APPROXIMATING HARMONIC AMPLITUDE ENVELOPES OF MUSICAL INSTRUMENT SOUNDS WITH PRINCIPAL COMPONENT ANALYSIS

by

Robert G. Laughlin
B.Sc. Carleton University 1972

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
in the School
of
Computing Science

SIMON FRASER UNIVERSITY
October 1989

APPROVAL

Name: Robert G. Laughlin

Degree: Master of Science

Title of thesis: Approximating Harmonic Amplitude Envelopes of Musical Instrument Sounds with Principal Component Analysis

Examining Committee:

Dr. Robert F. Hadley
Chair

Dr. Brian Funt
Associate Professor, School of Computing Science
Simon Fraser University
Senior Supervisor

Dr. Barry Truax
Associate Professor, Dept. of Communication
Simon Fraser University
Supervisor

Dr. Keith Hamel
Assistant Professor, Dept. of Music
University of British Columbia
External Examiner

Date Approved: Oct. 3, 1989

PARTIAL COPYRIGHT LICENSE

Title of Thesis/Project/Extended Essay

Approximating Harmonic Amplitude Envelopes of Musical Instrument Sounds

with Principal Component Analysis.

Author: ______________________

(signature)

Robert G. Laughlin

(name)

October 4, 1989

(date)

Abstract

This thesis presents an alternate way to represent harmonic amplitude envelopes of musical instrument sounds using principal component analysis. Analysis reveals considerable correlation between the harmonic amplitude values at different time positions in the envelopes. This correlation is exploited in order to reduce the dimensionality of envelope specification. It was found that two or three parameters provide a reasonable approximation to the different harmonic envelope curves present in musical instrument sounds. Approximations to harmonic amplitude envelopes can be quickly reconstructed from a set of bases (common to all envelopes) and a set of scalar weights derived from the principal component analysis. The representation is suitable for the development of high level control mechanisms for manipulating the timbre of resynthesized harmonic sounds.

Acknowledgements

I would like to thank Brian Funt for allowing me to follow my interests and do research in an area that only partially overlapped with his own. Without exposure to various analysis and synthesis techniques common in computational vision this thesis topic would have been overlooked. I would also like to thank Barry Truax for keeping me on track with his considerable knowledge of sound synthesis and for introducing me to the fascinating discipline of psychoacoustics. I am also indebted to Jim Delgrande for encouraging me to apply for an NSERC scholarship and to NSERC for supporting me throughout this research.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Digital sound synthesis is a rapidly evolving technique useful in musical applications and psychoacoustic research. A sound generating algorithm creates waveforms (stored in computer memory as a series of numbers) which are output to a digital-to-analog converter for conversion to correspondingly varying voltages. These voltages are readily transformed to sound waves via common analog hardware (amplifiers and speakers).

One advantage of representing sound waveforms in a digital form is the flexibility inherent in the algorithmic manipulation of the numbers making up the waveform. Virtually any sound, including those produced by acoustic musical instruments such as guitars, pianos, wind instruments, etc., can be created and altered given an appropriate algorithm. Unfortunately many of the existing digital synthesis methods applicable to acoustic instrument sounds are limited in scope or too cumbersome and low level to be useful to sound designers. The problem is the large amount of information that is required to fully specify the sound.

This thesis will attempt to develop a method (and a rudimentary implementation of it) that addresses this problem. The focus will be on the data reduction of one particular aspect of the sound specification—the harmonic amplitude envelopes of instruments with a harmonic spectrum. The method may also be useful for research on timbre.

1.1 Sound and Music

Sound is characterized by physicists as a variation over space and time of the density of the medium transmitting the sound. Sound typically results from pressure variations in air emanating from a vibratory source as longitudinal waves. At a fixed point in space (for example the human eardrum) a record of the air pressure variation as a function of time completely characterizes the sound signal. The amount of deviation in air pressure from a stable equilibrium is related to the *intensity* of the sound. Although the sound signal is not necessarily periodic, the record of pressure variations over time is referred to as a waveform. Periodic or quasi-periodic variations in air pressure have a *frequency* (or frequencies) associated with them (the inverse of the period).

After processing by perceptual and cognitive mechanisms the sound signal yields a percept. *Loudness* is a property of this percept and is related logarithmically to the intensity of the sound. If the sound signal is periodic (or quasi-periodic) in nature, the percept may have a *pitch* associated with it. The relationship between pitch and frequency is complex. Pitch is often directly related to the fundamental frequency of musical sounds.

A third property, *timbre*, is commonly associated with the perception of some sounds. Musical instruments produce a sound that is heard as possessing a distinctive and recognizable timbre. Timbre is closely associated with the *spectrum* of a sound (the proportional mixture of frequencies making up the sound). Due to the physical nature of certain types of musical instruments, the frequencies making up their sound spectrum are (approximately) integer multiples of some fundamental frequency.

For example, with stringed instruments the fundamental frequency is directly related to the tension and mass of the string and to the length of string free to vibrate between two fixed ends. The string will also tend to vibrate in integral fractional lengths and corresponding frequency multiples (harmonics). String stiffness and inertia may result in slightly detuned harmonics. A soundboard connected to one end of the string provides a good impedance match for string vibrations and amplifies the vibratory string motion—which in turn causes air to vibrate. The air vibrations enter the ear, and the ear and brain together construct the conscious awareness of sound.

1.2 Sound Specification

A specification of the sound signal has two representations, a *time domain* representation and a *frequency domain* representation. The time domain representation characterizes the variation in air pressure over time. The frequency domain representation characterizes the harmonic frequency content of the sound—which may also vary over time. The two domains are related by the Fourier transformations.

One useful strategy for the digital synthesis of natural instruments is the *Analysis and Resynthesis* method [57]. Actual instrument sounds are analyzed and information is extracted that can be used to resynthesize the sounds[1]. Most musical instrument sounds have a harmonic frequency content that changes over time. This information can be specified with an *amplitude envelope* for each harmonic in the sound. The amplitude envelope specifies how the amplitude (intensity) of a particular harmonic changes over the duration of the sound. The changing spectrum of a sound—captured by the set of amplitude envelopes for the constituent harmonics—plays a significant role in the perception of timbre. The harmonic frequencies also tend to *fluctuate* slightly over time. Frequency fluctuations can also be extracted from an analysis of sound.

Sounds can be resynthesized using the harmonic amplitude envelopes, the harmonic frequency fluctuations, a fundamental frequency value, and an appropriate *sampling rate* (see equation 2.1 on page 16). Since analyzing a sound requires that it

[1] Digital resynthesis entails constructing the numerical representation of the time domain waveform—usually from frequency domain information.

already be in a digital format[2], a reasonable question might be: *Why bother analyzing a sound and then resynthesizing it when the sound waveform is already available?* The answer is that it is easier to manipulate a sound in the frequency domain, and the goal is to not only *replicate* natural sounds but to be able to *alter* them to produce desired effects. The timbre of a sound is one of the qualities that can be manipulated.

As with computational vision, it would be helpful to understand how physical information is processed by perceptual mechanisms—in this case the extraction of timbral percepts from variations in air pressure. This may provide clues as to reasonable strategies to employ in the manipulation of the signal parameters to produce the desired timbral effects. These issues will be discussed in Chapter 2, *Physical Correlates of Timbre.*

1.3 Data Reduction

In order to resynthesize sound with realistic (natural) timbres, a great deal of information is required. If we assume for the moment that each harmonic amplitude envelope could take the form of an arbitrary curve, then resynthesizing 3 seconds of sound with 40 harmonics—where the harmonic amplitude values are specified at 50 ms intervals—would require 2400 values for just the amplitude curves alone. This poses more of a *conceptual* problem than a computational problem. It is extremely difficult to alter the timbre of a sound (in a predictable fashion) when so much data specification is involved.

One of the existing methods of reducing envelope information is with line segment approximations (see page 18). While this method does reduce the amount of information involved, it does not entirely solve the problem of how to manipulate the envelope curve data in a meaningful way. Extracting line segments from analyzed data is also not a trivial problem (see page 19).

[2] Other forms of analysis are possible that do not require the sound to be digitized, however, they have been largely superceded by more powerful digital analysis techniques.

Empirical evidence from sound analysis (for both vowel speech and musical instrument sound, see page 29) indicates that the *spectra* of sound (the frequency content) can be considerably reduced in dimension by a *principal component analysis* of the class of sounds under consideration. The sounds analyzed in these studies are steady-state sounds (unchanging frequency content).

Since many of the interesting properties of musical sound result from the variation of the spectral content over time [56] and considerable work has already been done on characterizing musical sound in terms of harmonic amplitude envelopes [27, 48, 49, 50], the proposal in this thesis is to apply principal component data reduction to the harmonic amplitude envelopes. In addition to being reduced in dimension, the envelope specification will be *standardized* which may be useful for manipulating the envelopes to facilitate changes in timbre.

1.4 Chapter Contents

The following is a short summary of the contents of each chapter.

Chapter 1 Introduction

Chapter 1 contains a brief introduction to musical sound and digital synthesis. The general background and motivation for the research undertaken is presented.

Chapter 2 Physical Correlates of Timbre

Chapter 2 surveys the research literature on the psychoacoustics of timbre and its application to musical sound synthesis. Some of the existing strategies for the analysis and resynthesis of musical instrument sound are discussed.

Chapter 3 An Alternate Envelope Representation

Chapter 3 begins with a summary of the relevant research from chapter 2. The proposed representation of harmonic amplitude envelopes based on a principal component analysis of actual instrument envelopes is introduced. The chapter also contains an explanation of principal component data reduction using an intuitive geometric model.

Chapter 4 Data Collection and Analysis

Chapter 4 outlines the methods used to collect the data (harmonic amplitude envelopes of 280 musical sounds) and the mathematical manipulations involved in principal component analysis. The sounds included in the analysis are described in this chapter.

Chapter 5 Analysis Results and Interpretation

Chapter 5 discusses the results obtained from the principal component analysis of harmonic amplitude envelopes. The results are interpreted and discussed with respect to previous research and potential applications.

Chapter 6 Conclusions

Chapter 6 briefly summarizes the results from chapter 5 and suggests ways that the analysis could be improved. The chapter ends with suggestions for future research.

1.5 How to Read this Thesis

There are several ways to avoid reading the entire thesis. Readers familiar with research on timbre and sound synthesis could skip chapter 2. Chapter 3 contains a short summary of the relevant research and introduces the proposed envelope representation. Reading chapter 3 and a summary of the results in chapter 6 should be sufficient to provide a concise overview of the thesis.

Chapter 4 is an in-depth look at the analysis methods used (both the extraction of harmonic amplitude envelopes and the principal component analysis of them) and is not required reading for understanding the analysis results presented in chapter 5.

Chapter 5 should be read by those who are considering using the methods developed in this thesis. Appropriate references are given to other sections that explain envelope reconstruction (mostly in chapter 4). Chapter 5 also goes into more detail on some of the proposed uses for the method.

Information on principal component analysis is scattered throughout the thesis. Those unfamiliar with the method might want to read the geometric interpretation of principal component data reduction in chapter 3 before they read the *Principal Component Analysis of Spectra* section or the *Multidimensional Scaling* section in chapter 2. Chapter 4 outlines the mathematics of principal component analysis.

Liberal cross-referencing is used throughout the thesis (thanks to LaTeX) which should make it possible to read only the sections of interest.

Chapter 2

Physical Correlates of Timbre

Timbre is an aural attribute of sounds whose tone *quality* evokes—in some sense—a unified percept. Although timbral qualities could be attributed to any cohesive sound, in practice the term is usually applied to the sounds produced by the human voice and some musical instruments, particularly ones generating sounds that exhibit a harmonic spectrum.

A long-standing but vague definition of timbre is that it is the quality of a sound *not* accounted for by loudness or pitch. Hence two sounds of the same loudness and pitch may be discriminated on the basis of timbral differences, for example, the sounds produced by a trumpet and a trombone.

Conversely, sounds of different pitch or loudness—produced by the same instrument (or voice)—may be recognized as originating from the same source because of timbral *similarities*.

Identification of sounds by their timbre is surprisingly resistant to distortions in the sound signal (including spectral distortions) produced by environmental factors, for example, a saxophone sounds like a saxophone whether played in a reverberant concert hall or significantly distorted by a toy radio.

This raises several questions:

- Do we simply *learn* to identify sounds as originating from the same instrument using some sort of cognitive pattern matching?
- Have perceptual mechanisms evolved to extract (or construct) *invariant* features in the physical world to tie similar sounds together?
- If these physical invariants exist, what are they?

This chapter will look at some of the research pertinent to the last question.

The digital computer has made it possible to manipulate the *internal microstructure* of sound with the same freedom as more conventional musical structures (intervals, rhythm, melody, etc.). While constructing timbral sounds (digitally) involves manipulating a physical model of sound, the *effect* of that manipulation can only be assessed in the perceptual realm. Rather than relying on trial and error or heuristics to produce a given aural effect, it would be more useful to explore the relationship between the physical processes that produce sound and the perception of sound with the aim of understanding the underlying mechanisms that relate them.

2.1 Psychophysics

Psychophysics is the study of the effect of physical processes on the mental processes of an organism (psychoacoustics is the subset of psychophysics restricted to the sense of hearing). The attempt to establish quantitative psychophysical relationships is one of the oldest branches of psychology [8]. To illustrate some of the inherent problems, consider the relatively simple relationships involved in *pitch perception*. The same problems will be apparent in the psychophysics of timbre.

2.1.1 Simple Tone Pitch Perception

It is easy to demonstrate that there is a monotonic relationship between the frequency of a simple (pure sine wave) tone and the pitch it evokes. As the frequency increases or decreases the perceived pitch does the same (or at least stays the same if the frequency change is small). Beyond this monotonic relationship between frequency and pitch—and a corresponding ordinal relationship between pitches—further quantification of the link between the physical and psychological phenomena involved in pitch perception is questionable[1].

2.1.2 Physiological Mechanisms

If the intervening physiology of the pitch detection mechanisms in the inner ear are considered, the psychophysical relationship between simple tone frequency and pitch can be better understood.

Pitch is detected by unique resonance regions on the basilar membrane (of the cochlea) that correspond to the frequency components of the complex vibrations transmitted to the cochlear fluid—in the form of a standing wave—by the eardrum and the middle ear bone chain [13, 59]. The basilar membrane resonances in turn activate hair cells connected neurally with the brain.

The relative positions of the resonance regions on the basilar membrane bear a roughly logarithmic relationship to the frequencies of the vibrations enervating those regions. A simple *Place Theory of Hearing* [59] has been proposed to account for the pitch perception of simple tones[2].

2.1.3 Complex Tone Pitch Perception

When a complex harmonic sound is heard, *many* resonance regions are activated on the basilar membrane—one region for each harmonic component of the sound. Due to the logarithmic relationship between frequency and resonance position, and the

[1] Attempts to quantify the relationship have been made, for example, the relationship between the *mel* pitch scale and frequency [72]. The relationship is questionable because of the dubiousness of the mel scale.

[2] The Place Theory of Hearing cannot account for the degree of accuracy (e.g. jnd) observed in human pitch perception [72]. It does however play an important role in the perception of timbre.

nature of the harmonic series (geometric), an invariant spatial *pattern* is activated on the basilar membrane for all harmonic sounds.

The logarithmic scale of resonance positions results in higher harmonics being grouped into progressively smaller areas on the basilar membrane. Since the membrane has limited spatial resolution, frequencies falling within the same resolution region cannot be resolved as distinct pitch (or timbral) components[3]. Consequently frequencies can be lumped together into groups corresponding to the *critical band* frequency resolution of the ear[4] [59].

The intriguing aspect of pitch perception of harmonic sounds is that a *unified percept* (one and only one pitch) is associated with each sound. While it is possible, with practice, to "hear out" frequency components of a harmonic sound, the default percept is a pitch at the *periodicity pitch* of the related harmonic components (periodicity pitch corresponds to the *lowest possible frequency* that could have the sound's component frequencies as harmonics [59, 65]). As a result, a complete set of harmonics—or even a fundamental—is not required for consistent pitch perception of harmonic sounds. Pitch is *invariant* under a wide variety of conditions and stimuli.

The physiological mechanisms involved in the pitch perception of complex tones are not as well understood as for simple tones. It seems likely that higher neural processes come into play to recognize spatial resonance patterns on the basilar membrane. Time domain signal information is also transmitted to the brain via neural pulse rates and may also contribute to periodicity pitch detection through time distribution analysis [59, sections 2.9 and 4.8].

