11.407                                 38/1

unit duration, min= .284, max= .319 seconds
```

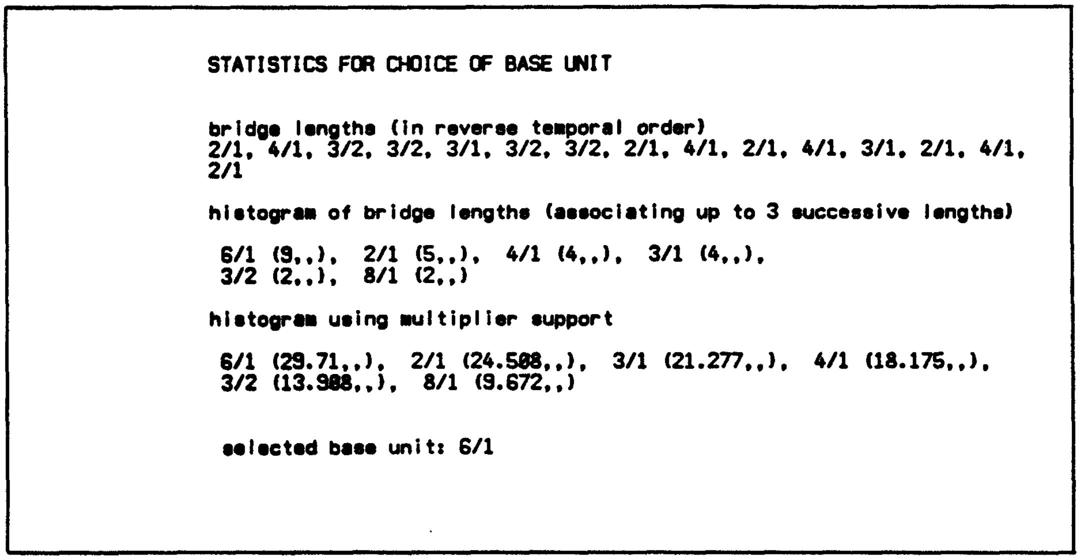
Figure 3.25. The first approximation of a tempo line.

3.3.4. Tempo Line

Upon completion of the rule-based bridge vs. pulse decisions, it is possible to create a first approximation of a tempo line, which is shown in Fig. 3.25. Now, the UNITS column is a direct summary of the units in each bridge.

The MPOS column is a running sum of these chosen metric units, whose durations fluctuate slightly around the value .30 seconds (UDUR). We can now compute musical time as a metronome marking for each bridge (MM column). It is evident from the MM column that this example is relatively stable with respect to tempo, but even the amount of tempo variation found here would be enough to perturb results were it not for this process.

The choice of the reference unit is based mostly on statistical analysis of individual note durations, which makes sense. There should be also be a way to look for a longer periodicity, composed of *groups* of the metric unit, that we call the *base unit*. It corresponds roughly to the level of a measure in music that has bars (in Class 3 music, it would correspond to the meso-period, modulo a power of 2).



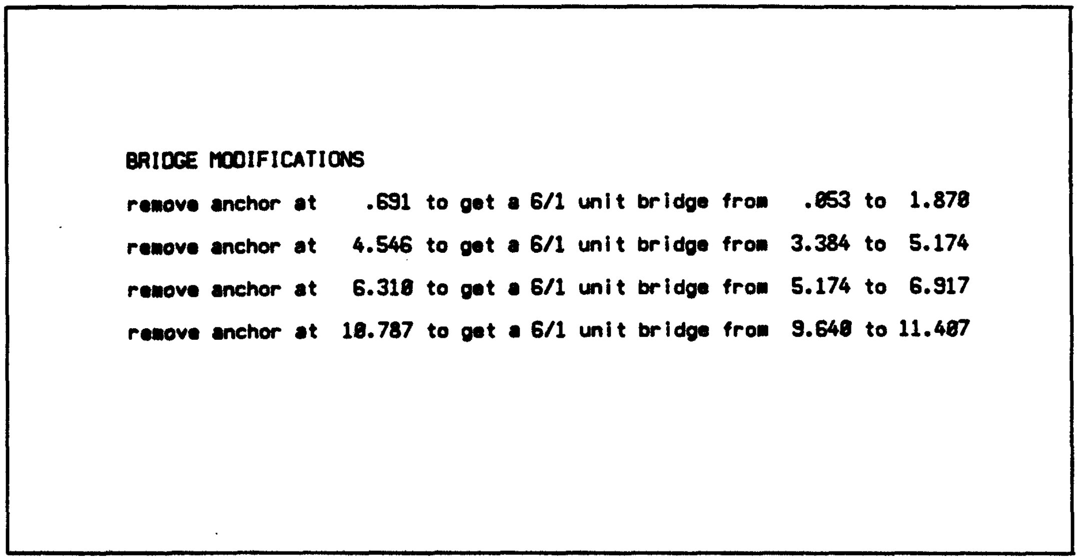
```
STATISTICS FOR CHOICE OF BASE UNIT

bridge lengths (in reverse temporal order)
2/1, 4/1, 3/2, 3/2, 3/1, 3/2, 3/2, 2/1, 4/1, 2/1, 4/1, 3/1, 2/1, 4/1,
2/1

histogram of bridge lengths (associating up to 3 successive lengths)

 6/1 (9,,),  2/1 (5,,),  4/1 (4,,),  3/1 (4,,),
 3/2 (2,,),  8/1 (2,,)

histogram using multiplier support

 6/1 (29.71,,),  2/1 (24.508,,),  3/1 (21.277,,),  4/1 (18.175,,),
 3/2 (13.908,,),  8/1 (9.672,,)


 selected base unit: 6/1
```

Figure 3.26. Searching for the choice of Base Unit.

The approach here is first to count the metric units concatenated that are shown in Fig. 3.25, column 3. In this example, the candidate 2/1 occurs four times. Looking for frequency is important for a base unit, but equally important is that it must be a central ratio in terms of itself and its simple divisors and multiples (again the idea of salient units being commensurate). We call this *multiplier support*, as shown in Fig. 3.26. This is a measure of occurrence of the unit and its closest multiplicative neighbors. In this example, the base unit found (6/1) corresponds to six eighth notes, which implies that the meter is in a group of six units, an encouraging result at this point.

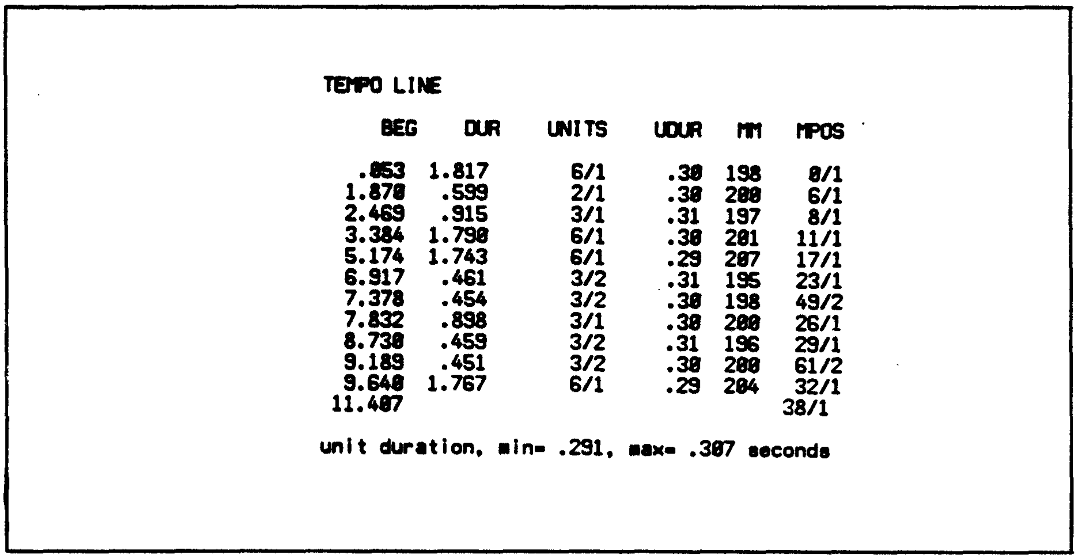
3.3.5. Tempo Line Refinement

A refinement of the tempo line from the base unit is shown in Fig. 3.27. Here we are trying to divide the bridges into smaller pieces that are some multiple of the base unit. This division is repeated until either the new interval is shorter than the base unit, or there is no event whose onset is close enough to the target division. Conversely, if the bridge length is smaller than the base unit, an attempt is made to omit the intermediate anchor if another anchor point can be found whose value is an integral number of units beyond. The result is shown in Fig. 3.28. The new



```
BRIDGE MODIFICATIONS

remove anchor at    .691 to get a 6/1 unit bridge from   .053 to  1.870

remove anchor at   4.546 to get a 6/1 unit bridge from  3.384 to  5.174

remove anchor at   6.310 to get a 6/1 unit bridge from  5.174 to  6.917

remove anchor at  10.787 to get a 6/1 unit bridge from  9.640 to 11.407
```

Figure 3.27. Trying to divide the bridges into smaller pieces that are some multiple of the chosen base unit.