[3] Searle [66] notes that the frequency analyzing power of the ear is cleverly arranged to give good frequency resolution in low frequency (slowly-varying formant) regions and good *temporal* resolution at high frequencies (short wide-band consonant bursts). Good temporal resolution in the high frequency region of the basilar membrane follows from the ear's poor high frequency resolution in the same region—since time and frequency resolution are inversely proportional to each other, even for the ear.

[4] Grouping of frequencies (or harmonics) falling within a critical band may be acceptable for pitch detection analysis but not necessarily acceptable for timbre judgements since different arrangements of closely spaced harmonics can alter the "texture" of a sound.

2.1.4 Important Issues in Psychophysics

The relatively simple psychophysical relation between signal frequencies and pitch perception illustrates some important issues that provide a useful theoretical footing to examine the more complex phenomena involved in timbre perception.

Inter and Intra Relationships

In establishing psychophysical relationships, the principal relationship is between the physical and mental phenomena. In other words the *bridge* between the two domains is the focus of interest. Relationships also exist *within* each domain. For example, with pitch detection a simple ordinal relationship exists between simple as well as complex sounds. The intra relationships between harmonic sounds in the physical domain are more complex.

Dimensionality

Pitch can be classified with respect to one dimension—the "tone height" of the pitch. Dimensionality usually applies to intra relationships. As we shall see, timbre perception can better be captured with multiple dimensions[5].

Invariance

A many-to-one mapping is common for psychophysical relationships. Perception seems to rely heavily on invariants—either by *constructing* them from disparate sensory input, or by tuning into existing invariants in the physical world. James J. Gibson [22] is the major proponent of the latter position.

[5] The other principal attribute of sound—loudness—is also (to a first approximation) a one dimensional phenomenon. The loudness of a sound is correlated with the total sonic energy arriving at the ear. Note that these one dimensional relationships only hold if all other factors are kept constant, for example, the loudness of a sound also depends on its frequency content.

For pitch detection of complex harmonic sounds, the invariant would be the periodicity "pitch" of the frequency components. Examples of invariants in visual perception are *color constancy* (the ability to separate the color of an illuminant from the intrinsic color of an illuminated object—the signal reaching the eye is a synthesis of the two phenomena) and the highly developed skill of *facial recognition* despite radical changes due to age, cosmetic changes, novel perspectives, etc.

Data Reduction and Redundancy

In many cases much of the information associated with physical events does not appear to be necessary to form an unambiguous percept. For example, many of the harmonic components of a complex sound may be unnecessary to establish the periodicity pitch (they may however be required for timbre perception). Data reduction is made possible by redundancy in the physical signal. Redundant information can often serve as backup for unusual or error prone situations.

Perceptual Fusion

Many different physical components may be involved in the production of a single percept. For example, the harmonic components of a complex sound fuse into one unified pitch percept.

Given the efficacy of evolutionary theory, it seems reasonable that our perceptual apparatus evolved in such a way as to focus on essential environmental information and to recognize similar situations across widely varying conditions. The most efficient way to lighten the perceptual workload would be some form of *synthesis* of relevant sensory input. Perceptual fusion is closely related to dimensionality, invariance, redundancy and data reduction.

Physiological Link

Understanding the physiological link can clarify the psychophysical relationship. In the case of simple tone pitch perception, the physiological mechanisms are well understood and provide a more quantitative basis for the relationship they mediate.

Quantification Problem

Even with the relatively straightforward relationship between frequency and pitch and the known logarithmic physiological connection, it is hardly reasonable to posit a log function, mapping frequencies to pitch[6]. For example it makes little sense to say that one pitch is $1\frac{1}{2}$ times greater than another one.

Another problem is that psychophysical relationships are often confounded with other psychophysical phenomena. For example, perceived pitch can be altered to some degree by the amplitude and the duration of a sound [59]. Psychophysical relationships are also susceptible to widely varying individual differences and the human ability to alter perceptual skills through learning.

2.2 Two Approaches to Studying Timbre

Since psychophysics bridges two domains, the physical and the mental, it is not surprising that timbre research has been approached from two different perspectives and disciplines—physics and psychology.

One approach begins with a model of sound and by manipulating it attempts to find features of the physical model that are relevant to timbre perception. Much of the work here is focused on creating adequate sound models that capture—with as few parameters as possible—a wide variety of timbres. This approach will be referred to as *Manipulating a Sound Model* (see page 15).

[6] In fact it has been suggested by Schoenberg (quoted in [17]) that pitch perception be subsumed under the overriding influence of the overtone series.

The second approach focuses on timbre as a multidimensional percept and attempts to partition it into a meaningful set of orthogonal components—and optionally look for physical correlates of these components. This approach will be referred to as *Categorizing Timbral Percepts* (see page 32).

Clearly the two approaches complement each other. In many cases effort is prudently divided between them.

2.3 Manipulating A Sound Model

2.3.1 Additive Model of Harmonic Sound

Timbre perception has been associated with the *spectrum* of a sound[7] since the pioneering research of von Helmholtz [75]. Powerful digital analysis tools developed in the last 30 years [46] reveal that the spectrum of a harmonic sound can change considerably over the duration of the sound. Analysis also reveals that harmonic frequencies fluctuate somewhat over time [56] and that inharmonic components, noise for example, are present in natural harmonic sounds [48, 49, 50].

Digital techniques can also be used to *resynthesize* musical sounds, based on information extracted from an analysis. A *mathematical model* of harmonic sound is constructed from a theoretical knowledge of sound—taking into consideration empirical information gained from analysis. Resynthesis allows the efficacy of the physical model to be assessed by comparing the resynthesized sounds to the originals on which the analyses were based. A physical model also allows information extracted from an analysis to be parameterized and manipulated before being resynthesized.

This methodology, referred to as the *Analysis Resynthesis Method* [57], is a powerful tool for studying timbre. The mathematical model most commonly used with it

[7] In general a time domain representation of sound is not suitable as a physical model in timbre research since it is very complex (the most effective way to reduce the complexity is to convert the signal to the frequency domain). Waveforms are also not suitable for exploring the physical correlates of timbre since they are greatly altered by changes in phase, while timbre perception is not [9, 41, 53]. Synthesis methods that manipulate waveforms *directly* are usually assessed in the frequency domain.

is the *additive synthesis* model[8].

Digital sound synthesis requires the construction of a *sampled time domain* representation of the sound. This numerical representation is converted to analog voltages (via digital to analog converters) and then to air pressure waves by amplifying the signal and sending it to appropriate transducers (speakers). The time domain representation can be constructed from frequency domain parameters using the following variant [27, 46] of the additive synthesis model[9],

$$X(n, \Delta t) = \sum_{k=1}^{M} A_k(n, \Delta t) \sin\{2\pi n \Delta t[k\omega + F_k(n, \Delta t)]\} \tag{2.1}$$

Where,

- n is the sample number ($0 \leq n < N$), N is the total number of samples.
- Δt is the time in seconds between consecutive samples and $n\Delta t$ is the time of occurrence of the nth sample.
- X is a function of n and Δt and is the time domain signal at time $n\Delta t$.
- k is the harmonic number ($1 \leq k \leq M$).
- M is the number of harmonics.
- ω is the fundamental frequency of the sound in Hz.

[8] While a variety of methods exist for the digital synthesis of sound [16], most are not general enough to serve as a *model* of harmonic sound. Additive synthesis serves both as a synthesis technique and as a model general and flexible enough to represent many varieties of harmonic sound. Additive synthesis has an additional advantage when studying the psychophysics of timbre since the ear is also capable (in a limited fashion) of analyzing time domain sound signals into (crude) frequency components. Models of musical sound have been proposed that are based on the underlying physics of instrument behaviour [29, 30]. These models are powerful synthesis tools (given a set of initial conditions) and (potentially at least) very general, but they are extremely complex and not particularly suited to timbre research. Risset and Wessel [57] classify these models as *acoustic models*, to distinguish them from *perceptual models* such as additive synthesis.

[9] Several factors are not accounted for in this model. These include the effect of onset transients, noise, and other inharmonic components.

- A_k is a function of n and Δt and is the amplitude of harmonic k at time $n\Delta t$. A_k is assumed to vary slowly with time.
- F_k is a function of n and Δt and is the frequency *deviation* (in Hz) of harmonic k—at time $n\Delta t$—from the theoretical harmonic frequency $k\omega$. F_k is assumed to vary slowly with time.

A time domain signal can be constructed with equation 2.1 (for each value of n) by choosing a sampling period[10] Δt, a fundamental frequency ω, harmonic amplitude envelope curves A_k, and harmonic frequency fluctuation curves F_k. These curves can be extracted from a spectral analysis of an actual sound or constructed algorithmically.

The model bears a superficial resemblance to Fourier series synthesis except that:

- The synthesis is bandlimited (the frequency response of the ear is also bandlimited).
- Phase considerations are ignored.
- The Fourier series sine coefficients are replaced with slowly time-varying functions.
- The harmonic frequencies are allowed to vary slowly with time.

Phase information is ignored since its effect on timbre perception is relatively small [9, 41, 53] and is somewhat obviated given the fluctuating harmonic frequencies. Risset and Wessel [57] point out that this insensitivity to phase makes sense from an evolutionary perspective since phases are significantly distorted in a reverberant environment.

The slowly varying behaviour of A_k and F_k has been determined empirically [20, 48, 49, 50, 56]. Note however that A_k and F_k tend to exhibit small but rapidly varying fluctuations. Risset and Mathews [56] characterize these variations in F_k as "quasi-random".

[10] The sampling period is the inverse of the sampling rate, for example, a sampling period of 25 microseconds corresponds to a sampling rate of 40,000 samples per second.

The *magnitude* of the harmonic frequency variations also tends to be small—except for effects such as vibrato, deliberate frequency glides, or intonation problems at the onset of wind and brass instrument sounds[11].

2.3.2 Line Segment Envelope Approximations

The mathematical model of harmonic sound defined by equation 2.1 (page 16) is a simplification of physical sound in that it ignores subtle inharmonic components. However, if the amplitude envelope functions A_k and frequency fluctuation functions F_k are unconstrained for *each* sample[12], then no further simplification or *data reduction* would result from using this model. In other words to specify the sound would require $O(n)$ values of A_k and F_k—the same complexity required to specify the time domain signal X in equation 2.1 (page 16).

The fact that A_k varies *slowly* with time can be used to reduce the information required to specify the sound. For example, A_k can be approximated with a set of straight line segments since changes in A_k are gradual and the first derivatives (slopes) are relatively constant over some regions of the sound. A straight line produces a reasonable approximation to the amplitude envelopes over these regions.

Risset and Mathews were the first to propose this method of data reduction in a study of trumpet tones [56]. Their study and subsequent ones [27, 48, 49, 50] indicate that line segment approximations to harmonic amplitude envelopes result in resynthesized sounds that in many cases are indistinguishable from the sounds resynthesized with the original (fully specified) amplitude envelopes.

[11] The transients produced at the onset of a sound tend to defy the assumptions of slowly-varying functions and small frequency fluctuations. However, these effects usually occur only within the first 10–100 ms of a sound's onset [40].

[12] This is not strictly possible when A_k is derived from an actual analysis since frequency resolution is inversely proportional to time resolution [66]. In order to obtain a frequency resolution Δf that is adequate to separate harmonics and ascertain their amplitudes, T discrete time domain signals are required, where $T = \lceil \frac{1}{\Delta f \Delta t} \rceil$. The values of A_k extracted would then represent the frequency information for harmonic k averaged over T samples. Of course values for A_k can be *interpolated* for each Δt interval in equation 2.1 (page 16).

Mixed results have been reported when the *original* digitized sounds are compared with either variety of resynthesized sound. Risset and Mathews [56] claim they were indistinguishable but Grey and Moorer [27] found that original tones could be discriminated to some extent from the resynthesized versions[13].

The success of line segment approximations tends to eliminate the *microstructure* of amplitude envelopes from consideration as a perceptually relevant physical correlate of timbre[14] (this may not be true for the onset transient portion of a sound). In addition, envelope simplification affords the researcher a clearer picture of the physical stimuli taking part in the psychophysical relationship. A simpler physical model should make it easier to discover salient features to tie to perceptual phenomena.

Problems with Line Segment Extraction

One difficulty with the line segment approach is the methods available to extract them from the original curves. In some cases [25, 27, 50, 56] the line segments were extracted manually (the optimal number of breakpoints required depends on the shapes of the original curves). This method produces non-standardized results in that it is not always clear what heuristics are being used in placing breakpoints. Deciding how to cope with local non-monotonic variations in the envelopes (for example the "blips" that occur for the onset portion of brass instrument sounds [40, 50, 70, 71]) is a particularly difficult problem. The method is also very tedious if a large number of samples are being approximated.

Automating Line Segment Extraction

Strawn [69] suggests several methods of automating this process using approximation theory and pattern recognition algorithms. One method uses a predetermined number

[13] This was attributed to various noise and random factors present in the original sounds. Grey and Moorer also used more than one type of instrument while Risset and Mathews based their conclusions solely on the trumpet.

[14] The microstructure referred to includes both small rapidly varying amplitude modulations as well as the larger scale "curve straightening" introduced by the line segment approximations.

of breakpoints and error minimization techniques to fit first-order splines (straight line segments) to the curve data. Another method allows the specification of an error threshold and has the algorithm select the number of line segments required to meet these conditions.

Strawn states that it has proven impossible to find a suitable paradigm for choosing a threshold that preserves salient perceptual features while eliminating extraneous features. He concludes that no single algorithm of this type will be sufficient for systematically exploring timbre and data reduction, primarily because global and local considerations are not both handled adequately.

Strawn also experimented with hierarchical syntactic analysis techniques from Artificial Intelligence research. The basic idea is that envelope curves would be parsed by a suitable grammar terminating in a set of primitive features and resulting in a tree-like data structure. This method appears better suited to the extraction of features at different levels.

2.3.3 Interharmonic Relationships

One problem with the general additive model is the potentially unconstrained *independence* of the harmonic components. This problem remains even when the individual harmonic amplitude envelopes are simplified. Attempts have been made to find dependent relationships among harmonic components in order to obtain a more global perspective on the resulting sound.

Formants versus Fixed Spectral Shape

A longstanding issue in timbre research is the role played by *formants* in identifiying sounds over widely varying conditions. Formant structures are fixed-frequency amplification patterns induced by passive sound-producing mechanisms. Natural resonances of the air cavity in the vocal tract [14] or in musical instrument bodies [2, 5]—as well as other transducer resonances (such as the natural modes of vibration of stringed instrument soundboards [31])—tend to reinforce frequency components in *fixed* regions of the spectrum. Typically there are several formant regions and each region is distributed over some range of frequencies, with a pronounced peak at a particular frequency.

Our ability to discern speech vowels (essentially based on timbral distinctions) is largely due to the *varying* formant structures resulting from changes to the size and shape of the vocal tract air cavity by positioning the tongue, lips, and teeth [14]. Formant structures for musical instruments are less variable and could conceivably be the invariant physical correlate that makes it possible to identify instrument sounds under different conditions.

An alternate theory is that the *shape* of an instrument's sound spectrum is relatively invariant. For example, according to this theory an instrument might be identified because the odd harmonics are pronounced with respect to the even harmonics, irrespective of the fundamental frequency of the note played. The formant theory would claim, on the other hand, that the pronounced harmonics would be the ones that happen to fall into the fixed resonance regions of the instrument. As the fundamental frequency changes, the pattern of pronounced harmonics would also change[15]. These invariant resonance regions would then be the aural clues to instrument identification.

Slawson [67] argues that the primary stimuli for timbre research should be the sounds produced by the human voice—reasoning that whatever timbre perception mechanisms are in place probably evolved to handle sounds produced by the human vocal tract. Subjects in Slawson's study were asked to assess either the differences in the vowel quality or the differences in musical timbre between two synthesized stimuli (the same stimuli were presented to both groups).

The stimulus pairs were constructed and varied so as to determine which of the two theories—the formant theory or the fixed spectral shape theory—would be supported. The formant theory was clearly superior.

[15] It is important to keep in mind that the harmonic amplitudes are also changing over time. Hence the interharmonic relationships will also change to some extent over the duration of the sound.

Small changes in the peak frequencies of the two lowest formants also resulted in large differences in timbre. Slawson speculated that timbre invariance is more likely to be the result of built-in genetic mechanisms than learned responses, and that musical timbre and vowel quality are based on the same set of physical correlates.