```
TEMPO LINE

      BEG     DUR     UNITS    UDUR   MM    MPOS

     .053   1.817      6/1     .30   198    0/1
    1.870    .599      2/1     .30   200    6/1
    2.469    .915      3/1     .31   197    8/1
    3.384   1.790      6/1     .30   201   11/1
    5.174   1.743      6/1     .29   207   17/1
    6.917    .461      3/2     .31   195   23/1
    7.378    .454      3/2     .30   198   49/2
    7.832    .898      3/1     .30   200   26/1
    8.730    .459      3/2     .31   196   29/1
    9.189    .451      3/2     .30   200   61/2
    9.640   1.767      6/1     .29   204   32/1
   11.407                                  38/1

unit duration, min= .291, max= .307 seconds
```

Figure 3.28. Revised tempo line, based on new bridge lengths. The UNITS column has been simplified from that of Fig. 3.25.

tempo line has, in the UNITS column, much simpler ratios than the previous tempo line shown in Fig. 3.25.

A set of possible rhythmic values for each note in the example can now be assigned in terms of the reference unit. This is tantamount to generating rational approximations for each duration, based on the unit length in seconds of Fig. 3.28. There will typically be several candidates for a given value, that decrease in closeness of fit to the target value. Figure 3.29 shows, in the APPROX field, collections of such rational approximations to the field called RDUR, which is the smoothed duration (Sdur) divided by the unit duration (UDUR from Fig. 3.28).

The VAL field in Fig. 3.29 is partially filled. This first pass at filling it in is realized by trying the closest fitting rational approximation (which is therefore defined to be lowest cost), for each note, and seeing whether the collection of these during each bridge adds up to the total number of bridge units. If so, these values are written in. This process does not involve a search; it simply checks for obvious answers, and in so doing, creates a strong context.

It is helpful to use the context created by the values already found. Figure 3.30 shows how we will bias the rational approximation generator by finding the "popularity" of the possible units, and use this to change the cost of choosing one rational candidate over another. The COUNT field is just the total of occurrences in the VAL field plus occurrences in the first position of APPROX field in Fig. 3.29. The COST is 1/(COUNT+1).

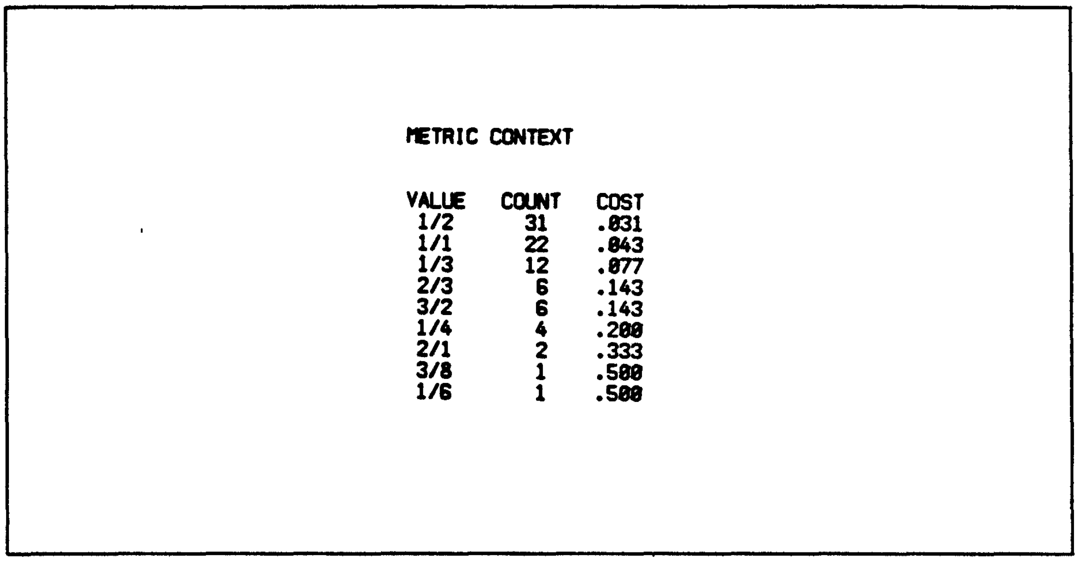
3.3.6. Determining Normalized Rhythmic Values

Figure 3.31 shows how the APPROX field has been altered and pruned because of the change of priorities of the rational approximations. For example, in bar 1, we see that for the fifth note, the rational approximations are (1/4, 1/6, 1/3), whereas in Fig. 3.29 they were (1/6, 1/4, 1/8, 1/3). The change is due to the lower penalty for 1/4 (cost = .20) as opposed to a cost of .5 for 1/6 (see Fig. 3.30). Often the different order of possible values for rational approximations will allow more VAL entries to be filled in without a search, if the first values in the APPROX field sum correctly. In this particular case, we do not have more VAL entries in Fig. 3.31 than in Fig. 3.29, but since the wanted value (WVAL) for this note is 1/4, the new first choice for the rational approximation is in fact correct, and this will facilitate the next step.

At this point there are still several VAL entries that have not been filled, because the sum of the first entries in each APPROX field does not equal the correct number