The stimuli in Slawson's study were sounds produced by an analog speech synthesizer modeled on vocal tract mechanisms (a pulse train filtered to emulate the vocal tract air cavity [14]). While Slawson's conclusions may be valid for these types of stimuli, it is not necessarily the case that all instrument sounds are a proper subset of vowel-like sounds.

In fact evidence from other studies indicates that for many harmonic instruments the formant theory does not appear valid—or only valid to a limited degree[16]. The results are somewhat contradictory. In some studies (of a variety of brass and woodwind instruments) evidence of formant structures was found [36, 56, 70, 71]. Other studies on some of the same instruments [40, 50] found little or no evidence of formant structures. A study by Saldanha and Corso [61] was inconclusive. It does appears that some instruments exhibit more of a formant structure than others (oboe [49], bassoon [36]).

Because of the inconclusive (and often contradictory) results, a formant analysis of musical instruments may not be as useful as it is for speech vowels. The fixed spectral shape theory is also questionable since the spectrum of an instrument's sound (and its temporal evolution) can be altered significantly by both the note played (register) and player controlled dynamics (intensity) [36, 39, 40, 56, 70, 71].

Spectrum and Intensity Variations

Luce [39] has systematically studied the effect of intensity variations on the spectrum of musical instrument sounds. In general, it appears that high frequency content increases with intensity, however, the amount of increase is instrument specific and

[16] Formant structures are a part of any resonant chamber, however, their effect on the resulting sound may be inconsequential because of other more powerful influences.

for some instruments the relationship does not hold.

Nonlinear Relationships

Several methods have been developed that overcome the numerous degrees of freedom inherent in the additive model. Risset, in follow-up work to [56] (discussed in [57]), related the amplitude envelopes of all upper harmonics to the envelope of the fundamental. The proposed interharmonic relationships are nonlinear and specific to the brass instrument family—and closely tied to the physics of sound production in these instruments (see also Beauchamp [3]).

In more recent work, Beauchamp [4] used the spectral "centre of gravity" (the midpoint of the spectral energy distribution)—as it changes over the duration of a sound—as a parameter for resynthesis. The spectral centre of gravity is highly correlated with the "brightness" or "sharpness" of a sound [26, 73, 74]. Beauchamp's synthesis technique is compatible with efficient synthesis methods (such as nonlinear/filter synthesis or FM) and the results can be compared mathematically with additive analysis data.

Envelope Data Reduction

Charbonneau [10] has studied the perceptual effects of various forms of data reduction on additive parameters (see also the section on *Frequency Fluctuations* on page 27). Amplitude envelopes were simplified by averaging *all* the harmonic amplitude envelopes (extracted from a time-varying spectral analysis of a given sound) and reconstructing the sound from four parameters:

- The amplitude mean curve (normalized to a peak amplitude of 1).
- The starting time of each harmonic.
- The ending time of each harmonic.
- The maximum amplitude of each harmonic.

Each harmonic curve was reconstructed by weighting the mean curve with the maximum amplitude for that harmonic and shifting and scaling the weighted curve's time duration in accordance with the starting and ending times of the harmonic (envelope reconstruction is nonlinear). The harmonics were then combined in the normal additive fashion.

The success of the resynthesized sounds (when compared to the originals) varied considerably over the range of instrument sounds analyzed. In many cases the resynthesized sound was surprisingly close to the original. The discrepancies between the resynthesized and original sounds are probably due to the lack of envelope "shape" variations—in particular the time occurrence of the peak harmonic amplitudes is not preserved in the resynthesis.

Schindler [64] has combined features of Charbonneau's amplitude envelope data reduction with Strawn's hierarchical syntactic structures [69] (see page 19). The result is a real-time control scheme for additive synthesis—with intuitive control parameters aimed at sound designers/musicians. The control parameters are embedded within a multidimensional model of *timbre space* (see also Wessel [77], described on page 39). The hierarchical nature of the control scheme allows a sound designer to rough in a sound and then refine it with lower-level control features.

Grouping Harmonics

Recent work by Kleczkowski [34] attempts to combine the conceptual simplicity of the additive model with data reduction, by grouping harmonics based on a criterion of similarity. Since Charbonneau's research [10] indicated that data reduction of amplitude data has more of a perceptual effect than reduction of frequency data, Kleczkowski chose to group harmonics with respect to amplitude envelope similarity. The specific grouping criterion used was the sum-of-squares distance between two amplitude envelopes (line segment approximated envelopes were used).

A common envelope shape (the average of the grouped harmonics' envelopes) is weighted by the average amplitude of each harmonic in the group. The result is summed (in the manner of equation 2.1 on page 16) to get the contribution of that group to the waveform. The groups are then summed to get the complete waveform.

A common frequency fluctuation factor is used for all harmonics in a group. Alternate methods allow the individual harmonics (within a group) to be scaled in time (to preserve onset and ending times) and provide a variety of simpler ways to compute the common frequency fluctuations. The algorithm becomes more complicated when grouped harmonics are not all adjacent.

The model is aimed both at decreasing computational time for additive synthesis and reducing the data specification required. The sounds produced by three groupings of harmonics were deemed sufficiently close to the original (line segment approximated) sounds for several instruments. Four or five groupings were required for other instruments. Reducing the number of groupings also appears to result in gradual timbre alterations.

2.3.4 Attack Transients

Psychological research on the perception of attack transients indicates that they play a significant role in the recognition of instrument timbre [7, 61, 76][17]. Moorer and Grey [49] cite research [23, 38] showing that the attack portion alone is more useful in identifying instrument timbre than the steady state portion alone.

Grey and Moorer [27], using multidimensional scaling of timbre comparisons, found that attack information comprises a distinct perceptual dimension (see also *Timbre Spaces* on page 35). Attack information was taken to be the low-amplitude inharmonic components at the onset of a sound—for example, the "breathy" quality at the start of a woodwind sound—rather than the initial steep rise of harmonic amplitudes.

[17] A study by Saldanha and Corso [61] indicates that the ***final transients*** in a sound provide ***no*** information useful for identifying instruments. Final transients occur when the energy source driving an instrument is removed—allowing the sound to decay naturally. The nature of final transients varies considerably depending on how the instrument is activated. The instruments in Saldanha and Corso's study were either wind or bowed instruments, for which the sound quickly dies out when the energy source is removed. Final transients obviously play a much more significant role in plucked or struck instruments such as the guitar or piano.

The attack transients of the brass instrument family have been studied extensively [39, 40, 50, 56, 71]. The steep initial rise of harmonic amplitudes is generally considered part of the "attack" in these studies. According to Luce and Clark [40] all the brass instruments included in their analysis of 900 trombone, trumpet, tuba and French horn sounds—with the exception of the French horn—had similar attack characteristics. These included:

- Significant frequency and amplitude modulation occurs during the first 50 ± 20 ms of a tone. The duration is independent of fundamental frequency. The amplitude modulation frequency is very high (about 20% of the fundamental).
- Higher harmonics start later and have slower rise times.
- The amplitude envelopes exhibit characteristic "blips" at the end of the transient period. The "blips" are more pronounced for lower notes, higher harmonics, and louder sounds.

Luce and Clark suggest that the commonality in attack characteristics of brass instruments may be one of the invariant physical correlates that uniquely identifies this family of "brassy-sounding" instruments.

Some of the attack characteristics described above do not accord well with the additive model described by equation 2.1 (page 16). However the success of line segment approximated resynthesis [27, 49, 50, 56] seems to indicate that the rapid amplitude modulations are not perceptually significant. Moorer and Grey [50] also claim that it is not necessary to include the characteristic amplitude "blips" when resynthesizing trumpet sounds[18].

It is reasonable to assume that the microstructure of attack transients would vary substantially for different types of instruments (guitar, piano, brass, bowed strings) since attack characteristics result from the physics of the initially unstable interaction

[18] These are the same "blips" that gave Strawn's line segment extraction algorithms so much trouble [69] (see page 19).

between the energy source driving the instrument and the instrument itself (see Luce [39] for a clear description of this interaction for brass instruments).

Attack transients are good candidates for the physical invariants of recognition of certain instrument *classes*. The classes could likely be discriminated on the basis of the method used to input energy to the instrument (struck, blown into, bowed, etc.), however, it may be difficult to adequately capture the perceptual subtleties resulting from attack transients within the framework of the additive model.

2.3.5 Frequency Fluctuations

The analysis methods used by various researchers (pitch-synchronous analysis [42], heterodyne filter [45], and the phase vocoder [47]) yield time-varying frequencies as well as harmonic amplitude variations. Analysis of harmonic sound reveals that the harmonic frequencies are to some degree inharmonic (not exact integer multiples of the fundamental frequency)[19]. Over and above this *average* (macro) inharmonicity the frequencies also *fluctuate* over time [27, 48, 49, 50, 56].

Risset and Mathews [56] resynthesized trumpet sounds using *constant* harmonic frequencies and added "quasi-random" fluctuations to simulate the effect of the original frequency variations. The resynthesized versions apparently sounded much like the originals.

Grey and Moorer [27] conducted experiments to test the "perceptual distance" between *a)* sounds reconstructed with constant frequencies in place of the original time-varying functions (the amplitude envelopes were approximated by line segments), and *b)* sounds resynthesized with *all* the analysis information or sounds resynthesized with the original time-varying frequency functions and line segment amplitude approximations. The constant frequency sounds were easily discriminated from the other two. They concluded that this degree of simplification is too drastic (for all but

[19] An early study by Fletcher [20] on piano tones indicates that inharmonicity in the piano is perceptually important and adds to the "warmth" of the sound (inharmonicity is often missing in electric pianos and other imitations).

one instrument—the English horn).

Charbonneau [10] in his study of the effects of data reduction on additive synthesis parameters (see also *Envelope Data Reduction* on page 23) arrived at a slightly different conclusion. Charbonneau data reduced the harmonic frequency functions (F_k in equation 2.1 on page 16) by using the fundamental frequency function F_1 as a reference function. For each harmonic k, the k**th** harmonic frequency function was reconstructed from F_1 by multiplying it by k at each time point for which the k**th** harmonic amplitude was nonzero. The resynthesized sound thus lost any *average* inharmonicity present in the original and all harmonics fluctuated in synchronicity.

These sounds were compared to the fully-specified, frequency-varying sounds (amplitude envelopes were approximated by line segments for both sounds). Subjects judged the two sounds to be "perhaps slightly different" on average (discriminability varied over instruments and subjects). Charbonneau concluded that the ear is relatively insensitive to *individual* harmonic frequency fluctuations, although some sort of variation is required in the harmonic set as whole.

A useful extension to Charbonneau's experiments might be to assess the perceptual impact of algorithmically generated frequency varying functions—or one reference frequency function for a variety of notes and playing intensities on the same instrument (or instrument family). This would reduce the sound specification requirements even further. It seems plausible that unique F_1 analysis data is not required for *every* sound reconstructed by analysis and resynthesis.

2.3.6 Non-additive Models

Non-additive synthesis (such as FM [11] or Karplus-Strong plucked string synthesis [32, 33]) has great appeal because of its computational efficiency and simple control parameters. However additive models are usually preferred over non-additive models for timbre research since they are more flexible and the control parameters transfer well to the perceptual realm (although there are often too many of them).

Recent advances in waveshaping synthesis [1, 35, 63] may change this bias against non-additive techniques. Waveshaping synthesis essentially uses one function to modify another one in a nonlinear fashion[20], to produce the waveform of the sound directly. Two parameters (a scalar a and a sampled function f) are sufficient to specify all waveforms.

The power of the technique comes from the ability to produce a waveform corresponding to *any* given harmonic spectrum—and change it over time—by manipulating the parameter a and the function f. f controls the steady state spectrum and changing the value of a over the duration of the synthesis alters the spectrum.

While the spectrum can be changed over time by changing only *one* parameter, it does not (yet) appear to be possible to create *arbitrary* spectral changes over time [35, page 263]. The major difficulty is the complexity of the mathematics involved—what is easy to do in additive synthesis (fully specifying a dynamically changing spectra) becomes very complex in waveshaping.

The method is very efficient and shows promise as a flexible alternative to the additive synthesis model. However, additive models may still be more appropriate for studying the psychophysics of timbre given current knowledge of how the ear processes harmonic sound.

2.4 Principal Component Analysis of Spectra

Principal component analysis is a multivariate statistical technique useful in reducing the dimensionality of large data sets comprised of many variables. Principal component analysis has been used by researchers to data reduce vowel spectra information [52, 66, 78] and to classify timbre percepts (using multidimensional scaling—a cousin of principal component analysis) [24, 26, 27, 44, 51, 74, 76, 77] (see also *Multidimensional Scaling* on page 32, and *Timbre Spaces* on page 35).

A geometric interpretation of principal component analysis on page 48 provides an intuitive explanation of what the analysis does.

[20] Waveshaping as described by Le Brun [35] subsumes FM synthesis.

Plomp, Pols and van de Geer [52] used a principal component analysis to reduce the information necessary to specify vowel spectra. 15 vowel sounds produced by 10 subjects were analyzed by $\frac{1}{3}$ octave filters into 18 frequency bands. A 100 ms section was used to compute spectra for each vowel sound. When the spectral data was submitted to principal component analysis, the first 4 principal components accounted for 84% of the variance.

Plomp suggests that principal component analysis has advantages over formant analysis since it takes into account the *whole spectrum* (even though it is reduced to only 4 parameters). Formants typically are specified by their peak frequency and ignore the shape of the distribution of frequency energy surrounding the peak. The first two principal components were in fact related to the first two formants. A 2-dimensional plot of the first formant frequencies versus the second formant frequencies (for all 15 vowels) had the same *configuration* as a 2-dimensional plot of the the first two principal component weights (coordinates in the transformed space).

Principal component analysis has potential for automated speech recognition. Vowel sounds analyzed by extracting principal components would be identified by their proximity to previously analyzed and known vowels in a 2-dimensional space of the first two principal component weights. Plomp points out that it is much easier to analyze 18 frequency bands and extract principal component weights than to determine formant frequencies. It was suggested that the method be extended to analyze speech spectra at short time intervals (to see how they change over time) and to include consonants in the analyses.

Zahorian and Rothenberg [78] used an analysis-resynthesis paradigm to test the effect of principal component data reduction on the intelligibility of speech. One of the aims of their research was to determine the type of amplitude data best suited for data reduction. As they put it, "[the assumption]... *that data variance is equivalent to data "information"... depends strongly on the proper scaling of the data."* Their work indicates that logarithmic amplitude coding is superior to linear coding.

Zahorian and Rothenberg found that speech that was resynthesized with 3 principal components was judged to be 70% intelligible (on a standard test). The intelligibility went up to 85% for 5 components. The principal component basis vectors (rotated axes) were very similar for speakers of the same sex and the first few basis vectors were largely speaker independent.

It is noteworthy that the intelligibility scores, for a given number of components, compare roughly to the spectral variance accounted for, for the same number of components. This seems to indicate that the data variance corresponds somewhat to perceptually relevant information and that a loss of spectral "information" has a concomitant effect on aural perception.

Searle [66] performed a principal component analysis of speech spectra on a running sample of speech and concluded that speech spectra have perhaps 5 degrees of freedom. The component basis vectors (which are weighted and then summed to reproduce the approximated spectra) were *roughly* the shape of a half cosine series[21]. However, when the spectra were reconstructed with *exact* half cosine series basis vectors, the error was substantially increased. Therefore, establishing a *data dependent* set of basis vectors appears to be worth the effort.

Searle made an interesting observation about consonants. Vowels tend to show up as clusters in the transformed space since they change slowly over time and the changes are small. Stop consonants on the other hand are better represented as *trajectories* through the space as a function of time. As a direction for future research Searle recommends that a *temporal* dimension be included in the transformation (by coordinate rotations over time) so that trajectories will appear as clusters instead of lines.

[21] The first basis (the constant term) is a straight line, the second is a cosine shape from 0 to π, the third a cosine shape from 0 to 2π, etc.

2.5 Categorizing Timbral Percepts

The section on *Manipulating a Sound Model* (page 15) looked at timbre research that employs a common paradigm of systematically manipulating sound parameters and observing the perceptual effects. The research outlined in this section approaches the problem from the other end. Timbre is first examined as a purely psychological (or perceptual) phenomenon. Some of the questions to be answered are:

- Are there different *dimensions* of timbre classification? If so what are they?
- To what degree is it possible to *interpolate* between dimensions in "timbre space?"
- What are the verbal dimensions of timbre distinctions? Is language capable of categorizing timbral differences or are judgements made at a more basic cognitive level?

If timbre percepts can be categorized in a meaningful way it may be possible to work backwards and discover physical correlates of these categories. This could conceivably give sound designers (and musicians) low-dimensional control over sounds by the manipulation of intuitive perceptual parameters. The details of the actual signal construction could be transparent and take advantage of the established psychophysical relationships.

2.5.1 Multidimensional Scaling

Studying psychological phenomena, particularly complex ones like timbre, requires powerful analysis tools. Factor analysis (which includes principal component analysis, see pages 29 and 48) is one such method[22]. Since factor analysis is a mathematical technique, it requires data to be in numerical form. Unfortunately psychological

[22] Both the physical and psychological domains can benefit from dimension reduction. Wedin and Goude [76] applied dimension reduction to timbre comparisons and sound stimuli in order to test the correlations between them.

judgement scales are quite different from quantifiable physical scales in that they are not as "number-crunchable" (for example, how would you rate the similarity of two sounds on a meaningful numerical scale?)[23]. Multidimensional scaling is a technique that has been developed to allow factor analytic methods to be applied to judgements of similarity [6].

In order to quantitatively assess similarity judgements with multidimensional scaling, all possible pair-wise comparisons between the stimuli must be made[24]. These similarity judgements are taken to represent subjective "distances" when the stimuli are (conceptually) mapped into a space whose dimension is the same as the number of stimuli. Each stimulus defines a point in this space.

If there are common strategies underlying the similarity judgements then a factor analysis on the sums of cross-products of all the paired comparisons [6, chapter 6] will reduce the number of dimensions required to assess the stimuli differences[25].

The reduced dimensions will reflect the range of stimuli involved, therefore it is important to choose a representative sample of stimuli if the results are to be generalized. For example, timbral dimensions could be assessed with stimuli restricted to harmonic sounds, or percussive and inharmonic sounds could be included as well. In the latter case, *homogeneity* of stimuli may be important. For example, large (perceptual) distances between instrument families (e.g. percussive and non-percussive) can cause degenerate multidimensional scaling solutions. This can be avoided by including sounds that bridge the gap between the two families. Nonhomogeneity of stimuli can also result in poor intra-family distinctions [77].

The number of stimuli will in general affect the distance *resolution* within the

[23] Psychological variables can be discrete or continuous and are typically categorized as nominal, ordinal, interval, or ratio. Ratio variables inherit all the properties of the number system that are taken for granted in most physical variables—such as equal intervals, equality of ratios (an implied absolute zero), etc. Judgement scales employ, at best, interval variables—ordinal variables are more common [18]. *Sones* (loudness) and *mels* (pitch) are two of the aural scales that have been developed [72].

[24] This imposes a practical limit on the number of stimuli in an experiment since for n stimuli $\frac{n!}{2!(n-2)!}$ comparisons are required ($n(n-1)$ if order of presentation is included).

[25] The original space (with dimension equal to the number of stimuli) allows for the possibility that *every* comparison could be based on a different criterion.

reduced dimension space. Wessel [77] indicates that 10 stimuli are a minimum for reduction to 2-dimensions (15 stimuli for 3-dimensions), although better results will be obtained with more stimuli. Stimuli should also be equalized for potentially confounding influences—for example, loudness, pitch, and duration should be the same when timbral similarities are being judged.

Distances between stimuli will be preserved (with some degree of approximation) in this reduced dimension space[26]. Reduction to 2 or 3 dimensions allows stimuli distances to be visualized so that patterns can be more easily seen—for example, clusters of stimuli points would indicate that these stimuli are perceived as similar. Each dimension corresponds to some as yet unnamed criterion of similarity.

At this point it is up to the researcher to interpret the dimensions. Typically this is done by correlating stimuli positions (on a particular dimension) with other phenomena—either other psychological factors or in the case of perception, physical correlates.

One way of interpreting the dimensions resulting from timbre comparisons is by answering the question: *What factors in the sound signal are causing timbres to be discriminated on this* one *dimension?* If for example the amount of high harmonic content[27] in the signal was found to be linearly correlated with the arrangement of stimuli on a particular dimension axis—and other studies indicated that sounds with high harmonic content are heard as "shrill"—then there is a case to be made for labeling the axis as one of "shrillness" discrimination.

Multidimensional scaling requires no *a priori* knowledge of judgement criteria, hence, the choice of stimuli is not constrained by preconceived ideas of what is being measured. A variation of multidimensional scaling uses *semantic scales* rather than similarity judgements. Subjects rate stimuli differences based on proximity to a pair (usually several pairs) of bipolar adjectives. This restricts subject responses to those

[26] As with *variance accounted for* in principal component data reduction (see page 51), mulitidimensional scaling computes an overall error between the original similarity judgements and their dimension reduced approximations.

[27] The amount of high harmonic content would have to be reduced to some scalar measure in order to correlate it with the distance relations on a perceptual axis.

that the researcher feels are relevant, hence, a priori knowledge of similarity criteria is useful. One hazard of semantic scales is that relevant (original) dimensions may be omitted—or even more serious—that some stimuli discriminations may not be easily translated into words [23].

An advantage of multidimensional analysis of similarities is that subjects are required to make a *synthetic* judgment as opposed to an *analytic* one [21]. Synthetic judgements are ones in which subjects assess stimuli as a whole rather than having to concentrate on a single aspect of the stimuli. This allows the use of a wide range of natural stimuli instead of ones artificially created to represent analytic categories. Freed and Martens [21] consider synthetic judgements to be more natural for perception research.

2.5.2 Timbre Spaces

Wedin and Goude [76] reduced similarity judgements between 10 different instrument sounds to 3 dimensions with a variant of multidimensional scaling (3 dimensions accounted for 75% of the variance). The perceptual dimensions did not discriminate the instrument families involved (woodwind, brass, and strings). However when the *names* of the instruments were presented as stimuli instead of their sounds, the dimensions coincided quite well with the three instrument families. Wedin and Goude concluded that the "cognitive structure" of timbre distinctions does not coincide with the "perceptual structure."

The steady state spectra of the stimuli were also factor analyzed (3 factors accounting for 95% of the variance). Correlations between the acoustic basis vectors and the perceptual basis vectors revealed that the first perceptual dimension was related to "overtone richness," the second to "overtone poorness," and the third to a low fundamental combined with increasing intensity of the first overtones. Wedin and Goude also found that the presence or absence of the attack portions of sound affected instrument identifiability but did *not* alter the dimensional structure. Subsequent studies have not supported these results [24, 26, 77] except that in all cases the *first* dimension is in some way related to the sound spectrum.

Miller and Carterette [44] used artificial tones with carefully manipulated characteristics. Multidimensional scaling revealed three perceptual dimensions for timbral similarity judgements. The first two dimensions were correlated with the number of harmonics in the artificial sound and the shape of the temporal energy envelope, respectively. The use of simplistic stimuli in this experiment limits the usefulness of these results. Freed and Martens [21] state, *"It seems intuitively obvious that musically relevant timbre research cannot employ musically useless timbres as stimuli."* They go on to cite Gibson's claim [22] that perceptual systems require complexity in order to function properly—the justification being that perception targets physical properties of sound sources, which are inherently complex.

The most comprehensive work in categorizing the dimensions of timbre has been done by Grey [23, 24, 26, 27] and Wessel [77].

Grey

Grey [24] analyzed 16 instrument sounds from 12 instruments of the woodwind, brass, and stringed instrument families. Multidimensional scaling of all-pairs timbre comparisons (over 35 subjects) was performed to obtain 2, 3, and 4-dimensional timbre spaces. The 2-dimensional space was difficult to interpret and the 4-dimensional space yielded no additional useful information over the 3-dimensional space. Since the stimuli were sounds resynthesized from time-varying spectral analyses (with line segment envelopes), it was possible to informally relate the perceptual dimensions to properties of a simplified additive model.

A preliminary analysis revealed that the first dimension was related to the *spectral energy distribution* of the sounds. The second and third dimensions appeared to correspond to *temporal* features of the sounds.

The second dimension was associated with a lack of attack synchronicity of harmonics and an accompanying spectral fluctuation over time[28]. Spectral fluctuation

[28] Apparently these two signal characteristics occurred in tandem for the instruments analyzed. For

here refers to macro level changes in the *shape* of the spectrum over time. The second dimension also discriminated the woodwind, brass, and stringed instrument families, with two exceptions—the bassoon was grouped with the brass, and the flute was grouped with the strings.

The third perceptual dimension was related to low-amplitude, high-frequency, inharmonic energy in the attack portion of a sound. Brasses tend to have very little while strings and some woodwinds have considerably more.

The distribution of instrument stimuli in the 3-dimensional space had another interpretation. Grey also analyzed the instrument spatial relationships with a hierarchical clustering algorithm—grouping instruments by their inter-distances in the space, irrespective of dimension. The analysis revealed that instrument families tended to cluster together[29]. Instrument family clusters were configured in a roughly cylindrical pattern around the first (spectral energy distribution) dimension. Therefore, combining the second and third dimensions appears to group instruments by temporal qualities, which are somewhat related to family characteristics. The exceptions indicate that temporal attributes of an instrument's sound override traditional family groupings.

Grey and Gordon [26] confirmed and extended some of the dimension interpretations of Grey's original study [24]. Eight of the original sounds were altered by exchanging spectral envelopes in pairs. This consisted of keeping the same harmonic amplitude envelope "shapes," but altering the peak harmonic amplitudes attained, in accordance with the other sound. The same bandwidth for the altered sounds was maintained by not swapping peak values for harmonics that were not present in the original sounds. The point of swapping spectral envelopes in this manner was

example woodwind upper harmonics tended to enter as a group, with the spectral shape somewhat invariant over time when all harmonics were present. Strings and brass, on the other hand, tended to have staggered patterns of harmonic onset and exit (see *Attack Transients*, page 25) with more spectral variation over time.

[29] The exceptions to family clustering (bassoon with the brass; trumpet, French horn, and flute with the strings) could be explained by the uncharacteristic articulatory patterns of these instruments (with respect to other members of their families). Grey speculates that temporal qualities of the attack may override familial groupings.

to exchange spectral energy distributions without perturbing temporal qualities (the change in harmonic envelope *slopes* did alter temporal qualities to some extent).

A multidimensional analysis of the 8 altered and 8 unaltered sounds resulted in the same dimension configurations as the original study [24]—except that the altered sounds swapped positions along the first dimension. Positions in the other two dimensions were only slightly altered. This supports the original interpretation of the first dimension as resulting from the spectral energy distribution of sound and indicates that the first dimension is indeed independent of the other two (as it should be given the orthogonal multidimensional scaling solution).

The altered sounds were *heard* as a hybrid of the original and its swapping partner. The articulatory qualities of the original were combined with the vowel-like tone "color" of the other.

Grey also confirmed the first dimension interpretation by deriving a quantitative measure of spectral energy distribution and correlating it with positions on the first dimension axis. Several alternate strategies of assessing spectral energy distribution were tried. They all resulted in high correlations with the first axis positions. The best correlation resulted from deriving a line spectrum from the harmonic (linear scale) amplitudes averaged over time—with critical band loudness corrections. The scalar value for energy distribution was derived from the line spectrum by taking its centroid.

Grey suggested some subjective interpretations for the two temporal dimensions. The second dimension (which correlated with spectral fluctuation) was said to capture the quality of *static* versus *dynamic* tones.

The third dimension was associated with *attack* characteristics. Low-amplitude, high-frequency, inharmonic content in the attack has a noiselike quality which, according to Wessel (cited in [26]), results in a subjective impression of a soft, long lasting attack (more time is in fact taken to reach maximum amplitude). On the other extreme, low noise attacks—where the lower harmonics come in quickly—have a fast, harder, more explosive sounding attack.

In Grey and Moorer's study [27] of the effects of three types of data reduction on instrument resynthesis quality (see also *Attack Transients* on page 25), the differences between the test sounds were analyzed with multidimensional scaling to see if different similarity criteria played a part in discriminating the various types of data reduction[30].

A one-dimensional solution resulted in the following ordering along the dimension, original tone, complex resynthesis (*all* the analysis information included), line segment approximation, and constant frequency approximation (with line segment amplitude envelopes).

The two-dimensional solution revealed that a similar criterion was used to discriminate all but the "cut-attack" approximation which had the second dimension all to itself. This supports the conjecture that the attack portion of a sound (at least a noisy one) forms a separate basis for categorizing instrument sounds that cuts across instrument family boundaries.

Grey has also looked at timbre discrimination in musical contexts [25] and interpolation between positions in timbre space [23].

Wessel

Wessel [77] constructed a model of timbre space with the intent of making it a useful tool for the compositional control of timbre in musical contexts. The ultimate goal is high level control of timbral sounds using perceptual parameters.

The 24 instrument sounds used by Grey (16 natural instruments plus the 8 hybrids used in [26])[31] were rated for similarity (by Wessel himself). A 2-dimensional timbre

[30] Similarity judgements were also analyzed to determine the degree to which the tones could be distinguished in pairs. The least discriminable were the complex resynthesis and line segment approximations. The most discriminable were the "cut-attack" approximations (removal of low-amplitude, high-frequency, inharmonic energy in the attack)—with everything else. The constant frequency approximation was found to be highly discriminable with the complex resynthesis and the line segment approximations. The original tone was somewhat discriminable from the complex resynthesis and the line segment approximated sounds.

[31] These 16 instrument sounds, originating from *CCRMA*, have been widely used in timbre research [10, 23, 24, 25, 26, 27, 34, 46, 49, 50, 64, 69, 77]. The sounds are short (280-400 ms) E♭ (≈ 311 Hz) notes of 12 different instruments—oboe, English horn, bassoon, E♭ clarinet, bass clarinet, flute,

space was constructed from the similarity judgements using multidimensional scaling.

The first dimension was interpreted as a perceptual correlate of *spectral energy distribution* (as with Grey) and the second as a correlate of *onset transient* characteristics. Sounds on the first dimension varied from "bright" to "mellow." The second dimension characterized the quality of the "bite" in the attack. The first dimension correlated well with a centroid measure of the spectral energy distribution (similar to Grey's measure [26]).

To illustrate the usefulness of the timbre space parameters in musical contexts, Wessel demonstrated *auditory streaming effects* [43] created by varying timbres with respect to timbre space location. A 3 note ascending sequence of notes was played repeatedly. The timbre of alternate notes could be altered (in unison) by choosing sounds at different positions along the spectral energy distribution dimension of the timbre space. When the timbres of adjacent notes were similar (in terms of timbre space position), the ascending pitch line of the notes dominated perception. However, as the two timbres were moved apart (in timbre space), the perceptual organization of the note pattern split into *two* ascending series of notes.

Repeating the experiment with timbres varying on the "attack bite" dimension resulted in the *rhythm* of the musical pattern moving from even to irregular. Wessel concluded that altering the attack characteristics of a sound affects the subjective onset time.

Wessel also experimented with *timbral analogies*. Subjects were presented with two sounds, A and B, which had different timbre space positions. A third sound C (at another timbre space location) was presented and subjects were asked to decide which of 4 additional sounds was the best *analogy* with respect to C of the timbral relationship between A and B. The results indicated that the sound selected had a timbre that came the closest to completing a parallelogram in the 2-dimensional timbre space. In addition, the rank ordering of the alternatives corresponded with

alto saxophone, soprano saxophone, trumpet, French horn, muted trombone, and cello. The sounds include a complete attack and a natural decay. The additional 4 sounds were produced with the same instruments under different playing conditions. Sounds were recorded to tape and digitized at 12 bit resolution with a sampling rate of 25,600 samples per second.

their proximity to the ideal parallelogram position.

An interactive graphics program was developed to synthesize sounds that would correspond to a given position in timbre space—by manipulating line segment envelopes (spectral energy distribution)—and attack characteristics (second dimension). Choices are presented on a 2-dimensional grid and by selecting a point in this space, a sound with the desired characteristics can be synthesized. Wessel reports that perceptually smooth transitions result when moving around in the space. This structural continuity suggests that a reasonably quantitative psychophysical timbre relationship has been developed—one that holds up well in a musical context. Wessel states that an efficient control scheme for manipulating envelopes is required to facilitate more complex forms of sound manipulation in timbre space.

2.5.3 Timbre Semantics

Work from the previous section on *Timbre Spaces* arranged timbres in a space without resorting to verbal attributes (although the resulting spaces could be described in those terms). This section looks at how language discriminates timbre.