```
   ART     BEG   DUR    SBEG   SDUR  Bu   Rdur  VAL    OK  WVAL  WBAR   WrPOS APPROX;
 PLAY;
 ∂ BAR 1 (from hints);
  HSLAP   .053  .455   .053  .455  0/1 1.502   ?     ?    3/2    1   0/1 (3/2, 2/1, 1/1);
  HBASS   .508  .183   .508  .183   /   .604   ?     ?    1/2    1   3/16(2/3, 1/2, 3/4);
  LOPEN   .691  .285   .691  .285   /   .941   ?     ?    1/1    1   1/4 (1/1, 3/4, 2/3, 4/3, 5/4, 3/2, 1/2, 5/3);
  HBASS   .976  .071   .976  .071   /   .234   ?     ?    1/4    1   3/8 (1/4, 1/3, 1/8);
  LSLAP  1.047  .057  1.047  .057   /   .188   ?     ?    1/4    1  13/32(1/6, 1/4, 1/8, 1/3);
  LMUFF  1.104  .130  1.104  .130   /   .429   ?     ?    1/2    1   7/16(1/2, 3/8, 1/3);
  LBASS  1.234  .162  1.234  .162   /   .535   ?     ?    1/2    1   1/2 (1/2);
  HOPEN  1.396  .166  1.396  .166   /   .548   ?     ?    1/2    1   9/16(1/2, 2/3);
  HOPEN  1.562  .308  1.562  .308   /  1.017   ?     ?    1/1    1   5/8 (1/1, 4/3, 3/4, 5/4, 3/2, 2/3, 2/1, 5/3);
 ∂ BAR 2 (from hints);
  LOPEN  1.870  .435  1.870  .435  2/1 1.452  3/2   1/1   3/2    2   0/1 (3/2, 4/3, 2/1, 1/1);
  HOPEN  2.305  .164  2.305  .164   /   .548  1/2    /    1/2    2   3/16(1/2, 2/3);
  HOPEN  2.469  .300  2.469  .300  3/1  .984  1/1    /    1/1    2   1/4 (1/1, 3/4, 4/3, 5/4, 2/3, 3/2, 1/2, 5/3);
  LSLAP  2.769  .148  2.769  .154   /   .504  1/2    /    1/2    2   3/8 (1/2);
  LSLAP  2.917  .165  2.923  .154   /   .504  1/2    /    1/2    2   7/16(1/2);
  HOPEN  3.082  .147  3.077  .154   /   .504  1/2    /    1/2    2   1/2 (1/2);
  HOPEN  3.229  .155  3.230  .154   /   .504  1/2    /    1/2    2   9/16(1/2);
  LOPEN  3.384  .294  3.384  .294  0/1  .985   ?     ?    1/1    2   5/8 (1/1, 3/4, 4/3, 5/4, 2/3, 3/2, 1/2, 5/3);
 ∂ BAR 3 (from hints);
  HMUFF  3.678  .079  3.678  .079   /   .265   ?     ?    1/3    3   0/1 (1/4, 1/3, 1/2);
  HMUFF  3.757  .107  3.757  .107   /   .359   ?     ?    1/3    3   1/24(1/3, 3/8, 1/2);
  LOPEN  3.864  .186  3.864  .186   /   .623   ?     ?    2/3    3   1/12(2/3, 1/2, 3/4, 1/1);
 COMMENT MISSING          1/3 ;
  LOPEN  4.050  .213  4.050  .213   /   .714   ?     ?    2/3    3   1/6 (2/3, 3/4, 1/1, 1/2);
 COMMENT MISSING          1/3 ;
  HOPEN  4.263  .283  4.263  .283   /   .949   ?     ?    1/1    3   1/4 (1/1, 3/4, 2/3, 4/3, 5/4, 3/2, 1/2, 5/3);
  LSLAP  4.546  .444  4.546  .444   /  1.488   ?     ?    3/2    3   3/8 (3/2, 4/3, 2/1, 1/1);
  HSLAP  4.990  .184  4.990  .184   /   .617   ?     ?    1/2    3   9/16(2/3, 1/2, 3/4);
  LOPEN  5.174  .287  5.174  .287  0/1  .988  1/1   1/1   1/1    3   5/8 (1/1, 3/4, 4/3, 5/4, 2/3, 3/2, 1/2, 5/3);
 ∂ BAR 4 (from hints);
  HMUFF  5.461  .097  5.461  .097   /   .334  1/3    /    1/3    4   0/1 (1/3, 1/2);
  HOPEN  5.558  .185  5.558  .185   /   .637  2/3    /    2/3    4   1/24(2/3, 1/2, 3/4, 1/1);
 COMMENT MISSING          1/3 ;
  HOPEN  5.743  .102  5.743  .102   /   .351  1/3    /    1/3    4   1/8 (1/3, 3/8, 1/2);
  HOPEN  5.845  .086  5.845  .086   /   .296  1/3    /    1/3    4   1/6 (1/3, 1/4, 3/8, 1/2);
  HMUFF  5.931  .097  5.931  .097   /   .334  1/3    /    1/3    4   5/24(1/3, 1/2);
  LOPEN  6.028  .282  6.028  .282   /   .971  1/1    /    1/1    4   1/4 (1/1, 3/4, 4/3, 2/3, 5/4, 3/2, 1/2, 5/3);
  LSLAP  6.310  .449  6.310  .449   /  1.546  3/2    /    3/2    4   3/8 (3/2, 4/3, 5/3, 2/1, 1/1);
  HOPEN  6.759  .158  6.759  .158   /   .544  1/2    /    1/2    4   9/16(1/2, 2/3);
  LOPEN  6.917  .310  6.917  .310  3/2 1.068  1/1    /    1/1    4   5/8 (1/1, 4/3, 3/4, 5/4, 3/2, 2/3, 2/1, 5/3);
 ∂ BAR 5 (from hints);
  HMUFF  7.227  .151  7.227  .151   /   .491  1/2    /    1/2    5   0/1 (1/2);
  HSLAP  7.378  .287  7.378  .287  3/2  .948  1/1    /    1/1    5   1/16(1/1, 3/4, 2/3, 4/3, 5/4, 3/2, 1/2, 5/3);
  LSLAP  7.665  .167  7.665  .167   /   .552  1/2    /    1/2    5   3/16(1/2, 2/3);
  HOPEN  7.832  .292  7.832  .292  3/1  .978  1/1    /    1/1    5   1/4 (1/1, 3/4, 4/3, 2/3, 5/4, 3/2, 1/2, 5/3);
  LSLAP  8.124  .155  8.124  .152   /   .506  1/2    /    1/2    5   3/8 (1/2);
  HSLAP  8.279  .155  8.275  .152   /   .506  1/2    /    1/2    5   7/16(1/2);
  HMUFF  8.434  .150  8.427  .152   /   .506  1/2    /    1/2    5   1/2 (1/2);
  LOPEN  8.584  .146  8.578  .152   /   .506  1/2    /    1/2    5   9/16(1/2);
  LOPEN  8.730  .289  8.730  .289  3/2  .944  1/1    /    1/1    5   5/8 (1/1, 3/4, 2/3, 4/3, 5/4, 3/2, 1/2, 5/3);
 ∂ BAR 6 (from hints);
  HSLAP  9.019  .170  9.019  .170   /   .556  1/2    /    1/2    6   0/1 (1/2, 2/3);
  LSLAP  9.189  .290  9.189  .290  3/2  .965  1/1    /    1/1    6   1/16(1/1, 3/4, 4/3, 2/3, 5/4, 3/2, 1/2, 5/3);
  LSLAP  9.479  .161  9.479  .161   /   .535  1/2    /    1/2    6   3/16(1/2);
  HOPEN  9.640  .283  9.640  .283  3/1  .893   ?     ?    1/1    6   1/4 (1/1, 3/4, 2/3, 1/2);
  HMUFF  9.903  .073  9.903  .073   /   .248   ?     ?    1/3    6   3/8 (1/4, 1/3, 1/8);
  HMUFF  9.976  .077  9.976  .077   /   .261   ?     ?    1/3    6   5/12(1/4, 1/3, 1/2);
  HOPEN 10.053  .089 10.053  .089   /   .302   ?     ?    1/3    6  11/24(1/3, 1/4, 3/8, 1/2);
  HOPEN 10.142  .102 10.142  .102   /   .346   ?     ?    1/3    6   1/2 (1/3, 3/8, 1/2);
  LOPEN 10.244  .095 10.244  .095   /   .323   ?     ?    1/3    6  13/24(1/3, 1/2);
  HOPEN 10.339  .116 10.339  .116   /   .394   ?     ?    1/3    6   7/12(3/8, 1/3, 1/2);
  HOPEN 10.455  .332 10.455  .332   /  1.127   ?     ?    1/1    6   5/8 (1/1, 4/3, 5/4, 3/4, 2/1, 2/3);
 ∂ BAR 7 (from hints);
  LSLAP 10.787  .820 10.787  .820   /  2.105  2/1   1/1   2/1    7   0/1 (2/1);
 FINISH;
```

Figure 3.29. Notelist with Rational Approximations, and the first guess at the value (VAL) for each duration. ART—Articulation. BEG—Begin time. DUR—Duration. SBEG, SDUR—Smoothed beg and dur. Bu—Bridge Unit. Rdur—Relative Unit. VAL—Rational Approximation for durations. OK—For debugging. '1/1' or '/' means correct value. WVAL—Wanted value. WBAR—For debugging. Tells what measure you're in. WrPOS—Wanted metric position in each bar. (Also for debugging). APPROX—List of rational approximations ordered by closeness to normalized duration given.



```
METRIC CONTEXT

VALUE   COUNT   COST
 1/2     31    .031
 1/1     22    .043
 1/3     12    .077
 2/3      6    .143
 3/2      6    .143
 1/4      4    .200
 2/1      2    .333
 3/8      1    .500
 1/6      1    .500
```

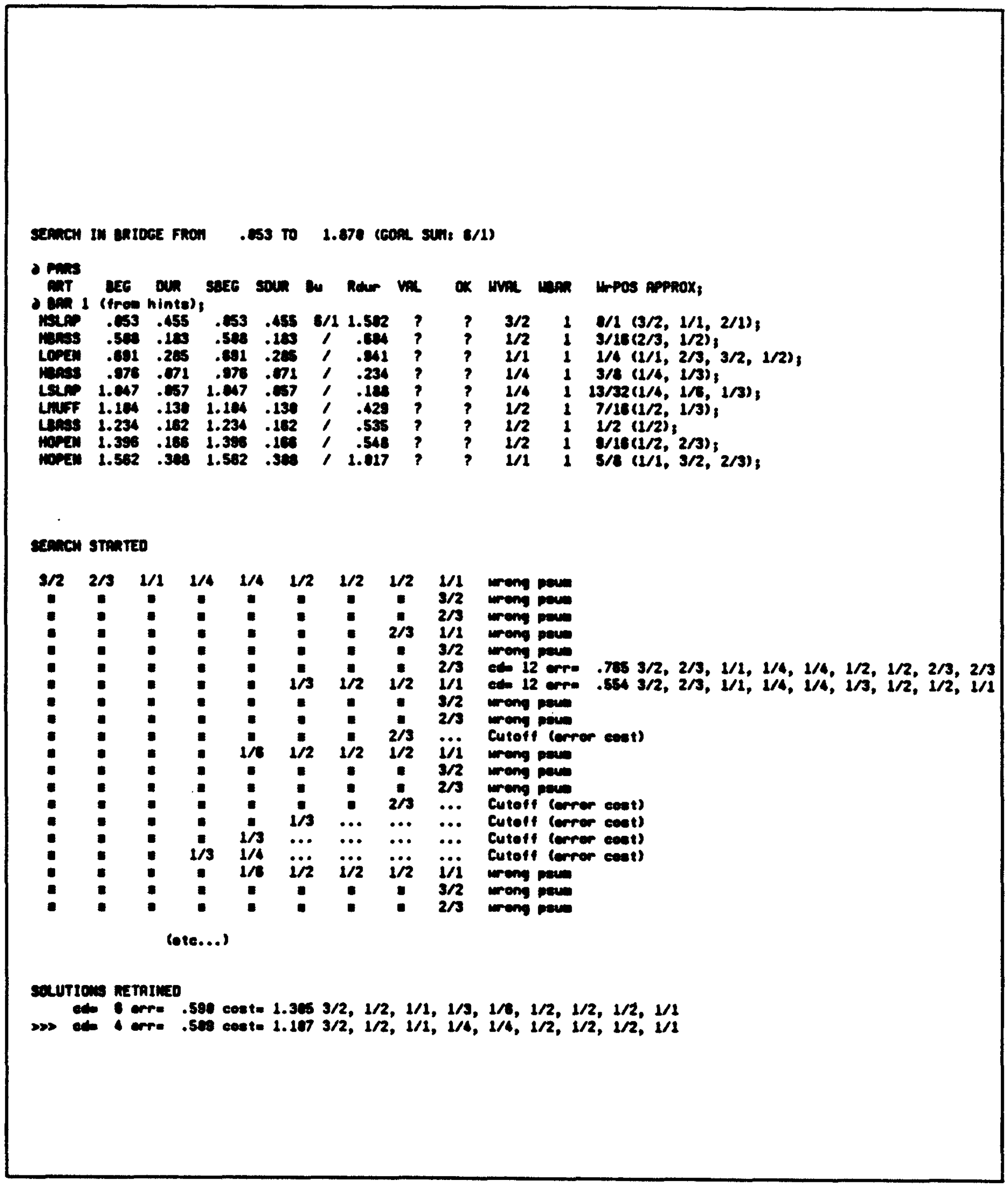
Figure 3.30. Note value statistics. Find the "popularity" of the possible base units, and use this to recalculate the rational approximations for durations.

of units in the respective bridge. Now, a combinatorial search through all the values in APPROX is needed to find the correct solutions.

Thus, the last step in finding the normalized metric values for the performed durations involves a recursive search of possible intra-bridge combinations using the refined cost measure as shown in Fig. 3.32. When this is finished, we see in Fig. 3.33 that the VAL field is completely filled in, and all values sum correctly.

```
  ART     BEG    DUR    SBEG   SDUR  Bu   Rdur  VAL    OK  WVAL  WBAR   WrPOS APPROX;
PLAY;
∂ BAR 1 (from hints);
 HSLAP    .053  .455    .053  .455  0/1 1.502   ?     ?    3/2    1   0/1 (3/2, 1/1, 2/1);
 HBASS    .508  .183    .508  .183   /   .604   ?     ?    1/2    1   3/16(2/3, 1/2);
 LOPEN    .691  .285    .691  .285   /   .941   ?     ?    1/1    1   1/4 (1/1, 2/3, 3/2, 1/2);
 HBASS    .976  .071    .976  .071   /   .234   ?     ?    1/4    1   3/8 (1/4, 1/3);
 LSLAP   1.047  .057   1.047  .057   /   .188   ?     ?    1/4    1  13/32(1/4, 1/6, 1/3);
 LMUFF   1.104  .130   1.104  .130   /   .429   ?     ?    1/2    1   7/16(1/2, 1/3);
 LBASS   1.234  .162   1.234  .162   /   .535   ?     ?    1/2    1   1/2 (1/2);
 HOPEN   1.396  .166   1.396  .166   /   .548   ?     ?    1/2    1   9/16(1/2, 2/3);
 HOPEN   1.562  .308   1.562  .308   /  1.017   ?     ?    1/1    1   5/8 (1/1, 3/2, 2/3);
∂ BAR 2 (from hints);
 LOPEN   1.870  .435   1.870  .435  2/1 1.452  3/2   1/1   3/2    2   0/1 (3/2, 1/1, 2/1);
 HOPEN   2.305  .164   2.305  .164   /   .548  1/2    /    1/2    2   3/16(1/2, 2/3);
 HOPEN   2.469  .300   2.469  .300  3/1  .984  1/1    /    1/1    2   1/4 (1/1, 2/3, 3/2, 1/2);
 LSLAP   2.769  .148   2.769  .154   /   .504  1/2    /    1/2    2   3/8 (1/2);
 LSLAP   2.917  .165   2.923  .154   /   .504  1/2    /    1/2    2   7/16(1/2);
 HOPEN   3.082  .147   3.077  .154   /   .504  1/2    /    1/2    2   1/2 (1/2);
 HOPEN   3.229  .155   3.230  .154   /   .504  1/2    /    1/2    2   9/16(1/2);
 LOPEN   3.384  .294   3.384  .294  0/1  .985   ?     ?    1/1    2   5/8 (1/1, 2/3, 3/2, 1/2);
∂ BAR 3 (from hints);
 HMUFF   3.678  .079   3.678  .079   /   .265   ?     ?    1/3    3   0/1 (1/4, 1/3, 1/2);
 HMUFF   3.757  .107   3.757  .107   /   .358   ?     ?    1/3    3   1/24(1/3, 1/2);
 LOPEN   3.864  .186   3.864  .186   /   .623   ?     ?    2/3    3   1/12(2/3, 1/2);
COMMENT MISSING              1/3 ;
 LOPEN   4.050  .213   4.050  .213   /   .714   ?     ?    2/3    3   1/6 (2/3, 1/1, 1/2);
COMMENT MISSING              1/3 ;
 HOPEN   4.263  .283   4.263  .283   /   .949   ?     ?    1/1    3   1/4 (1/1, 2/3, 3/2, 1/2);
 LSLAP   4.546  .444   4.546  .444   /  1.488   ?     ?    3/2    3   3/8 (3/2, 1/1, 2/1);
 HSLAP   4.990  .184   4.990  .184   /   .617   ?     ?    1/2    3   9/16(2/3, 1/2);
 LOPEN   5.174  .287   5.174  .287  0/1  .988  1/1   1/1   1/1    3   5/8 (1/1, 2/3, 3/2, 1/2);
∂ BAR 4 (from hints);
 HMUFF   5.461  .097   5.461  .097   /   .334  1/3    /    1/3    4   0/1 (1/3, 1/2);
 HOPEN   5.558  .185   5.558  .185   /   .637  2/3    /    2/3    4   1/24(2/3, 1/2, 1/1);
COMMENT MISSING              1/3 ;
 HOPEN   5.743  .102   5.743  .102   /   .351  1/3    /    1/3    4   1/8 (1/3, 1/2);
 HOPEN   5.845  .086   5.845  .086   /   .296  1/3    /    1/3    4   1/6 (1/3, 1/4, 1/2);
 HMUFF   5.931  .097   5.931  .097   /   .334  1/3    /    1/3    4   5/24(1/3, 1/2);
 LOPEN   6.028  .282   6.028  .282   /   .971  1/1    /    1/1    4   1/4 (1/1, 2/3, 3/2, 1/2);
 LSLAP   6.310  .449   6.310  .449   /  1.546  3/2    /    3/2    4   3/8 (3/2, 2/1, 1/1);
 HOPEN   6.759  .158   6.759  .158   /   .544  1/2    /    1/2    4   9/16(1/2);
 LOPEN   6.917  .310   6.917  .310  3/2 1.000  1/1    /    1/1    4   5/8 (1/1, 3/2, 2/3);
∂ BAR 5 (from hints);
 HMUFF   7.227  .151   7.227  .151   /   .481  1/2    /    1/2    5   0/1 (1/2);
 HSLAP   7.378  .287   7.378  .287  3/2  .948  1/1    /    1/1    5   1/16(1/1, 2/3, 3/2, 1/2);
 LSLAP   7.665  .167   7.665  .167   /   .552  1/2    /    1/2    5   3/16(1/2, 2/3);
 HOPEN   7.832  .292   7.832  .292  3/1  .976  1/1    /    1/1    5   1/4 (1/1, 2/3, 3/2, 1/2);
 LSLAP   8.124  .155   8.124  .152   /   .506  1/2    /    1/2    5   3/8 (1/2);
 HSLAP   8.279  .155   8.275  .152   /   .506  1/2    /    1/2    5   7/16(1/2);
 HMUFF   8.434  .150   8.427  .152   /   .506  1/2    /    1/2    5   1/2 (1/2);
 LOPEN   8.584  .146   8.578  .152   /   .506  1/2    /    1/2    5   9/16(1/2);
 LOPEN   8.730  .289   8.730  .289  3/2  .944  1/1    /    1/1    5   5/8 (1/1, 2/3, 3/2, 1/2);
∂ BAR 6 (from hints);
 HSLAP   9.019  .170   9.019  .170   /   .556  1/2    /    1/2    6   0/1 (1/2, 2/3);
 LSLAP   9.189  .290   9.189  .290  3/2  .965  1/1    /    1/1    6   1/16(1/1, 2/3, 3/2, 1/2);
 LSLAP   9.479  .161   9.479  .161   /   .535  1/2    /    1/2    6   3/16(1/2);
 HOPEN   9.640  .263   9.640  .263  0/1  .883   ?     ?    1/1    6   1/4 (1/1, 2/3, 1/2, 3/4);
 HMUFF   9.903  .073   9.903  .073   /   .248   ?     ?    1/3    6   3/8 (1/4, 1/3);
 HMUFF   9.976  .077   9.976  .077   /   .261   ?     ?    1/3    6   5/12(1/4, 1/3, 1/2);
 HOPEN  10.053  .089  10.053  .089   /   .302   ?     ?    1/3    6  11/24(1/3, 1/4, 1/2);
 HOPEN  10.142  .102  10.142  .102   /   .346   ?     ?    1/3    6   1/2 (1/3, 1/2);
 LOPEN  10.244  .095  10.244  .095   /   .323   ?     ?    1/3    6  13/24(1/3, 1/2);
 HOPEN  10.339  .116  10.339  .116   /   .394   ?     ?    1/3    6   7/12(1/3, 1/2, 3/8);
 HOPEN  10.455  .332  10.455  .332   /  1.127   ?     ?    1/1    6   5/8 (1/1, 2/3, 2/1);
∂ BAR 7 (from hints);
 LSLAP  10.787  .620  10.787  .620   /  2.105  2/1   1/1   2/1    7   0/1 (2/1);
FINISH;
```

Figure 3.31. Notelist with APPROX field recalculated based on new metric context.