Lichte [37] in an early study (1941) found three attributes of complex tones (other than loudness and pitch). "Brightness" was related to the mid-point of the spectral energy distribution, "roughness" to the presence and location of partials above the sixth, and "fullness" to the relative presence of odd and even partials. All the stimuli were artificial.

Solomon studied the semantics of auditory perception of complex tones [68]. Navy sonarmen were asked to rate 20 actual sonar sounds on 50 seven-point scales of bipolar adjectives (heavy-light, smooth-rough, clean-dirty, etc.). All the participants were experienced at interpreting sonar sounds and had developed their own informal vocabulary in order to communicate to others the qualities of sonar sounds. Solomon used a standard test, the Semantic Differential (similar to multidimensional scaling), to determine the essential "semantic space" of the sonarmen judgements of sound quality.

Eight factors accounted for 42% of the variance. Factors were correlated with the bipolar adjectives to yield an interpretion of the semantic dimensions. The first three dimensions were found to be "magnitude," aesthetic judgement, and "clarity." A reasonable conclusion is that the first dimension is related to either pitch or loudness (corresponding adjectives were heavy, large, rumbling, low, wide, etc.), and the second and third dimensions to timbral qualities. Unfortunately the results are likely confounded with what the sonar sounds usually *mean*—for example, the size of a submarine. No attempt was made to relate semantic dimensions to acoustic qualities of the sonar sounds.

Von Bismarck investigated the verbal attributes of timbre of steady state sounds [74]. The Semantic Differential scales were preselected by the subjects for their appropriateness to the task. Von Bismarck's hypothesis was that timbres can be uniquely described by a small number of verbal categories.

The 35 artificial timbres were found to be categorizable with only 4 of the original 30 scales. The first factor, labeled "sharpness," was the only one deemed consistent enough to be a general attribute of timbre. Sharpness judgements were correlated with the frequency location of the energy concentration in stimuli spectra. The other factors were plagued by large individual differences among the subjects. The features (dimensions) of timbre not accounted for by sharpness did not appear to have obvious verbal labels.

In a subsequent study Von Bismarck explored the psychophysical relationships involved in the "sharpness" dimension [73]. Sharpness judgements were scaled to several acoustic parameters including limiting frequencies of broadband sounds and spectral envelope slopes. Sharpness scales were consistent over a reasonable range of physical parameters when sounds were equalized for loudness and pitch. The interaction of loudness and pitch with sharpness judgements was complex.

Research on the linguistic categories of timbre has not been as fruitful as the timbre space research. The prearranged categories for subject judgements in semantic studies is a much less useful paradigm than the unspecified judgement criteria in timbre space research.

The major reason for the poor results from semantic studies may simply be that words do not capture the experience of timbre. The perceptual mechanisms responsible for vision and hearing are several orders of magnitude anterior to language mechanisms. Visual and aural experience may use a language that cannot be put into words.

Chapter 3

An Alternate Envelope Representation

The following section summarizes some of the results from the previous chapter, *Physical Correlates of Timbre.*

3.1 Data Reduction and the Additive Model

Virtually any harmonic sound can be digitally constructed using the additive synthesis model, including sounds that mimic acoustic instruments. Unfortunately the complexity of the model makes this extremely cumbersome. There are simply too many degrees of freedom.

Analysis of acoustic instrument sound reveals that there is a considerable amount of redundancy in the signal specification. This redundancy results from the cohesiveness of physical processes and can be used to reduce the dimensionality of model data [10, 34, 64]. In terms of the additive model expressed by equation 2.1 (page 16), this redundancy results in autocorrelation within the harmonic amplitude function A_k—both over time and over the k harmonics. Frequency fluctuations F_k appear to be more random.

Factor analysis, of which principal component analysis and multidimensional scaling are two variants, is a powerful method of exploiting redundancy to reduce model dimensionality. Principal component analysis of vowel speech indicates that the frequency information contained in steady-state vowel spectra [52, 66, 78] can be usefully reduced to 3 to 5 dimensions (see page 29). Similar reductions occur for musical sound spectra [76]. Multidimensional scaling of timbral similarities also uncovers low-dimensional *perceptual* processes [24, 26, 27, 44, 51, 74, 76, 77] (see page 35).

The results obtained by Grey [23, 24, 26, 27] (see page 36) and Wessel [77] (see page 39) with *timbre spaces* indicates that the low-dimensionality of both physical and perceptual processes can be put to good use in constructing musically interesting timbres with a small number of perceptually relevant parameters.

However, the global parameters used by Grey and Wessel to capture physical dimensions (centroid of the spectral energy distribution, attack profiles, etc.) are fairly crude[1]. Finer control over additive parameters is desirable, without any loss of high level access. A hierarchical control scheme as proposed by Strawn [69] (see page 19) and Schindler [64] (see page 24) would be desirable.

Within the additive model expressed by equation 2.1 (page 16), two possible sources exist for significant data/dimension reduction, harmonic amplitude envelopes A_k and harmonic frequency fluctuations F_k.

3.1.1 Harmonic Frequency Fluctuations

Good results have been obtained for the reduction of frequency fluctuation data. While Grey [27] concluded that constant frequency approximations to harmonic components were highly discriminable, further work by Charbonneau [10] (see page 28) and Kleczkowski [34] (see page 24) indicates that frequencies need not be allowed to fluctuate *independently* over harmonics. Risset and Mathews' original research [56]

[1] These parameters are aimed primarily at providing interpretations of timbre space dimensions (by correlating the parameters with the distribution of sounds in timbre space) rather than as useful resynthesis parameters.

also indicates that "quasi-random" approximations are an adequate replacement for detailed fluctuation functions. In general, some sort of harmonic frequency fluctuation appears to be desirable for capturing the nature of realistic acoustic instrument sound. Frequency fluctuations may be suitable for an *algorithmic* treatment, rather than having to be derived from sound analyses.

3.1.2 Attack Transients

Research indicates that this aspect of instrument sound is very important for instrument recognition [7, 61, 76] (see page 25). It also appears that a distinct perceptual dimension is involved in processing the onset characteristics of sound [23, 24, 26, 27, 77] (see page 35). Attack characteristics vary widely over instruments and instrument families and are dictated somewhat by the method used to input energy to the instrument. Studies of instrument specific attack transients indicate that considerable additive parameter fluctuation occurs [19, 39, 40, 50, 56, 71]—so much in fact that it may be questionable to refer to attack reconstruction as an additive (harmonic) process.

It is not clear how to incorporate attack transients into the additive model. Some transient characteristics can be captured with amplitude envelope and frequency fluctuation parameters[2]. However, it may be necessary to add additional attack detail with some other module or method, in order to satisfy the ear's high discriminability for attack characteristics.

3.1.3 Harmonic Amplitude Envelopes

Envelope approximation with line segments has been quite successful [27, 50] (see page 18). Further data reduction is possible using line segment envelopes as a base, by taking advantage of interharmonic correlations [10, 34, 64] (see page 23).

[2] A spectral analysis of sound digitized at a high sampling rate facilitates the extraction of rapidly changing onset characteristics. However, the frequency components are likely to be both inharmonic and unstable.

Line segment approximations, on the other hand, are difficult to obtain [69] (see page 19), and while manipulating breakpoints (of individual harmonics) may be straightforward, the technique is heuristic and offers no standardized way to compare and manipulate envelopes across harmonics and sounds. Strawn proposes some solutions to the problem of line segment extraction in [69] (see page 19).

As with attack transients, the information provided by amplitude envelopes is largely *sound specific* (envelopes will vary over instrument, intensity, note register, player parameters, room acoustics, etc.). For the foreseeable future at least, reproducing subtle differences in acoustic instrument timbre will likely require the use of information obtained from a pre-analysis of actual sounds.

The problem then becomes how to best reduce the information contained in individual (analyzed) envelopes and *combine* that information on a larger scale over all the harmonics of a sound. Instrument specific methods (mostly brass) exist for exploiting interharmonic relationships [3, 4, 57] (see page 23), as well as more general methods [10, 34, 64] (see pages 23 and 24). The general methods rely on line segment envelopes with their attendant problems. Wessel [77] states that a more efficient means of representing and manipulating envelopes would be a great asset in utilizing the power of timbre spaces (see page 39).

Since harmonic envelopes are themselves multidimensional phenomena, it is somewhat surprising that the dimension-reducing capabilities of principal component analysis have not been applied to harmonic envelopes[3]. Reducing the dimensionality of envelope data with principal component analysis would also have the advantage of expressing envelope variance in a standard form.

[3] The method has been used by Wedin and Goude [76] for *spectra* reduction (see page 35) but only for the purpose of correlating physical and perceptual dimensions.

3.2 Geometric Interpretation of Principal Component Analysis

Principal component analysis is a multivariate statistical technique useful in reducing the dimensionality of large data sets comprised of many variables[4]. Principal component analysis has been used by researchers to data reduce vowel spectra information (see page 29), and to classify timbre percepts (see page 32).

While principal component analysis is realized by eigenvalue solutions to the covariance (or correlation) matrix of a set of variables (see page 71), it also has a geometric interpretation that aids in understanding what it does and why it is useful.

3.2.1 Spectral Data Reduction Example

To illustrate the technique, consider the *spectral shape* of a steady-state harmonic sound (ignore variations over time for the moment).

For any given harmonic sound, graphing *SPL* (sound pressure levels [72]) on the Y axis versus frequency on the X axis will result in points plotted on the graph at each harmonic frequency (the "spectral shape" can be considered to be the curve resulting from connecting adjacent points with straight line segments). As discussed previously (page 20), spectral shape is useful in characterizing the timbre of a sound.

If the sounds are composed of k harmonics, then k *SPL* values would be required to fully specify the spectral shape. These k *SPL* values can also be conceptualized as a *vector* in k-dimensional space. Each of the k axes (one for each component of the vector) would record the *SPL* levels for a particular harmonic over all sounds[5]. If we

[4] See [6, 60] for a readable introduction to multivariate methods and applied factor analysis, and [12, 55] for a mathematical treatment.

[5] It is important to realize that these coordinate axes (each defined by a particular harmonic) are ***not orthogonal*** (perpendicular to each other). They are not orthogonal since the harmonic *SPL* values are ***correlated*** with each other. For example, if it happened to be the case (for the set of sounds we are analyzing) that a high *SPL* value for the 2nd harmonic generally indicates that the 4th harmonic *SPL* value will be high then the 2nd harmonic is said to be correlated with the 4th harmonic. This correlation means that the vectors that represent the axes are ***statistically dependent***

limit the number of harmonics to 3 (so that $k = 3$) then it is easy to visualize the 3 *SPL* values as defining a point in the 3-dimensional space we are accustomed to.

3.2.2 Comparing Spectral Shapes

Since the hypothesis is that spectral shape is a correlate of timbre, it would be useful to analyze *many* harmonic sounds and plot each spectral shape (now represented as a k-vector) in our k-dimensional space. Sounds with similar spectral shapes will be close together in this k-space (you can demonstrate this by drawing several similar shapes for $k = 3$ and plotting them).

If the sounds represented by points that cluster together are determined to all have a similar timbre (we could stipulate that pitch and loudness are the same for all sounds) then we could conclude that small variations in spectral shape (the physical correlate) lead to small variations in timbre (the perceptual attribute)—this is in fact the case but somewhat trivial.

A geometric interpretation allows us to conceptualize in a simple way the relationships between points in this space—and hence the relationships between timbres. For example, it might be the case that *different* cluster groups (Plomp refers to them as "clouds" [52]) would correspond to instrument families with similar timbre. We could now compare this physical correlate of timbre (spectral shape) over different instrument families.

However the geometric interpretation is somewhat illusory. The simple "point" in the geometric model is a convenient spatial abstraction. What we really have is a large collection of numbers—k-vectors—to describe each point. What is needed is a way to reduce the number of numbers involved. This is where principal component analysis comes in.

on each other. Dependent vectors are not orthogonal. Part of what principal component analysis does is to replace the original axes with a set of orthogonal axes.

3.2.3 Orthogonal Rotated Axes

What principal component analysis does essentially is to move the *origin* of the k-dimensional space to the "center of gravity" of all the points in the space[6] and then *rotate* the axes about the new origin so that the *principal axis* minimizes the variance[7] of all the points about that axis. For a 3-dimensional space, the cluster of points might all be contained within an ellipsoid and the principal axis would be the major axis of the ellipsoid. The second axis is also aligned to minimize the variance—subject to the constraints that the first (principal) axis is now fixed in position and the second axis must be orthogonal to the first axis. The next $k-3$ axes are aligned in a similar fashion (the last axis has no choice in the matter since there are only $k-1$ degrees of freedom).

While the original axes have a simple interpretation (for example, positions on the first axis represent the *SPL* values of the fundamentals) the newly translated and rotated axes do not. The new axes "cut through" many of the dimensions of the original axes and no longer represent something measurable in the physical world.

The advantage of the transformation is that if the set of points has a non-random configuration (some sort of k-dimensional "shape" to it) then the new principal axis will capture more information as to how the original set of points varied. For example, with 3 dimensions and an ellipsoidal set of points, the points will be spread out to a large degree along the principal axis since the principal axis coincides with the major axis of the ellipsoid.

[6] The origin is *translated* to the "center of gravity" of the points if each of the *SPL* values for a particular harmonic is subtracted from the *mean* of all the *SPL* values (for the same harmonic) over all sounds. Principal component analysis can also be done on the *raw* values. This is more difficult to visualize in the geometric interpretation.

[7] The variance is the sum—over all points—of the squared shortest distances between the points and the axis in question, divided by the number of points. Minimizing the variance with respect to the axis essentially means fitting the "best" straight line through the set of points.

3.2.4 Variance Accounted For

The spread of points is *maximized* along the principal axis. This is a side effect of the variance minimization when setting the first rotation. For this reason the largest *variance accounted for* is attributed to the principal axis. Since the next rotation (for the second axis) repeats the process of minimizing variance, the next largest variance accounted for is due to the second axis... and so on[8].

3.2.5 Reducing the Number of Dimensions

The point of the axes rotation is that it may now be possible to *approximate* the original set of data (points in k-dimensional space or k-vectors) by considerably fewer dimensions.

For example, the coordinate of a point *with respect to* the first principal component axis is an approximation to the point since moving from the origin to that coordinate brings us closer to the point. If from here we now move in the direction of the second principal component axis by an amount equal to the point's coordinate for that axis, we get even closer to the point... and so on for the remaining axes. Since the variance accounted for is inversely proportional to the principal component dimension number, adding progressively more dimensions to the *reconstruction* of the point's position has increasingly less impact as we zero in on the original point[9].

Depending on the degree of *dimension reduction* possible with the rotated axes, the approximations may be quite accurate. The amount of dimension reduction that occurs will depend on:

- The nature of the data set. Data that has a great deal of redundancy built in is well suited to principal component analysis.

[8] The *total* variance is the sum of the squared distances of all points from the translated origin, divided by the number of points. Since the total variance is the sum of the variances for each dimension (the Pythagorean theorem applies since the axes are now orthogonal), the variance about each axis can be said to "account for" some portion of the total variance.

[9] The original points can always be reconstructed *exactly* from their new coordinates in the rotated space by using *all* the dimensions.

- The relative amount of variance accounted for by the first few components.
- The number of principal components used in the reconstruction.

For example, if 3 principal components capture 85% of the data set variance and the number of original axes was 20 (harmonics) then only 3 numbers are required to approximate the spectral shapes—down from the original 20. This is a considerable data reduction. The shapes can be reconstructed with 85% accuracy (averaged over all the spectral shapes originally analyzed).

Reconstructing an approximation to the original spectral curves involves a simple linear transformation using the principal component axes (the new axes are k-vectors in the *original* space) and the coordinates[10] of the point with respect to those axes (see equation 4.6 on page 72).

Principal component analysis is a *statistical* technique commonly applied to data with stochastic properties, hence, some points—out in left field so to speak (hopefully a small number)—may *not* be well approximated with a small number of components.

3.2.6 Factor Analysis

Factor analysis begins with a principal component analysis (or something similar) but often goes further in assigning *meaning* to the newly rotated axes by observing which of the original dimensions are correlated with them[11]. This usually makes more sense when the dimensions (variables) measure different phenomena (for example, different economic indicators). However, it is useful when using homologous variables (such as harmonic *SPL* values) to look for correlations between principal components and phenomena *other* than the original variables. For example, the second component in the spectral shape example might weight the high harmonics and be correlated with "brightness of tone" judgements.