```
SEARCH IN BRIDGE FROM    .053 TO   1.870 (GOAL SUM: 6/1)

∂ PARS
  ART    BEG   DUR   SBEG  SDUR  Bu   Rdur  VAL    OK  WVAL  WBAR   Wr-POS APPROX;
∂ BAR 1 (from hints);
 MSLAP   .053  .455   .053  .455  6/1 1.502  ?     ?    3/2    1   0/1 (3/2, 1/1, 2/1);
 HBASS   .508  .183   .508  .183   /   .604  ?     ?    1/2    1   3/16(2/3, 1/2);
 LOPEN   .691  .285   .691  .285   /   .941  ?     ?    1/1    1   1/4 (1/1, 2/3, 3/2, 1/2);
 HBASS   .976  .071   .976  .071   /   .234  ?     ?    1/4    1   3/8 (1/4, 1/3);
 LSLAP  1.047  .057  1.047  .057   /   .188  ?     ?    1/4    1  13/32(1/4, 1/6, 1/3);
 LMUFF  1.104  .130  1.104  .130   /   .429  ?     ?    1/2    1   7/16(1/2, 1/3);
 LBASS  1.234  .162  1.234  .162   /   .535  ?     ?    1/2    1   1/2 (1/2);
 HOPEN  1.396  .166  1.396  .166   /   .548  ?     ?    1/2    1   9/16(1/2, 2/3);
 HOPEN  1.562  .308  1.562  .308   /  1.017  ?     ?    1/1    1   5/8 (1/1, 3/2, 2/3);



SEARCH STARTED

 3/2   2/3   1/1   1/4   1/4   1/2   1/2   1/2   1/1   wrong psum
  ■     ■     ■     ■     ■     ■     ■     ■    3/2   wrong psum
  ■     ■     ■     ■     ■     ■     ■     ■    2/3   wrong psum
  ■     ■     ■     ■     ■     ■     ■    2/3   1/1   wrong psum
  ■     ■     ■     ■     ■     ■     ■     ■    3/2   wrong psum
  ■     ■     ■     ■     ■     ■     ■     ■    2/3   cd= 12 err=  .785 3/2, 2/3, 1/1, 1/4, 1/4, 1/2, 1/2, 2/3, 2/3
  ■     ■     ■     ■     ■    1/3   1/2   1/2   1/1   cd= 12 err=  .554 3/2, 2/3, 1/1, 1/4, 1/4, 1/3, 1/2, 1/2, 1/1
  ■     ■     ■     ■     ■     ■     ■     ■    3/2   wrong psum
  ■     ■     ■     ■     ■     ■     ■     ■    2/3   wrong psum
  ■     ■     ■     ■     ■     ■     ■    2/3   ...   Cutoff (error cost)
  ■     ■     ■     ■    1/6   1/2   1/2   1/2   1/1   wrong psum
  ■     ■     ■     ■     ■     ■     ■     ■    3/2   wrong psum
  ■     ■     ■     ■     ■     ■     ■     ■    2/3   wrong psum
  ■     ■     ■     ■     ■     ■     ■    2/3   ...   Cutoff (error cost)
  ■     ■     ■     ■     ■    1/3   ...   ...   ...   Cutoff (error cost)
  ■     ■     ■     ■    1/3   ...   ...   ...   ...   Cutoff (error cost)
  ■     ■     ■    1/3   1/4   ...   ...   ...   ...   Cutoff (error cost)
  ■     ■     ■     ■    1/6   1/2   1/2   1/2   1/1   wrong psum
  ■     ■     ■     ■     ■     ■     ■     ■    3/2   wrong psum
  ■     ■     ■     ■     ■     ■     ■     ■    2/3   wrong psum

                (etc...)

SOLUTIONS RETAINED
     cd=  6 err=  .590 cost= 1.305 3/2, 1/2, 1/1, 1/3, 1/6, 1/2, 1/2, 1/2, 1/1
>>>  cd=  4 err=  .508 cost= 1.107 3/2, 1/2, 1/1, 1/4, 1/4, 1/2, 1/2, 1/2, 1/1
```

Figure 3.32. Example of recursive search through rational approximations, using refined cost measure, for bar 1. Final solution is correct.