[10] These coordinates are the *weights* referred to in the more technical explanation of principal component analysis on page 71. The newly rotated axes are the *bases*.

[11] Factors can be further rotated to facilitate this.

3.3 Proposed Envelope Representation

The explanation of principal component analysis in the previous section used as an example the data reduction of steady-state spectral curves. Variations of spectra over time were not considered. However, principal component analysis is capable of reducing the dimensionality of *any* set of data curves. Since harmonic spectrum variation over time plays a significant role in the perception of timbre, and spectral changes can be captured by the amplitude envelopes of individual harmonics, it may be of some advantage to apply principal component data reduction to the harmonic amplitudes as they vary over time.

For example, if the amplitude envelope curves (derived from an analyis of sound) were specified at 50 ms intervals over the duration of a sound, then 20 envelope points (per harmonic) would be required to specify the amplitude changes over 1 second of sound. If the amplitude curves are similar in shape, over different harmonics and sounds, then principal component analysis may be capable of reducing the number of parameters required to specify the curves (with some degree of approximation). The method used to reconstruct the curves is described on page 71.

One advantage of representing the envelope curves with this reduced representation is that all curves will be reconstructed from a set of bases (the rotated axes in the previous example) that are *fixed* for all envelope curves. Hence, the few parameters (weights) required to specify the curves will be directly comparable. This may be useful for the development of higher level control mechanisms that manipulate these curve-altering parameters to produce changes in timbre.

The other major advantage is that the extraction of bases and weights is automated. This makes it feasible to catalogue a large number of sounds with subtle differences in timbre. The standardized representation may allow the effects of various parameters on timbre (note register, note intensity, etc.) to be better understood.

The *Principal Component Analysis of Envelopes* section in Chapter 4 (page 69) goes into more detail on the representation and also outlines the mathematical manipulations involved.

Chapter 4

Data Collection and Analysis

In order to extract representative principal components of harmonic amplitude envelopes, a large number of envelope curves were required. Existing data was not available in sufficient quantity or suitable form. To obtain the curves, a program was written on a microcomputer to analyze digitized sounds. The sounds analyzed were musical instrument sounds of either 2.7 seconds duration (208 sounds), or .65 seconds duration (72 sounds). The instruments included were, trombone, E♭ clarinet, flute, tenor saxophone, piano, classical guitar, and steel string guitar (see *Sound Samples* on page 67).

4.1 Equipment

An AKG D1200E microphone was used to record the sounds to cassette tape. Sounds were input to an Amiga microcomputer with an inexpensive 8 bit analog to digital converter connected to the computer's parallel port. Sounds were digitized at a sampling rate of 28,185 samples per second, and a commercial program was used to edit and store the digitized sounds to disk (in a standard Interchange File Format—IFF).

4.2 Spectral Analysis Method

A C program was written to analyze the sound samples and extract time-varying spectra. Harmonic amplitude envelopes were obtained with a Fast Fourier Transform (FFT) applied at successive positions in the sound samples[1]. The FFT is an algorithm for computing the Discrete Fourier Transform (DFT) [54].

The DFT is the discrete version of the Fourier integral, hence a sampled time domain is transformed into a sampled frequency domain with a frequency resolution Δf derived from the DFT window size N (number of samples being considered), and the sampling interval Δt (see below). The equation for the DFT is,

$$X_f(k\Delta f) = \Delta t \sum_{n=0}^{N-1} X_t(n\Delta t) e^{-j2\pi k \Delta f n \Delta t} \tag{4.1}$$

or (using the identity $e^{\pm j\theta} = \cos\theta \pm j \sin\theta$),

$$X_f(k) = \frac{1}{N} \sum_{n=0}^{N-1} X_t(n) \frac{\cos 2\pi kn}{N} - jX_t(n) \frac{\sin 2\pi kn}{N} \tag{4.2}$$

Where,

- n is the time sample index, $n = 0, 1, 2, \ldots, N-1$.
- N is the number of samples in the DFT (or FFT) window.
- Δt is the time between samples in seconds.
- $N\Delta t$ is the window size in seconds.
- k is the frequency domain index, $k = 0, 1, 2, \ldots, N-1$.
- Δf is the frequency spacing (resolution) in Hz, where $\Delta f = \frac{1}{N\Delta t}$.
- X_t is the time domain values at time $n\Delta t$.

[1] The FFT was used because of its $O(n \log n)$ computational efficiency since a large number of sounds were to be analyzed.

- X_f is the frequency domain values obtained from the DFT of $X_t(n\Delta t)$.

DFT frequencies are, in general, positive and negative. For real valued functions (such as sound) the sampled frequencies that result are $0, \Delta f, 2\Delta f, 3\Delta f, \ldots, \frac{N}{2}\Delta f$.

4.3 Analysis and Resynthesis Program

The program is designed as an interactive tool for exploratory data analysis. A graphical user interface is incorporated for ease of use. A graphical display of data aids in the selection of optimal parameters to use for envelope extraction and is important for visually assessing the results of envelope reconstruction with principal components. Some of the program features are outlined below.

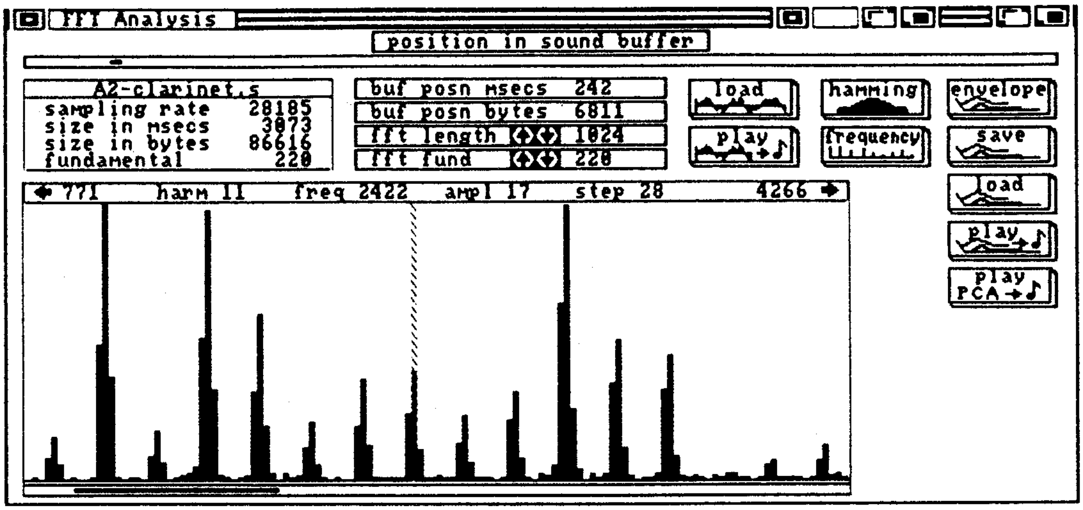
4.3.1 Frequency Resolution

The FFT was implemented with a power of 2 algorithm ($N = 2^i$ where i=1,2,3,...). The window size N can be specified[2] to obtain various frequency resolutions Δf. For a sampling rate of 28,185 samples per second, some practical window sizes and corresponding time and frequency resolutions are: 256 samples (time resolution 9 ms, Δf of 110 Hz), 512 samples (18 ms, 55 Hz), 1,024 samples (36 ms, 27.5 Hz)...4,096 samples (145 ms, 6.9 Hz)...65,536 samples (2,325 ms, 0.4 Hz). An FFT analysis took approximately 3 seconds for a 1,024 sample window.

4.3.2 Window Type

Two choices of window type are available, *rectangular* (square pulse) or *Hamming* (cosine from $-\frac{\pi}{2}$ to $\frac{\pi}{2}$, on a pedestal). The Hamming window compensates for artifacts (leakage error) introduced by the abrupt cutoff of the rectangular window [54, chapter 6].

[2] As the window size N is increased, the frequency resolution improves, however, the amplitude values will then be *averaged* over a longer time span, resulting in poorer time resolution.



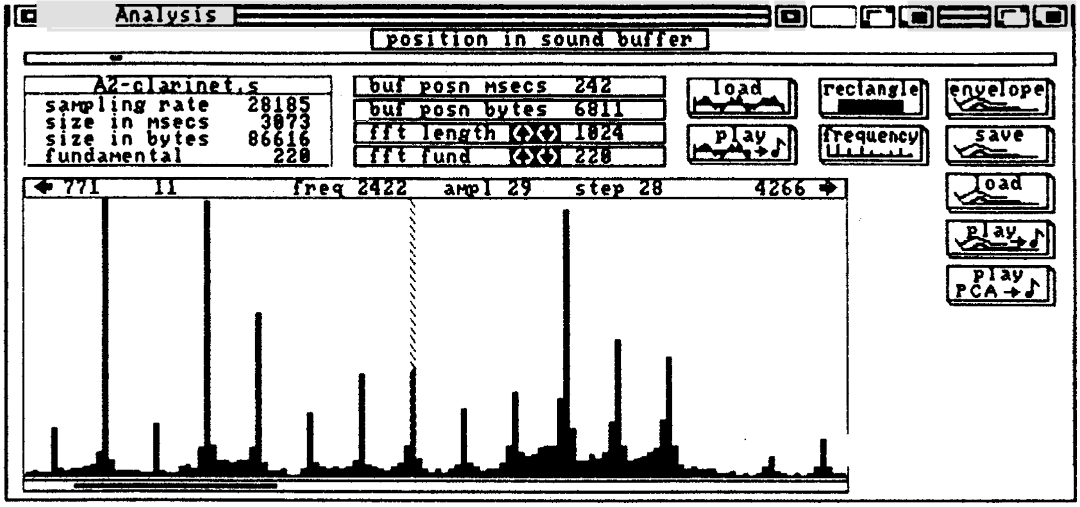
```
A2-clarinet.s
sampling rate      28185
size in msecs       3073
size in bytes      86616
fundamental          220

buf posn msecs    242
buf posn bytes    6811
fft length        1024
fft fund          220

771   harm 11   freq 2422   ampl 17   step 28   4266
```

Figure 4.1: Frequency domain resulting from a Hamming window FFT of an A note (220 Hz) played by an E♭ clarinet. The FFT frequency under the wavy vertical line has been selected with the mouse. Information on this frequency component is displayed immediately above the graph. This discrete FFT frequency is the one closest to the 11th harmonic frequency.

Figure 4.1 (page 57) shows a frequency domain display using a Hamming window and figure 4.2 (page 58) shows a frequency domain display using a rectangular window for the same sample, extraction parameters, and time position. Note the spectral line widening at the base of the rectangular window harmonic components (leakage error) and the narrower base of the Hamming window harmonic components. The rectangular window does have the advantage of "sharper" harmonic peaks. The amplitudes in the Hamming window are also scaled down in comparison to the rectangular window amplitudes, however, the relative amplitudes are not changed. Since the resynthesized sound will have to be scaled to the bit resolution of the playback device, only the relative amplitudes are important.

4.3.3 Time and Frequency Display

The time domain and frequency domain values (for the current FFT window size) can be graphically displayed. In addition, a numerical display of FFT discrete frequency,



A2-clarinet.s

sampling rate 28185
size in msecs 3073
size in bytes 86616
fundamental 220

buf posn msecs 242
buf posn bytes 6811
fft length 1024
fft fund 220

771 11 freq 2422 ampl 29 step 28 4266

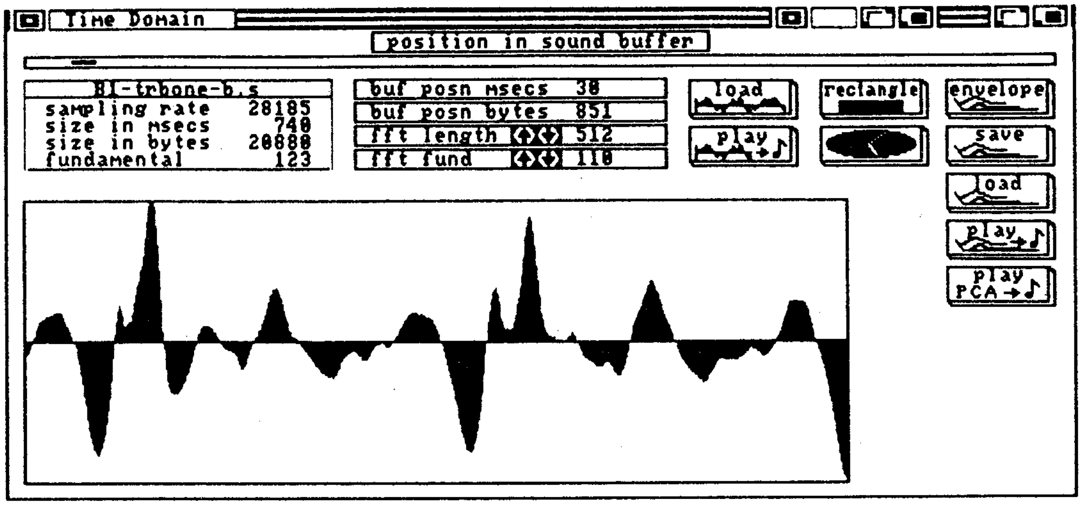
Figure 4.2: Frequency domain resulting from a rectangular window FFT. The frequencies tend to spread out at the base of harmonic peaks, in comparison to the Hamming window (see figure 4.1 on page 57). This is due to the convolution of the window's frequency domain transformation with the near impulse train of the harmonic sound. The frequency domain transformation of the Hamming window produces a narrower center spike with lower side lobes [54, page 141].

amplitude for that frequency, and harmonic number (if applicable), can be obtained by mouse clicking on the bar graph display of the frequency domain data.

Figure 4.2 (page 58) shows a frequency domain display and figure 4.3 (page 59) shows a time domain display. Harmonic frequency components are also highlighted in a different color (not shown in the figures).

4.3.4 Sound Buffer Positioning

Any position in the sound buffer can be selected for analysis by adjusting a pictorial slider consisting of a rectangular "knob" constrained to move in a container (illustrated near the top of figure 4.2 on page 58). The position of the slider knob in the slider container indicates the current position in the buffer. The numerical values of sample and time positions in the buffer are also displayed. The length of the slider knob (with respect to the length of the slider container) reflects the proportion of the



B1-trbone-b.s

sampling rate 28185
size in msecs 740
size in bytes 20888
fundamental 123

buf posn msecs 30
buf posn bytes 851
fft length 512
fft fund 110

Figure 4.3: Time domain display of a Trombone B note (123 Hz) starting at the 30 ms position of a short 740 ms sound. 512 samples are displayed since the current FFT window size is 512. Note the discrepancy between the fundamental frequency (123 Hz) and the closest FFT frequency (110 Hz).

buffer included in the current FFT window. The slider knob also moves through the container while a full FFT scan of the sound is underway in order to see the current analysis position. Precision positioning in the buffer can be accomplished by typing the sample number or time in ms (from the start of the sound) into string (integer) "gadgets".

4.3.5 Fundamental Frequency

A fundamental frequency value can be selected and the FFT frequencies *closest* to the corresponding harmonic frequencies will be highlighted on the frequency domain bar graph display. Alternately, the fundamental frequency of a sound can be stored along with the IFF sample information. This is the default fundamental frequency used if it is available. A simple fundamental frequency extraction algorithm could have been implemented (see Harris and Weiss [28]) but was not required since the fundamental frequencies of all sounds were known.

4.3.6 Sample Resolution

The program will accept 8, 12, or 16 bit samples. Only 8 bit samples were used since 12 or 16 bit digitizers were not available. The 8 bit samples had a noticeably poor signal to noise ratio ($\approx$ 48 dB) but otherwise sounded quite good—probably due to the relatively high sampling rate.

4.3.7 Sound Playback

The Amiga is only capable of playing back 8 bit samples since the 4 built-in digital to analog converters (DACs) are 8 bits (with 6 bit volume controls). The original and resynthesized samples can be played back through the DACs and out to a stereo system through RCA jacks. The playback sampling rate is also restricted to 28,185 samples per second due to DMA limitations (sound output is coprocessor based). Any portion of the sample buffer can be played back (for example the first 15 ms of a sound).

4.3.8 Envelope Extraction

The envelopes are obtained by moving an FFT analysis window through the sound sample[3] and extracting the harmonic amplitudes at each time position. The spacing between envelope points is determined by the window size N, the sampling interval Δt, and the amount of window overlap (spacing $= N\Delta t$ seconds for no overlap). When sounds are later resynthesized with the envelopes, the first envelope time position is placed at $\frac{N\Delta t}{2}$ seconds (amplitude values are linearly interpolated between envelope time positions).