```
  ART    BEG   DUR   SBEG  SDUR  Bu   Rdur  VAL    OK  WVAL  WBAR   WrPOS APPROX;
PLAY;
∂ BAR 1 (from hints);
 HSLAP   .053  .455   .053  .455 0/1 1.502  3/2   1/1   3/2    1   0/1 (3/2, 1/1, 2/1);
 HBASS   .508  .183   .508  .183  /   .604  1/2    /    1/2    1   3/16(2/3, 1/2);
 LOPEN   .691  .285   .691  .285  /   .941  1/1    /    1/1    1   1/4 (1/1, 2/3, 3/2, 1/2);
 HBASS   .976  .071   .976  .071  /   .234  1/4    /    1/4    1   3/8 (1/4, 1/3);
 LSLAP  1.047  .057  1.047  .057  /   .188  1/4    /    1/4    1  13/32(1/4, 1/6, 1/3);
 LMUFF  1.104  .130  1.104  .130  /   .429  1/2    /    1/2    1   7/16(1/2, 1/3);
 LBASS  1.234  .162  1.234  .162  /   .535  1/2    /    1/2    1   1/2 (1/2);
 HOPEN  1.396  .166  1.396  .166  /   .548  1/2    /    1/2    1   9/16(1/2, 2/3);
 HOPEN  1.562  .308  1.562  .308  /  1.017  1/1    /    1/1    1   5/8 (1/1, 3/2, 2/3);
∂ BAR 2 (from hints);
 LOPEN  1.870  .435  1.870  .435 2/1 1.452  3/2    /    3/2    2   0/1 (3/2, 1/1, 2/1);
 HOPEN  2.305  .164  2.305  .164  /   .548  1/2    /    1/2    2   3/16(1/2, 2/3);
 HOPEN  2.469  .300  2.469  .300 3/1  .984  1/1    /    1/1    2   1/4 (1/1, 2/3, 3/2, 1/2);
 LSLAP  2.769  .148  2.769  .154  /   .504  1/2    /    1/2    2   3/8 (1/2);
 LSLAP  2.917  .105  2.923  .154  /   .504  1/2    /    1/2    2   7/16(1/2);
 HOPEN  3.022  .147  3.077  .154  /   .504  1/2    /    1/2    2   1/2 (1/2);
 HOPEN  3.229  .155  3.230  .154  /   .504  1/2    /    1/2    2   9/16(1/2);
 LOPEN  3.384  .294  3.384  .294 0/1  .985  1/1    /    1/1    2   5/8 (1/1, 2/3, 3/2, 1/2);
∂ BAR 3 (from hints);
 HMUFF  3.678  .079  3.678  .079  /   .265  1/2   3/2   1/3    3   0/1 (1/4, 1/3, 1/2);
 HMUFF  3.757  .107  3.757  .107  /   .359  1/2    /    1/3    3   1/24(1/3, 1/2);
 LOPEN  3.864  .186  3.864  .186  /   .623  1/2   3/4   2/3    3   1/12(2/3, 1/2);
COMMENT MISSING           1/3 ;
 LOPEN  4.050  .213  4.050  .213  /   .714  1/2    /    2/3    3   1/6 (2/3, 1/1, 1/2);
COMMENT MISSING           1/3 ;
 HOPEN  4.263  .283  4.263  .283  /   .949  1/1   1/1   1/1    3   1/4 (1/1, 2/3, 3/2, 1/2);
 LSLAP  4.546  .444  4.546  .444  /  1.488  3/2    /    3/2    3   3/8 (3/2, 1/1, 2/1);
 HSLAP  4.990  .184  4.990  .184  /   .617  1/2    /    1/2    3   9/16(2/3, 1/2);
 LOPEN  5.174  .287  5.174  .287 0/1  .988  1/1    /    1/1    3   5/8 (1/1, 2/3, 3/2, 1/2);
∂ BAR 4 (from hints);
 HMUFF  5.461  .097  5.461  .097  /   .334  1/3    /    1/3    4   0/1 (1/3, 1/2);
 HOPEN  5.558  .185  5.558  .185  /   .637  2/3    /    2/3    4   1/24(2/3, 1/2, 1/1);
COMMENT MISSING           1/3 ;
 HOPEN  5.743  .102  5.743  .102  /   .351  1/3    /    1/3    4   1/8 (1/3, 1/2);
 HOPEN  5.845  .086  5.845  .086  /   .296  1/3    /    1/3    4   1/6 (1/3, 1/4, 1/2);
 HMUFF  5.931  .097  5.931  .097  /   .334  1/3    /    1/3    4   5/24(1/3, 1/2);
 LOPEN  6.028  .282  6.028  .282  /   .971  1/1    /    1/1    4   1/4 (1/1, 2/3, 3/2, 1/2);
 LSLAP  6.310  .449  6.310  .449  /  1.546  3/2    /    3/2    4   3/8 (3/2, 2/1, 1/1);
 HOPEN  6.759  .158  6.759  .158  /   .544  1/2    /    1/2    4   9/16(1/2);
 LOPEN  6.917  .310  6.917  .310 3/2 1.068  1/1    /    1/1    4   5/8 (1/1, 3/2, 2/3);
∂ BAR 5 (from hints);
 HMUFF  7.227  .151  7.227  .151  /   .491  1/2    /    1/2    5   0/1 (1/2);
 HSLAP  7.378  .287  7.378  .287 3/2  .948  1/1    /    1/1    5   1/16(1/1, 2/3, 3/2, 1/2);
 LSLAP  7.665  .167  7.665  .167  /   .552  1/2    /    1/2    5   3/16(1/2, 2/3);
 HOPEN  7.832  .292  7.832  .292 3/1  .976  1/1    /    1/1    5   1/4 (1/1, 2/3, 3/2, 1/2);
 LSLAP  8.124  .155  8.124  .152  /   .506  1/2    /    1/2    5   3/8 (1/2);
 HSLAP  8.279  .155  8.275  .152  /   .506  1/2    /    1/2    5   7/16(1/2);
 HMUFF  8.434  .150  8.427  .152  /   .506  1/2    /    1/2    5   1/2 (1/2);
 LOPEN  8.584  .146  8.578  .152  /   .506  1/2    /    1/2    5   9/16(1/2);
 LOPEN  8.730  .289  8.730  .289 3/2  .944  1/1    /    1/1    5   5/8 (1/1, 2/3, 3/2, 1/2);
∂ BAR 6 (from hints);
 HSLAP  9.019  .170  9.019  .170  /   .558  1/2    /    1/2    6   0/1 (1/2, 2/3);
 LSLAP  9.189  .290  9.189  .290 3/2  .965  1/1    /    1/1    6   1/16(1/1, 2/3, 3/2, 1/2);
 LSLAP  9.479  .161  9.479  .161  /   .535  1/2    /    1/2    6   3/16(1/2);
 HOPEN  9.640  .263  9.640. .263 0/1  .883  1/1    /    1/1    6   1/4 (1/1, 2/3, 1/2, 3/4);
 HMUFF  9.903  .073  9.903  .073  /   .248  1/3    /    1/3    6   3/8 (1/4, 1/3);
 HMUFF  9.976  .077  9.976  .077  /   .261  1/3    /    1/3    6   5/12(1/4, 1/3, 1/2);
 HOPEN 10.053  .089 10.053  .089  /   .302  1/3    /    1/3    6  11/24(1/3, 1/4, 1/2);
 HOPEN 10.142  .102 10.142  .102  /   .346  1/3    /    1/3    6   1/2 (1/3, 1/2);
 LOPEN 10.244  .096 10.244  .095  /   .323  1/3    /    1/3    6  13/24(1/3, 1/2);
 HOPEN 10.339  .116 10.339  .116  /   .394  1/3    /    1/3    6   7/12(1/3, 1/2, 3/8);
 HOPEN 10.455  .332 10.455  .332  /  1.127  1/1    /    1/1    6   5/8 (1/1, 2/3, 2/1);
∂ BAR 7 (from hints);
 LSLAP 10.787  .620 10.787  .620  /  2.105  2/1    /    2/1    7   0/1 (2/1);
FINISH;
```

Figure 3.33. The final notelist, with value field complete. Now musical values have been given for all durations, so the notation can follow from here automatically. The score is shown in Fig. 3.34.

3.3.7. The Musical Map

It now remains to decide whether these values are correct. This is not a trivial question, because for improvised music there is no score to verify the program's conclusion. Careful listening and introspection (the player is listening to his own improvisation) are required, and one advantage here is that in this case the performer is also the verifier and author, and presumably knows what his musical intentions were. With all due caution, it is safe to say that these are indeed the "correct" values.

The example can now be notated, as the VAL field reduces the task to a simple translation. The metric hierarchy does not yet indicate the meter. Figure 3.34, which shows the end result of the program's efforts, shows a bit more than the program actually provided, namely the time signature, bar lines, and beaming. All else is provided by the program automatically. Note that the stroke-types are labeled beneath the note heads, with an X to indicate damped and a regular notehead to indicate undamped.

This example looks simple once it is written; though it is quite regular, many experienced listeners had trouble parsing it. The program performed quite well in finding all the correct metric values, and tracking the inevitable local fluctuations in the performance.

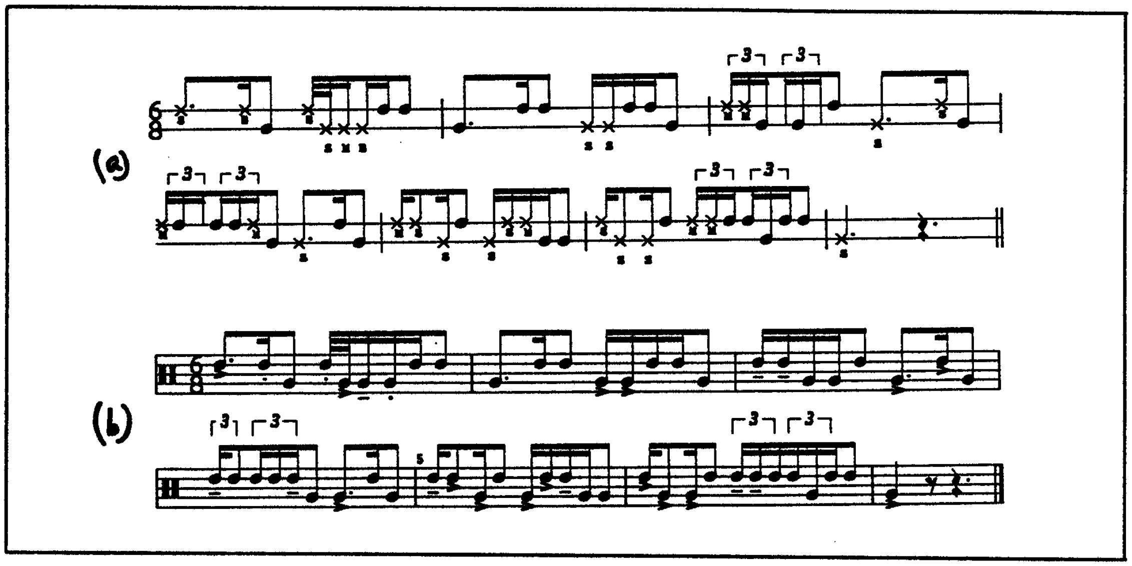


Figure 3.34. The score; the result of all the analysis described up to now.

a) A transcription done by ear. Extra information is coded in the following way:

High and low drums are notated on separate lines. Damped notes are notated with an x; undamped (open) with normal noteheads. Stroke-identification is denoted with italic letters beneath noteheads:

S—slap. *M*—muff. *B*—bass. Note that there are three note stems without noteheads. (Measure 3 has two and measure 4 has one). These are notated in this way because they were actually played, but not detected by the segmenter, and we need to know what the program did with them. (Refer to (b)).

b) A transcription done by the program. Note that this transcription is extremely close to the one done by ear. The strokes are coded with the following signs:

Accent—slap. *Underbar*—muff. *Dot*—bass. Note that the damped/undamped decision is implied by this notation—noteheads without marks are undamped. If we look at the places where the signal processing missed three attacks, the higher level gracefully notated the rhythm as it might be without these notes. (Measures 3 and 4). Segmentation omissions do not propogate beyond the beat they are in.

Chapter 4

Conclusions

"He who makes a mistake is still our friend;
He who adds to, or shortens, a melody is still our friend;
But he who violates a rhythm unawares
Can no longer be our friend."

—Ishaq ibn Ibrahim (767-850 A.D.)

4.1. Summary

In this dissertation, we discussed rhythm from a number of vantage points, from historical, theoretical, philosophical, pyschological, and musical. We have described an analysis system that expects a real performance as input, and generates a score automatically from the acoustic waveform. We have traced the system as it is applied to a short improvised drum example, and tried to verify that the result is correct.

It is not obvious what it means to prove a transcription is correct, if one is working from an improvised example, because if there is no score, there is no "answer in the back of the book." The best criterion for success is resynthesis at different levels, which is what is done in this thesis; most types of errors become quite obvious in the context of resynthesis. What this means is that in the context of this thesis, *the experiment and the transcription are completely linked.* Early attempts to do automatic transcription did not have the advantage of resynthesis capability. Although it is not the whole story, A.M. Jones said: "When a person can transcribe an African song or drumming *and is able to prove he is right,* he is well on the way to understanding the music." [Jones, 1959, p. 10]. He meant that verification of a transcription is not a trivial matter, especially if the original performer is not able or available to judge the results. Although transcription is a highly subjective matter, one wishes to impose some criterion for accuracy. The best method is resynthesis from the program's output.

The realm of percussion was chosen as a specific area out of a particular interest in rhythm, but the transcription system described is applicable in a more general musical context. The automatic segmentation and the tempo tracking methods are fairly general, and do not make many assumptions about the input.

4.2. Implications

There are two general areas to which this work contributes. The first is the basic problem of automatic transcription, some of which is solved herein (see Chapter 3). The other is in the more theoretical level of trying to define and characterize rhythm and meter. We have outlined rhythmic paradigms, and posited a special category for African and African-derived music.

One of the areas that has not been explored in previous automatic transcription projects is the problem of *automatic tempo tracking*. As described in Chapter 2, this problem was skirted in various ways by previous researchers, sometimes by just ignoring it, or typically, giving the system "hints" about meter and tempo.