[3] The FFT window can be moved through the sound sample at consecutive positions or in overlapping positions (double, triple, or quadruple).

4.3.9 Inharmonic Partials

An algorithm is used to extract slightly inharmonic partials since it was empirically determined that the average deviation from integral harmonics can be as high as 2% for some instruments ("dead" guitar strings for example).

A large-window FFT analysis (typically 8,192 samples with a Δf of 3.4 Hz, or 16,384 samples with a Δf of 1.7 Hz) is first performed to determine an accurate fundamental and an average difference between partial frequencies (Δh_{avg}). This is accomplished by using the integral harmonic frequency values to search for neighboring peaks of FFT frequencies close to the theoretical harmonic frequencies and then averaging the frequency differences between all harmonics to determine Δh_{avg}.

Δh_{avg} is used to set tentative harmonic frequencies for extraction of harmonic envelopes at a much lower frequency resolution[4] (typically 1,024 samples per window for a Δf of 27.5 Hz). The method follows[5].

A harmonic is searched for in a (settable) range of 10–60% of Δf from the estimated harmonic frequency (both sharp and flat harmonics were searched for). The FFT frequency component with the highest amplitude in this range is chosen as the "harmonic", and its amplitude recorded as the value for that harmonic. The next harmonic is searched for in a similar fashion—by starting at the previously determined harmonic frequency plus Δh_{avg}... and so on.

Note that even when little inharmonicity is present in the sound (Δh_{avg} = the fundamental frequency) the above search procedure is still required since the discrete FFT frequencies do not necessarily coincide with harmonic frequencies and the harmonic frequencies are likely to fluctuate over the sample.

All the numerical harmonic information that is extracted (for a single FFT analysis) can be displayed in order to compare it with the frequency domain bar graph display to see how well the algorithm is working (it works quite well).

[4] A much smaller window size is necessary to achieve decent time resolution for the extraction of harmonic amplitudes as they vary over time.

[5] The algorithms used to search for (slightly) inharmonic frequencies (both in large and "standard" sized FFT analysis windows) were complex, tedious, and heuristic. The details have been omitted.

4.3.10 Frequency Fluctuations

The FFT discrete frequencies closest to the harmonic components are not used for frequency fluctuation values due to the relatively large frequency resolution Δf in the actual envelope extraction[6]. Another method of analysis (heterodyne filter [45] or phase vocoder [47]) would be better suited to determining frequency fluctuations. Since the goal here was to extract a *large* number of amplitude envelopes, the FFT was used.

4.3.11 Number of Harmonics

Any number of harmonics can be extracted (up to the Nyquist limit of the sampling rate). The principal component analysis will accept a different number of harmonics for each sound. The number of harmonics extracted can also be determined by a user specified upper frequency limit.

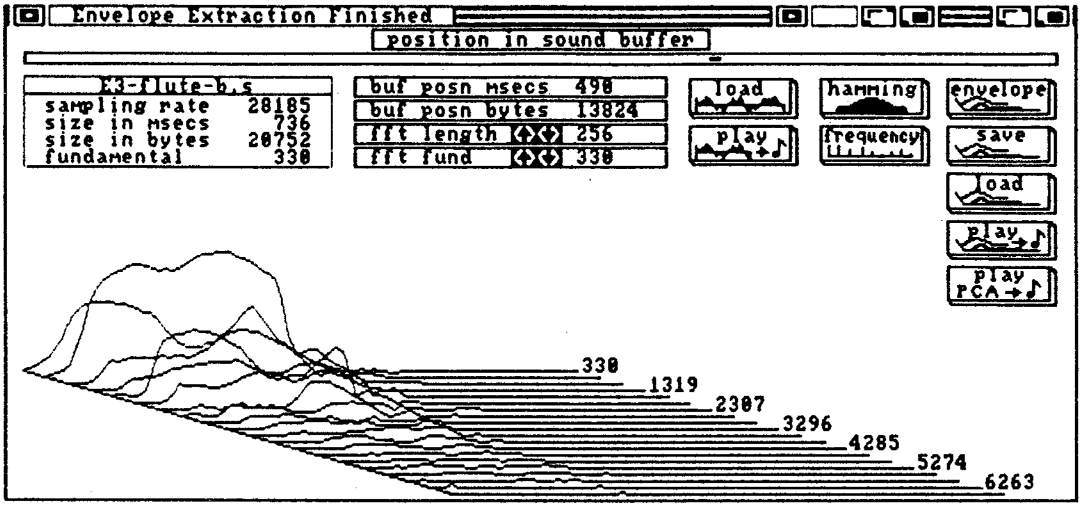
4.3.12 Number of Envelope Points

The maximum number of envelope points analyzed (one point for each FFT window) can be set by the user. Since the principal component analysis uses time positions as variables (see page 71), the same number of envelope points is used for all sounds in a particular principal component analysis (missing values are permitted but were not used).

4.3.13 Dynamic Envelope Display

The envelope curves are displayed as they are extracted. The extraction can be aborted at any time. The harmonic amplitude values (for all harmonics) are displayed

[6] A larger window (with a smaller Δf) would of course allow finer frequency fluctuation assessment, however, to capture the relatively small fluctuations involved would require a window so large that the time resolution would be too large to be useful.



Envelope Extraction Finished

position in sound buffer

E3-flute-b.s

```
sampling rate   28185
size in msecs     736
size in bytes   20752
fundamental       330
```

```
buf posn msecs    490
buf posn bytes  13824
fft length        256
fft fund          330
```

Figure 4.4: The dynamic envelope display just after the envelope extraction has completed. The sound is an E note (330 Hz) on the flute. A 256 sample window was used for the extraction (Δf is 110 Hz). The FFT discrete frequency closest to the fundamental just happens to be exactly the same as the fundamental.

for the first FFT window position in the sound on a "pseudo 3-D" graph. Similarly for the second FFT window position. . . and so on, until the end of the sound sample or the required number of envelope points has been extracted. Envelope curves are also displayed during resynthesis—in this case all of the first harmonic amplitude curve is drawn first (and added into the sample buffer), followed by the second amplitude curve. . . and so on. Figure 4.4 (page 63) illustrates the dynamic envelope display.

4.3.14 Static Envelope Display

When all the amplitude envelopes have been extracted they are displayed on a 16 color "pseudo 3-D" graph (not illustrated). The principal component reconstructed envelopes are displayed beside them (if available) for visual comparison. The principal component basis vectors can be optionally displayed, as well as a 2-dimensional plot of the first 2 principal component basis weights of all harmonic envelopes of a sound.

4.3.15 Automated Envelope Extraction

Envelope data is saved to disk in a standard IFF format (my own). Any number of digitized sounds (subject to disk space constraints) can be analyzed and saved to disk without user intervention (after setting up all the extraction parameters). This takes approximately 8 hours for 50 sounds of 76,800 samples each.

4.3.16 Resynthesis

The resynthesis algorithm reconstructs the sounds from the harmonic envelopes and corresponding harmonic frequencies.

The actual FFT (discrete) frequencies are not used in the reconstruction—only their amplitudes (due to the poor frequency resolution Δf). Instead, the slightly inharmonic frequencies extracted from the initial large-window FFT (with good frequency resolution) are used as the "harmonic" frequencies (see page 61). These frequencies are an *average* over a large part of the sound sample and hence reflect macro-level inharmonicity (if any is present)[7]. Sounds can also be reconstructed using the integral harmonic frequencies.

The harmonic frequencies are kept constant throughout the resynthesis. The section on *Frequency Fluctuations* (page 27) discusses the perceptual impact of using constant frequency harmonics.

A (settable) threshold is established below which harmonic amplitude values will not be included in the resynthesis. This considerably speeds up resynthesis when high harmonics die out quickly (if the harmonic amplitude later rises above the threshold it will be included again).

[7] The major reason for including this macro inharmonicity was that the *pitch* of a sound was affected when inharmonicity greater than $\approx 1\%$ was present (averaged over all harmonics). In this case a digitized sound would be perceived as ***sharp*** (inharmonic partials were usually sharp) with respect to the resynthesized version of it—if the integral harmonics were used. Inharmonic partials were common in steel string guitar sounds. The ***timbre*** did not appear to be affected by resynthesizing with integral harmonics. Unfortunately, resynthesizing with inharmonic partials often resulted in unpleasant beating.

Resynthesis was not done with the inverse FFT since *a)* a lot of information had been discarded and there was no point in computing with zero amplitude frequencies, *b)* it would not have been possible to interpolate amplitude values for each sample, and *c)* an integer math version of the FFT (with variable window size) is difficult to code and possibly inaccurate.

Sounds can be resynthesized with any number of harmonics as long as it is less than or equal to the number of harmonics extracted in the analysis.

4.3.17 Integer Math Resynthesis

Sounds were resynthesized using the following formula,

$$\sum_{k=1}^{M} \sum_{n=0}^{N-1} A_{kn} \sin(2\pi h_k n \Delta t) \tag{4.3}$$

Where, k is the harmonic number, M is the number of harmonics, n is the sample number, N is the total number of samples, A_{kn} is the amplitude value of harmonic k for the *nth* sample (interpolated), h_k is the *kth* harmonic frequency (integral or empirically determined), and Δt is the sampling interval.

Considerable computational saving (with minimal loss in accuracy) can be realized when using integer math[8] with equation 4.3. The sine term can be rewritten as $\sin[(\frac{2\pi}{S})h_k n]$ where S is the sampling rate ($S = \frac{1}{\Delta t}$). If a sine table is pre-computed at $2\pi\Delta t$ intervals with a size equal to S then one full sine period is in the table, sampled at discrete $2\pi\Delta t$ intervals. We can then use $[h_k n \bmod S]$ as an *index* into this table (array) to retrieve the value of $\sin[(\frac{2\pi}{S})h_k n]$ with no sine table sampling error (meaning that an *exact* sine value will be in the table for every possible value of $[(\frac{2\pi}{S})h_k n]$).

The inner loop in equation 4.3 can then be computed by *a)* adding h_k to a running total—each time through the loop—to compute $h_k n$, *b)* taking the modulus of $h_k n$ with respect to S, *c)* retrieving a sine array value using $[h_k n \bmod S]$ as the index, *d)* multiplying the retrieved array value by A_{kn} (A_{kn} is also incremented (or decremented)

[8] The computer that was used had no floating point coprocessor.

by an interpolation factor for each sample), and *e)* adding the result to the previous outer loop value.

Reducing computation in the inner loop of equation 4.3 is important since the loop will be executed MS times for each second of sound (computation is on the order of 10^6 operations for 2 seconds of (20 harmonic) sound at a sampling rate of 28,185 samples per second).

The sine table is set up by computing floating point values of sine at $2\pi\Delta t$ intervals and then scaling these sine values (-1.0 to 1.0) up to suitably large integer values in order to use 32 bit integer math. If a constant sampling rate is used then the sine table needs to be computed only once (computing a 28,185 size sine table takes approximately 30 seconds on the Amiga). Values of A_{kn} are also scaled up and h_k rounded off to an integer value. When all the 32 bit integer samples have been computed with equation 4.3, they are scaled down to the bit resolution of the playback samples (8 bits in this case).

However, even with the computational savings described above, resynthesis is still slow, $\approx$ 1 minute for each second of sound (20 harmonics) at a sampling rate of 28,185 samples per second.

4.3.18 Principal Component Envelopes

Principal component approximated envelopes can be reconstructed by using the basis weights and vectors extracted from the analysis (see equation 4.6 on page 72). An arbitrary set of envelopes can also be reconstructed by manipulating slider "gadgets" which alter the weights of the basis curves (see figure 5.1 on page 90). The new envelope curves can be constructed in real-time and the sound can then be resynthesized (unfortunately nowhere near real-time). These envelope manipulations are discussed on page 89.

4.3.19 Guitar Attack Algorithm

An experimental guitar attack algorithm is available. It is added on top of the harmonic reconstruction as a separate module (see page 95). The algorithm does not use any sound specific analysis information.

4.3.20 Envelope Extraction Parameters

After much experimentation the following envelope extraction parameters were used.

- Sample windows were 1,024 samples in size. This is a good compromise of time (36 ms) and frequency (27.5 Hz) resolutions for sounds with fundamental frequencies varying from 82 Hz to 1318 Hz (the range of sounds analyzed). Analyzing sounds with lower fundamentals naturally presents more of a problem for accurate envelope extraction since harmonics will be tightly packed in the FFT discrete frequencies—which do not necessarily (in fact rarely) coincide with harmonic frequencies.
- A Hamming window was used.
- Windows were not overlapped.
- 20 harmonics were extracted (if they were available).
- 75 envelope points were used for the 2.7 second sounds, and 18 envelope points for the .65 second sounds.
- The true partial frequencies were extracted with a 8,192 sample FFT (Δf of 3.4 Hz).

4.4 Sound Samples

A total of 280 sounds were digitized and analyzed—208 of 2.7 seconds duration and 72 of .65 seconds duration. All sounds were digitized at a sampling rate of 28,185 samples per second with 8 bit resolution.

Room acoustics varied considerably for the sounds. The wind instrument sounds (except the tenor saxophone) were recorded in a highly reverberant room with concrete walls. The guitar, piano, and tenor saxophone were recorded in a less reverberant environment. Room acoustics were dictated by the availability of instrument players.

4.4.1 Wind Instruments

The 2.7 second sounds consisted of 24 each of *trombone*, *E♭ clarinet*, *flute*[9] and *tenor saxophone*, in semitone increments (starting at F (87 Hz) for the trombone, D (147 Hz) for the clarinet, C♯ (277 Hz) for the flute, and B (123 Hz) for the tenor saxophone). These sounds had no decay since the analysis stopped before the end of the sound.

24 sounds of .65 seconds duration were also analyzed for the trombone, E♭ clarinet, and flute (over the same range of notes). The natural decay portion of these sounds was included in the analysis[10].

4.4.2 Piano

49 piano sounds of 2.7 seconds duration were analyzed, ranging from E (82 Hz) to E (1318 Hz) at semitone intervals. Since piano notes have variable decay rates, some of the higher notes had died out before the end of the analysis period (similarly for guitar high notes). This did not pose a problem for reconstructing the sounds with

[9] The flute was only analyzed for the first 1.8 seconds of sound (50 envelope points) since some notes had not been sustained for the full 2.7 seconds when recorded.

[10] The decay portions of the 2.7 second sounds of the wind instruments (trombone, clarinet, tenor saxophone, and flute) were not included in the analysis since the sounds that were recorded varied too much in duration. Wind generated sounds tend to decay rapidly when wind energy to the instrument is removed. It is not anticipated that excluding the decay portion of these sounds will significantly affect a principal component analysis of the harmonic amplitude envelopes (other than to alter the basis shapes towards the end of the sound). Research also indicates that the decay portion of an instrument's sound has little affect on the identification of the instrument (see footnote on page 25). Resynthesizing wind instrument sounds (with decay) with principal component bases could simply taper off the overall amplitude of the sound at the end—a better solution would be to repeat the analyses with sounds recorded with equal durations.

principal component bases and weights (at least when more than 1 basis was used—see page 86).

4.4.3 Guitar

63 guitar sounds of 2.7 seconds duration were analyzed, 9 for the classical guitar and 54 for the steel string guitar. A mixture of old and new strings was used with a variety of picking positions and plectrums (hard pick, fleshy part of the finger, etc.).

Guitar timbre can be substantially altered by the picking position (for example picking near the bridge results in a thin metallic sound and picking near the midpoint position on the string results in a full mellow sound). The picking position effects were quite audible in the original digitized versions—for extreme differences in position—much less so for slight and moderate picking position differences.

It was difficult to detect the plectrum type used to produce a guitar sound in the digitized versions. The difference in sound between old and new strings was also not readily apparent in the digitized sounds. A higher sampling rate or bit resolution would probably make these effects more audible.

4.5 Principal Component Analysis of Envelopes

Once all the harmonic amplitude envelopes had been extracted from the sounds, they were subjected to a principal component analysis (PCA). A variety of instrument groupings were used (see Table 5.1 on page 75). The PCA was done on amplitude envelopes over *time* rather than spectra (see page 48 for a PCA example using spectral curves).