In this transcription system, we track tempo automatically, and although the methods are not fully robust, they work in a large variety of examples. It is interesting to view the tempo-tracker as an "automatic foot-tapper," and in fact, it was surprising to observe that in test examples many listeners had trouble finding the meter and the beat, whereas the program was very close. The example used is nontrivial, and the meter is not obvious.

Usually, we would not expect the program to do better than a human listener; in fact, we should be suspicious when this happens. In the realm of Artificial Intelligence, we know that it is easy to do better than a person in particular formal tasks, but typically, perceptual skills are extremely difficult to emulate. When a particular machine skill has been developed and tailored primarily for appealing to a perceptual response, then "superior" performance by the machine certainly requires a critical view. It turns out that the program does better in this particular case, but there is a peculiar trade-off: the program needs to "hear" the whole piece before it knows how to "tap its foot." It does not process the data from left to right, whereas a listener always does. The listener, in the other hand, will start tapping fairly soon, and might applaud at the end, but certainly won't wait until the end of the piece to tap his or her foot! A real-time version of the program, that responds more as a person, would have to proceed from left to right.

4.3. Future Research

Though the system described herein covers new ground with respect to automatic transcription and analysis of rhythm, much terrain remains to be explored. Here, we will discuss briefly some of the intriguing directions possible.

4.3.1. Polyphony

The word "polyphony" is probably misleading here, because we do not mean polyphony as it is used in Western music, but rather a much more general concept of combining several simultaneous musical sources, voices, or layers, which will be found in almost any context. Dealing with polyphony is a challenge from the point of view of any automatic analysis system. Some of the methods, like the segmentation algorithm, can be applied to multiple-voice music, but will require a considerable amount of work to deal effectively with the problems introduced by polyphony. Another approach is to process each voice separately, which is possible if one has control over the recording process. It is so easy for people to distinguish different sources (for instance the "cocktail party effect"*) that one is motivated to try to duplicate this task at some level of competence.

In percussive music, one would like to be able to compare the different concurrent rhythmic components and examine exactly how they are related. This kind of work was done (by hand) by ethnomusicologists at UCLA (see [Koetting, 1970]), and it would be a significant advance to automate it.

Another aspect of the problem is identifying patterns in the music. This problem applies to single-voice music and is directly extendable to polyphonic music. Bernard Mont-Reynaud, working on the National Science Foundation grant at CCRMA described in Section 2.2.5, has developed several pattern detection schemes that are currently being tested on single-voice data, and will be used to help deal with polyphony and to determine meter by some kind of "minimum entropy" method.

---

* This refers to the well-known ability of people to comprehend a certain speaker in the midst of many other speakers, using one ear, or more effectively, both ears

4.3.2. Analysis of Style via Timing Studies

In Chapter 2, we reviewed numerous studies of timing in music. It is clear that timing is one of the most important aspects of performance practice in any musical tradition; for example, in piano music, it is the primary parameter.* The studies by Gabrielsson on rhythmic performance [Gabrielsson, 1974, 1980, 1983] point to many more experiments that can be done, and that would benefit a great deal from the analysis methods described herein (Section 3.2). The data would be both more accurate and more dependable, because more examples could be analyzed.

4.3.3. Synthesis of Percussive Sounds

It is clear that resynthesis is an important aspect of this work. Until recently, it was quite difficult to verify analysis results. Now, we can automatically generate aural instantiations of the analysis at any level. So far, we have been either resynthesizing with a plucked-string algorithm (see [Jaffe and Smith, 1983]), or reconstructing exact timings by concatenating small soundfiles (digitized individual drumstrokes) at the exact timings indicated by the analysis. It would be very desirable to have an adequate synthesis technique for percussive sounds. Most of the familiar ones are just not satisfying at the moment.

Subtractive synthesis methods that model the signal as a sum of damped sinusoids would seem to be promising for synthesizing drum sounds. Using Prony's method [Kay and Marple, 1981], we have tried to find pole-zero models for several drum tones, and up to now the results have not been very satisfying. It is likely that more effort to synthesize percussive tones using this method or others will be fruitful in the near future.

---

* The piano is the most obvious example of an instrument for which timing issues are the most important aspect of performance. That is, unlike most other instruments, the piano is not capable of any articulation and timbre control *within a single note*. (Pianists will likely disagree with this idea, but it is true.) What makes Richter sound different from Ashkenazy is attributable in large measure to complex time-intensity patterns. The result is that style can be approached via a careful study of timing issues. Obviously this won't reveal everything, but it is a crucial element.

4.3.4. Intelligent Editor of Musical Sound

As digital recording becomes more and more widespread and eventually becomes the standard recording medium, it will be more and more crucial to have sophisticated ***high-level*** editors of musical sound (see [Foster, et al., 1982] and [Chafe, et al., 1982]). It is now possible to edit digitized music down to the sample, which is delightfully accurate, but at 44.1 kHz per channel (sampling rate of the Sony PCM F1 digital tape recorder), the profusion of data is immense, and it is quite difficult to "find your way around" in the music. It is clear that what is needed is an ***intelligent*** editor of musical sound, an editor that knows about music, in much the same way that a text editor knows about written text (a structure editor). For instance, instead of saying "skip to the second word in the last paragraph," one could say "skip to the trumpet entrance after the fifty-second measure." Actually, the musical editor is a far more difficult task, because we wish to begin from the ***signal.*** It is analogous to a text editor that goes from the *spoken word*, in other words, a speech recognition program linked to a word processor, which is still a long way off.

The automatic transcription capability described in this thesis will be a good first step in the development of the intelligent editor. It is precisely these techniques of segmentation and tempo-tracking that, when generalized, will provide the first few levels of structure, e.g. beats, measures, sections, tempo. This is a vast subject which will undoubtedly be dealt with soon.

4.3.5. Interactive Performance

As mentioned earlier, one goal is to have a real-time version of the system. This is dependent as much on programming strategy as it is on having very fast processing. The idea of processing from "left to right" is very simple and appealing, but it is a major restriction that is not always possible to implement. Motivation is high for achieving a real-time version, and one of the most compelling reasons is the possibility for the composer of having an intelligent partner for *interactive* performance.

When a composer writes a piece for computer and instrument(s), it should no longer be necessary for the computer part to be prerecorded computer-generated music, which is very restrictive in performance because tape is a completely unexpressive, unresponsive medium incapable of adjusting to any tempo changes by the instrumentalist(s). There is little possibility of real ensemble (in the sense of interaction between the players) in any performance using tape, because performance is constrained by the tape part—tape is an absolutely "tyrannical" performer. The other performance choice, in which the players' sound is processed by electronic means, as was done in the electronic music of the 1960's, is too passive, offering little or no material that is not a direct result of what the players are doing.

Using the kind of musical knowledge available to the automatic transcription system, it will be possible for the computer to "listen" to the other players, and respond by making high-level *musical* decisions, predetermined at any level by the composer. Suddenly, the machine-performer is neither passive nor tyrannical, but becomes a viable performer in its own right, reflecting the intentions of the composer.

This aspect of computer music has barely been tapped to date. As computers become less expensive and more powerful, and more programs are written that expand on some of the issues raised in this thesis, we should see a new area of interactive musical performance emerge.

Appendix A. Contents of Tape of Musical Examples

A cassette with the following sound examples that accompany the text can be requested from the author by writing to:

Andrew Schloss
CCRMA/Music Department
Stanford University
Stanford, CA 94305

A. The Rhythmic Paradigm examples (Section 2.5).

1. Mozart: *Quintet in C, k* 515.
2. Stravinsky: *Les Noces* 1917.
3. Bulgaria: *Jove Malaj Mome*
4. Ghana: *Sogo* Dance (*Ewe* people).
5. Indian drumming: Alla Raka and Zakir Hussain (tabla).
6. Brazil: *Batucada.*

B. The roll examples (high-pass filtering).

1. Changing the cut-off frequency (see Figure 3.10).
2. Noise vs. phase discontinuity (see Figure 3.11).

C. The stroke types.

1. Individual strokes (see Figures 2.4–2.7).
2. Improvisation in free rhythm.
3. Resynthesis by concatenation of sound-files.
4. Unison w/synthesized guitar.

D. The transcription example.

1. Original performance (see Figure 3.34).
2. Resynthesis by concatenation of sound-files.
3. Unison w/synthesized guitar.
4. "Normalized" resynthesis from score.

References

Alette, Carl. *Theories of Rhythm*, Ph.D. Dissertation, Eastman School of Music, 1951.

Allan, Lorraine. "The Perception of Time," *Perception and Psychophysics* 26(5):340-354, 1979.

Askenfelt, A. "Automatic Notation of Played Music," *Speech Transmission Laboratory* 4, 1969.

Balzano, Gerald. "The Group-theoretic Description of 12-Fold and Microtonal Pitch Systems," *Computer Music Journal* 4(4):66-84, 1980.

Bartók, Béla and Albert B. Lord. *Serbo-Croation Folk Songs*, New York: Columbia University Press, 1951.

Bengtsson, Ingmar and Alf Gabrielsson. "Methods for Analyzing Performances of Musical Rhythm," *Scandanavian Journal of Psychology* 21:257-268, 1980.

Bengtsson, Ingmar and Alf Gabrielsson. "Analysis and Synthesis of Musical Rhythm," in *Studies of Music Performance*, Royal Swedish Academy of Music 39:27-59, 1983.

Chafe, C., B. Mont-Reynaud, and L. Rush. "Toward an Intelligent Editor of Digital Audio: Recognition of Musical Constructs," *Computer Music Journal* 6(1):30-41, 1982.

Chernoff, John Miller. *African Rhythm and African Sensibility,* Chicago: University of Chicago Press, 1979.

Cooper, Grosvenor and Leonard Meyer. *The Rhythmic Structure of Music*, Chicago: University of Chicago Press, 1960.

Creelman, C. "Human Discrimination of Auditory Duration," *Journal of the Acoustical Society of America* 34:582-593, 1962.

Diószegi, Vilmos. "Tuva Shamanism: Intraethnic Differences and Interethnic Analogies," *Acta Etnographica* 11:143-190, 1962.

Divenyi, Pierre. "The Rhythmic Perception of Micromelodies," *in E. Gordon, ed., Studies in the the Psychology of Music*, Vol. VII, Iowa City: University of Iowa Press, 1971.

Foster, S., W. A. Schloss, and A. J. Rockmore. "Toward an Intelligent Editor of Digital Audio: Signal Processing Methods," *Computer Music Journal* 6(1):42-51, 1982.

Fraisse, Paul. "Time and rhythm perception," *in E. C. Carterette and M. P. Friedman, eds., Handbook of Perception,* Vol. 8, New York: Academic Press, 1978.

Fraisse, Paul. "Rhythm and tempo," *in D. Deutsch, ed., The Psychology of Music,* New York: Academic Press, 1982.

Freedman, M.D. *A Technique for Analysis of Musical Instrument Tones,* Ph.D. Dissertation, University of Illinois, 1965.

Gabrielsson, Alf. "Similarity Ratings and Dimension Analyses of Auditory Rhythm Patterns," Parts I and II, *Scandanavian Journal of Psychology* 14: 138-176, 1973.

Gabrielsson, Alf. "Performance of Rhythmic Patterns," *Scandanavian Journal of Psychology* 15: 63-72, 1974.

Gabrielsson, Alf, "Experimental Research on Rhythm," *The Humanities Association Review* 30:69-92, 1979.

Gabrielsson, Alf, I. Bengtsson, and B. Gabrielsson. "Performance of Musical Rhythm in 3/4 and 6/8 Meter," *Scandanavian Journal of Psychology* 24: 193-213, 1983.

Getty, David J. "Discrimination of Short Temporal Intervals: A Comparison of Two Models," *Perception and Pyschophysics* 18(1):1-8, 1975.

Gordon, John *Perception of Attack Transcients in Musical Tones,* Ph.D. Dissertation, Department of Music, Stanford University, 1984.

Gordon, John and J. Strawn. "Introduction to the Phase Vocoder," *to appear in Roads, C. and J. Strawn, eds., Computer Music,* Cambridge: MIT Press, 1984.

Grey, John *An Exploration of Musical Timbre,* Ph.D. Dissertation, Stanford University, also Department of Music Report No. Stan-M-2 (1975)

Henderson. M.T. "Rhythmic Organization in Artistic Piano Performance," *in Carl E. Seashore, ed., University of Iowa Studies in the Psychology of Music,* Vol. IV. Iowa City: University of Iowa Press, 1936.

Henderson, M.T., J. Tiffin, C.E. Seashore. "The Iowa Piano Camera and its Use," *in Carl E. Seashore, ed., University of Iowa Studies in the Psychology of Music,* Vol. IV. Iowa City, University of Iowa Press, 1936.

Hirsh, Ira J. "Auditory Perception of Temporal Order," *Journal of the Acoustical Society of America* 31(6):759-767, June, 1959.

Hood, Mantle. *The Ethnomusicologist*, New York: McGraw-Hill, 1971.

Jaffe, David and J. O. Smith. "Extensions of the Karplus-Strong Plucked-String Algorithm," *Computer Music Journal* 7(2):56-69, 1983.

Jones, A. M. *Studies in African Music*, Vols. I and II. London: Oxford University Press, 1959.

Kay, Steven and Stanley Marple, Jr. "Spectrum Analysis—A Modern Perspective," Proceedings of the IEEE 69(11): 1404-1407, 1981.

Koetting, James. "Analysis and Notation of West African Drum Ensemble Music," Selected Reports, UCLA Institute of Ethnomusicology 1(3), 1970.

Kunst, Jaap. *Metre, Rhythm, Multi-Part Music,* Leiden: E.J. Brill, 1950.

Lindblom, B. and J. Sundberg. "Towards a Generative Theory of Melody," *Speech Transmission Laboratory* 4, 1969.

Longuet-Higgins, H.C. "The Perception of Melodies," *Nature* 263:646-653, 1976.

Longuet-Higgins, H.C. "The Perception of Music," *Interdisciplinary Science Reviews* 3(2):148-156, 1978.

Longuet-Higgins, H.C. and C.S. Lee. "The Perception of Musical Rhythms," Unpublished manuscript, 1978.

Luce, David A. *Physical Correlates of Nonpercussive Musical Instrument Tones*, Ph.D. Dissertation, MIT, 1963.

Lund, Max. *An Analysis of the "True Beat" in Music*, Ph.D Dissertation, Stanford University, 1938.

Lunney, H.W.M. "Time as Heard in Speech and Music," *Nature* 249:592, 1974.

Markel, J.D. and A.H. Gray. *Linear Prediction of Speech*, Springer Verlag, New York, 1976.

Michon, J.A. "Studies on Subjective Duration," *Acta Psychologica* 22:441-450, 1964.

Michon, J.A. *Timing in Temporal Tracking*, Soesterberg: Institute for Perception, 1967.

Miller, Ray E. "The Pitch Vibrato in Artistic Gliding Intonations," *in Seashore, Carl E., ed., University of Iowa Studies in the Psychology of Music*, Vol. I. Iowa City: University of Iowa Press, 1932.

Minton, Neil. Unpublished thesis, School of Music, Yale University.

Mont-Reynaud, Bernard. *Final Report, National Science Foundation Grant*, Grant No. MCS-7923282, 1984.

Moorer, James A. *On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer*, Ph.D. Dissertation, Stanford Universiy, 1975.

Moorer, James A. "The Use of the Phase Vocoder in Computer Music Applications," *Journal of the Audio Engineering Society* 26:42-45, 1978.

Morse, Philip M. and K. Uno Ingard. *Theoretical Acoustics*, New York: McGraw-Hill, 1968.

Needham, Rodney. "Percussion and Transition," *Man* 2:606-614, 1967.

Neher, Andrew. "A Physiological Explanation of Unusual Behavior in Ceremonies Involving Drums," *Human Biology* 34(2):151-161, 1962.

Oshinski, J.S. and S. Handel. "Syncopated Auditory Polyrhythms: Discontinuous Reversals in Meter Interpretation," *Journal of the Acoustical Society of America* 63(3):936-939, 1978.

Patterson, James H. and David M. Green. "Discrimination of Transient Signals Having Identical Energy Spectra," *Journal of the Acoustical Society of America* 48(4):894-905, 1970.

Pisczcalski, Martin. "Classifying Complex Sound Patterns as Simple Sequential Objects, Musical Notes," Unpublished document, University of Michigan Department of Computer Science.

Pisczcalski, Martin and B. Galler. "Automatic Music Transcription," *Computer Music Journal*, 1(4):24-31, 1977.

Pisczcalski, Martin and B. Galler. "Predicting Musical Pitch from Component Ratios," *Journal of the Acoustical Society of America* 66(3):710-720, 1979.

Pisczcalski, Martin, B. Galler, R. Bossemeyer, M. Hatamian, and F. Looft. "Performed Music: Analysis, Synthesis and Display by Computer," *Journal of the Audio Engineering Society* 29(1):38-46, 1981.

Pisczcalski, Martin and B. Galler. "A Computer Model of Music Recognition," *in M. Clynes, ed. Music, Mind and the Brain*, New York: Plenum Press, 1981.

Pressing, Jeff. "Cognitive Isomorphisms in Pitch and Rhythm in World Musics: West Africa, The Balkans, Thailand, and Western Tonality," Unpublished manuscript, 1979.

Riemann, Hugo. *Musikalische Dynamik und Agogik*, Hamburg, 1884.

Rossing, Thomas. "The Physics of Kettledrums," *Scientific American*, November, 1982.

Rossing, Thomas. "Acoustics of Percussion Instruments," *The Physics Teacher* 14(9):546-556, 1976, and 15(5):278-288, 1977.

Sachs, Curt. *Rhythm and Tempo*, New York: W.W. Norton and Co., 1953.

Salzer, Eric. *Structural Hearing*, New York: Dover, 1952.

Schroeder, Manfred R. "Parameter Estimation in Speech: a Lesson in Unorthodoxy," *Proceedings of the IEEE* 58(5), May 1970.

Seashore, Carl E., ed. *University of Iowa Studies in the Psychology of Music*, Volume I. *The Vibrato*, Volume III. Psychology of the Vibrato in Voice and Instrument, Volume IV. *Objective Analysis of Musical Perfomance*, Iowa City: University of Iowa Press, 1932, 1936, 1936.

Seeger, Charles. "An Instantaneous Music Notator," *Journal of the International Folk Music Council* III: 103-106, 1951.

Seeger, Charles. "Toward a Universal Music Sound-writing for Musicology," *Journal of the International Folk Music Council* IX: 63-66, 1957.

Seeger, Charles. "Prescriptive and Descriptive Music Writing," *Musical Quarterly*, XLIV(2): 185-195, 1958.

Shepard, Roger. "The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function," *Psychometrica* 27, 1962.

Stautner, John P. *The Auditory Transform*, MS thesis, MIT, 1982.

Steedman, Mark. "The Perception of Musical Rhythm and Metre," *Perception* 6: 555-569, 1977.

Sternberg, Saul, R.L. Knoll and P. Zukofsky. "Timing by Skilled Musicians," *in Deutsch, D., ed., The Psychology of Music*, New York: Academic Press, 1982.

Strawn, John. "Research on Timbre and Musical Contexts at CCMRA," *Proceedings of the 1982 International Computer Music Conference*. San Francisco: Computer Music Association, 1982.

Sundberg, Johan and B. Lindblom. "Generative Theories in Language and Music," *Cognition* 4:99-122, 1976.

Sundberg, Johan and Verrillo. "Anatomy of the Ritard," *Speech Transmission Laboratory* 2/3, 1977.

Tove, P.A., B. Norman, L. Isaksson and J. Czekajewski. "Direct-recording Frequency Ramp Meter for Analysis of Musical and Other Sonic Waveforms," *Journal of the Acoustical Society of America* 39:362-371, 1966.

Vernon, L.N. "Synchronization of Chords in Artistic Piano Music," *in Seashore, Carl E., ed., University of Iowa Studies in the Psychology of Music*, Vol. IV. Iowa City, University of Iowa Press, 1936.

von Békésy, Georg. *Experiments in Hearing*, New York: McGraw-Hill, 1960.

Vos, Joos and Rudolf Rasch. "Perceptual Onset of Musical Tones," *Perception and Psychophysics* 29(4):323-335, 1981.

Yeston, Maury. *The Stratification of Musical Rhythm*, New Haven: Yale University Press, 1976.