4.5.1 Vector Interpretation of Envelopes

The harmonic amplitude envelopes can be represented as vectors,

$$\vec{a}_k = (a_{k1}, a_{k2}, a_{k3}, \ldots, a_{kn}) \tag{4.4}$$

Where,

- $\vec{a}_k$ is the vector representing the *kth* amplitude envelope.
- The vector component a_{ki} $(i = 1, 2, 3, \ldots, n)$ is the amplitude of the *kth* envelope at the *ith* time position.
- There are a total of n time positions and a value for the harmonic amplitude at each position (a_{k1} is the amplitude of the *kth* harmonic envelope at the start of the sound and a_{kn} is the amplitude of the *kth* harmonic envelope at the end of the sound).

4.5.2 Principal Component Data Matrix

All the envelopes are grouped together (over all harmonics and all sounds being analyzed) into a data matrix which is input to the PCA. The data matrix is,

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \ldots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \ldots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \ldots & a_{3n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{h1} & a_{h2} & a_{h3} & \ldots & a_{hn} \end{pmatrix} \tag{4.5}$$

h is the number of envelope curves being analyzed. The matrix can be quite large. One of the PCA analyses had $n = 75$ and $h = 2{,}354$ (121 sounds, $\approx$ 20 harmonics per sound).

In terms of the geometric interpretation of principal component analysis (see page 48) each row in the data matrix (an envelope "shape") defines a point in an n-dimensional space where each of the n axes represent the harmonic amplitudes at a particular time position in the sound.

4.5.3 Extracting the Principal Components

PCA interprets each *column* in the data matrix as the values of a variable. The column vectors are not statistically independent (amplitude values at different time positions are correlated with each other), hence, the vector space defined by the n column vectors can be reduced in dimension (in a statistical sense).

The PCA computes an n x n covariance matrix C from the data matrix. Solving the determinant equation $|C - LI| = 0$ for L (where I is the identity matrix and L is the diagonal matrix of latent eigenvalue solutions) produces the eigenvalues. The eigenvectors corresponding to the eigenvalues are the *basis* vectors that will be used to reconstruct envelope curves. The size of the eigenvalues yield a measure of the *variance accounted for* by the corresponding eigenvectors (see page 51).

4.5.4 Reducing the Bases Representation

The number of bases (or eigenvectors) that can be produced by the PCA is n—the same as the original set of variables (n time positions). However, as discussed on page 51 a significantly fewer number of bases can be used to reconstruct the original envelope curves (with a small amount of error).

For this study only 5 basis vectors were extracted for each PCA of the envelope data[11] (5 bases typically accounted for 98–99% of the variance for these data sets).

4.5.5 Reconstructing the Envelopes with Weighted Bases

The PCA also produces a set of weights w_{ki} (one for each basis vector i) for each original envelope curve k in the data matrix ($k = 1, 2, 3, \ldots, h$). The original curves can be reconstructed (approximately) from the bases and weights as follows,

[11] The PCA was done on the raw data rather than mean corrected data using SHAZAM (an econometrics statistical analysis package) on the MTS system at Simon Fraser. The following SHAZAM options were used, LIST RAW PEVEC MAXFACT=5.

$$\begin{pmatrix} a'_{k1} \\ a'_{k2} \\ a'_{k3} \\ \vdots \\ a'_{kn} \end{pmatrix} = w_{k1} \begin{pmatrix} b_{11} \\ b_{12} \\ b_{13} \\ \vdots \\ b_{1n} \end{pmatrix} + w_{k2} \begin{pmatrix} b_{21} \\ b_{22} \\ b_{23} \\ \vdots \\ b_{2n} \end{pmatrix} + \cdots + w_{kR} \begin{pmatrix} b_{R1} \\ b_{R2} \\ b_{R3} \\ \vdots \\ b_{Rn} \end{pmatrix}$$

or equivalently,

$$\tilde{a}_k = \sum_{i=1}^{R} w_{ki}\vec{b}_i \tag{4.6}$$

Where,

- $\tilde{a}_k$ is the vector approximation to $\vec{a}_k$ (see equation 4.4 on page 70) and $\tilde{a}_k = (a'_{k1}, a'_{k2}, a'_{k3}, \ldots, a'_{kn})$.
- R is the number of basis vectors used in the reconstruction.
- w_{ki} is the scalar weight for the *ith* basis vector for envelope curve k.
- $\vec{b}_i$ is the *ith* basis vector, $\vec{b}_i = (b_{i1}, b_{i2}, b_{i3}, \ldots, b_{in})$ where b_{i1} is the *ith* basis value at the first envelope time position and b_{in} is the *ith* basis value at the last envelope time position. Figures 5.4 to 5.13 on pages 99 to 103 present a graphical display of some basis vectors.

4.5.6 Alternate Envelope Representation

Note that in the reconstruction of $\tilde{a}_k$ only the weights w_{ki} are specific to that envelope curve. The bases $\vec{b}_i$ are common to *all* the envelope curves. Hence, the approximated envelope curve $\tilde{a}_k$ can be represented by the set of weights w_{ki} $(i = 1, 2, 3, \ldots, R)$ which can also be expressed as a vector $(w_{k1}, w_{k2}, w_{k3}, \ldots, w_{kR})$. If $R = 2$ or $R = 3$ then the weights for a particular envelope curve can be plotted on a 2 or 3-dimensional graph.

This facilitates visualization and comparison of the envelope curves, particularly if some meaning can be attached to the basis vectors. Alternate ways of conceptualizing the weight information (over the sound as a whole) are presented in *Higher Level Control Mechanisms* on page 90.

4.5.7 Representing Envelopes not in the Original Analysis

Reconstructing envelopes with the bases is not limited to envelopes included in the original principal component analysis. For an arbitrary amplitude envelope vector $\vec{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n)$ and a set of R basis vectors $\vec{b_i} = (b_{i1}, b_{i2}, b_{i3}, \ldots, b_{in})$ $(i = 1, 2, 3, \ldots, R)$, if the time positions represented by j in α_j and b_{ij} are the same then taking the dot products of $\vec{\alpha}$ and $\vec{b_i}$ yields the weights w_i needed to compute $\tilde{\alpha}$—the approximation to $\vec{\alpha}$ (see equation 4.6 on page 72).

$$\begin{aligned} w_i &= \vec{\alpha} \cdot \vec{b_i} \\ &= \alpha_1 b_{i1} + \alpha_2 b_{i2} + \alpha_3 b_{i3} + \cdots + \alpha_n b_{in} \end{aligned} \tag{4.7}$$

This simple transformation is a result of the orthogonal nature of the basis vectors. The basis vectors need to be in normalized form[12] in equation 4.7 (each basis element b_{ij} is divided by $\sqrt{b_{i1}^2 + b_{i2}^2 + b_{i3}^2 + \cdots + b_{in}^2}$).

If a sufficiently general set of basis vectors is available then the PCA analysis need not be repeated to included new sounds in the reconstruction procedures. All that is required is that the harmonic amplitude envelopes $\vec{\alpha_m}$ for the sound be known (where m is the number of harmonics) at the same time positions j as the set of bases[13].

[12] Most PCA analysis packages return the basis vectors in normalized form.

[13] The envelope amplitudes could also be interpolated to produce values at the j time intervals.

Chapter 5

Analysis Results and Interpretation

The harmonic amplitude envelopes of 280 musical instrument sounds (see *Sound Samples* on page 67) were extracted and a principal component analysis (PCA) of them was performed for various instrument groupings. Table 5.1 (page 75) lists the different instrument groups analyzed. Initially only individual instruments were included in a PCA. The success of the data reduction of envelope shapes for single instruments suggested that more diverse groups of instruments be included in a PCA.

Table 5.1 also lists the variance accounted for by the first 5 basis vectors of all the instrument groupings (see page 51 for an intuitive explanation of variance accounted for).

Figures 5.4 to 5.13 (pages 99 to 103) show a graphical display of the basis vectors for the various instrument groupings.

Envelope curves are reconstructed by weighting each basis vector (with a value produced by the analysis) and summing the results over the number of basis vectors included in the reconstruction (see equation 4.6 on page 72). The more basis vectors used in the reconstruction the better the approximation to the original envelope curves will be.

Instruments included in each analysis	*Number of sounds per instrument*	*Cumulative % of variance accounted for by the bases*				
		1st	*2nd*	*3rd*	*4th*	*5th*
piano, saxophone clarinet, trombone	49, 24 24, 24	90.4	96.9	97.9	98.4	98.8
saxophone clarinet, trombone	24 24, 24	95.6	97.7	98.3	98.7	99.0
guitar, clarinet	15, 15	91.7	97.5	98.3	98.7	99.0
guitar	63	89.5	97.1	98.5	99.1	99.3
piano	49	89.1	94.1	96.2	97.4	98.2
trombone	24	95.7	97.6	98.3	98.7	99.0
clarinet	24	95.2	97.6	98.1	98.6	98.9
saxophone	24	96.8	98.1	98.8	99.1	99.3
flute[1]	24	98.3	98.9	99.3	99.5	99.6
trombone[2] flute, clarinet	24 24, 24	88.3	93.7	96.1	97.2	98.0

Table 5.1: A principal component analysis of harmonic amplitude envelope curves was performed on various groupings of instruments. The variance accounted for in each analysis is summarized here. Each sound consisted of 20 (analyzed) harmonics (or less if the Nyquist frequency was exceeded) therefore the total number of envelope curves included in each analysis is approximately 20 times the number of sounds in the analysis.

Figures 5.14 to 5.18 (pages 104 to 108) show the degree of approximation resulting from the envelope reconstruction of a clarinet sound with from 1 to 5 basis vectors. Figures 5.19 to 5.21 (pages 109 to 111) show clarinet envelopes reconstructed with alternate bases. Figures 5.22 to 5.28 (pages 112 to 118) illustrate additional envelope reconstructions for guitar, tenor saxophone, piano, trombone, and flute.

[1] Only the first 50 envelope points (1.8 seconds duration) were analyzed for the flute since some of the recorded notes had not been sustained for the full 2.7 seconds.

[2] All the sounds in this group were .65 seconds in duration, with a natural decay. The sounds in all the other groups (except the flute) were 2.7 seconds in duration. The 2.7 second tenor saxophone, clarinet, and trombone sounds, and the 1.8 second flute sounds, had no decay (see *Sound Samples* on page 67).

5.1 Basis Vectors

The basis vectors corresponding to the instrument groupings in Table 5.1 are illustrated in figures 5.4 to 5.13 on pages 99 to 103.

5.1.1 Variance Accounted For

As can be seen in Table 5.1 the variance accounted for is quite high for all instrument groups. The first basis accounts for an average of 93.1% of the variance (over all instrument groups), the second for 96.9%, the third for 98.0%, the fourth for 98.5%, and the fifth for 98.9% (these percentages are cumulative).

The section on *Aural Evaluation* (page 84) discusses the perceptual effect of reconstructing envelopes with different numbers of basis vectors. In general, sounds reconstructed with only a first basis approximation to the envelope curves possess a timbre that is characteristic of the instrument, but with a very uninteresting sound that is distinctly different from the sounds reconstructed from the original envelopes. The perceptual difference between sounds reconstructed with from 2 to 5 bases is much less noticeable. In some cases it is difficult to detect differences in sound when adding in bases after the second.

The high proportion of the variance accounted for by the first basis—and the low quality sound resulting from envelopes reconstructed with one basis—indicates that the variance accounted for does not necessarily translate into meaningful ratios in the perceptual realm (see Zahorian and Rothenberg's work [78] discussed on page 30).

5.1.2 Basis Inversion

Some of the bases have been inverted (along with the corresponding weights) to make interpretation easier. The first basis returned by the PCA analysis was always negative (with all negative weights) and it was common for the third and fifth bases to be negative at the start of the sound (with a mixture of positive and negative weights)[3]. The second to fifth basis vectors were inverted (if necessary) to produce positive spikes at the start of the vectors.

5.1.3 Data Dependent Bases

All the basis sets analyzed (figures 5.4 to 5.13 on pages 99 to 103) indicate that the envelope shape information that is captured is concentrated at the start (attack portion) of the sound. These *data dependent* bases bear a superficial resemblance to other orthogonal basis sets (such as the Fourier sine series) except that the oscillations are crowded into the (perceptually important) onset of the sound, where most of the data variance presumably occurs[4].

5.1.4 Basis Fluctuations

The piano bases (page 99) and guitar bases (page 102) are much smoother in appearance than the bases for the wind instruments (pages 99, 100, and 103). This follows from the generally smoother envelope curves of the piano and guitar sounds (see pages 112 and 115) as compared to the wind instrument envelopes (see pages 106, 114, 116, and 117).

Including more envelopes in a PCA also appears to smooth out the basis curves (compare the 72 sound PCA of sax, clarinet, and trombone sounds (page 101) with the 24 sound PCA for each instrument separately (pages 99 and 100))[5].

[3] For the bases as displayed in figures 5.4 to 5.13 (pages 99 to 103), the first basis weights are always positive (after being "flipped"). The second to fifth bases, when weighted with positive weights, add components to the envelope curves for positive regions of the bases and subtract components for negative regions of the bases. When the second to fifth bases are weighted with negative weights, the converse is true.

[4] Searle compared the amount of information captured with data dependent bases versus data *independent* bases [66] (discussed on page 31).

[5] The smoother bases for the three wind instruments combined may also be due to the averaging effects induced by including more than one instrument in the PCA.

5.1.5 Two Instrument Classes

An inspection of the first basis curves over all groups indicates that there are two distinct sound classes present here—sounds where the energy level can be sustained by the player (wind instruments) and sounds whose decay is not controlled by the player (plucked or struck stringed instruments). The first basis vectors of the latter class of sounds (piano on page 99 and guitar on page 102) taper off to reflect the overall amplitude decay of the sound. In contrast, the first basis vectors for the wind instrument sounds (pages 99, 100, and 103) are relatively constant over the duration of the sound (except for some attack information at the start).

The inclusion of the two classes of instrument sound in one PCA analysis—piano, tenor saxophone, clarinet, and trombone (page 101)—and guitar and clarinet (page 102), produced some interesting results. The sounds resynthesized with these bases did not differ significantly from the sounds resynthesized with the single instrument bases (with the exception of the 1-basis guitar and piano sounds, see *Aural Evaluation* on page 84). It appears that the overall decay information moved into higher order bases when the two instrument classes were combined in one PCA. The two sets of basis vectors on page 101 illustrate the effect of including piano sounds in a PCA of wind instrument sounds.

The basis vectors also replicate the early decay of higher harmonics as can be seen for harmonics 12 to 20 for the guitar (page 112) and harmonic 20 for the clarinet (page 106).

5.1.6 Interpretation of Bases

First Basis

The first basis vector produced by the PCA is based on the averages of the column vectors (time positions) in the data matrix (formula 4.5 on page 70). These averages are normalized to a unit vector. The weights for the first basis vector scale the vector to produce the best overall "fit" to the original curves, hence, this scalar weight is a reasonably good (relative) measure of a harmonic's amplitude averaged over the duration of the sound. A graphical illustration of this can be seen in the 1-basis approximations of figure 5.14 on page 104. This correspondence of the first basis to the overall harmonic amplitude levels indicates that the first basis weights, considered as a whole over all the harmonics in a sound, can be taken as a rough measure of the *spectral energy distribution* of the sound.

While the higher order bases can also alter the amplitude level of harmonics (in the process of adjusting the "shape" of the envelope curve), the effect is slight in comparison to the amplitude factor introduced by the first basis weights. For example, the *average* of the first basis weights (over all 20 harmonics) of the clarinet note depicted on page 105 is approximately 66. The average of the second basis weights is approximately 9 (the absolute values of the weights were used to compute the average). The third and higher-order basis-weight averages decrease (to a lesser extent) from the second basis-weight averages. This is a result of the decreasing variance accounted for by the higher order bases. Note also that second and higher order bases both add and subtract amplitude components at different time periods in the envelopes. These tend to cancel out, leaving the average amplitudes unchanged.

In a sense the first basis weights correspond to the von Helmholtz [75] conjecture that a steady-state, harmonic distribution characterizes the timbre of a sound. Resynthesis with one basis does in fact appear to allow instrument identification but results a dull lifeless sound (see *Aural Evaluation* on page 84).

It is also interesting that the first basis weights yield a measure (spectral energy distribution) that was used to interpret the first dimension of the timbre spaces of Grey [24, 26] and Wessel [77] (see *Grey* on page 36, and *Wessel* on page 39). The set of first basis weights (over all the harmonics of a sound) could be used to compute an approximation to the centroid of the spectral energy distribution.

Second Basis

Inspecting the second basis curves (figures 5.4 to 5.13, pages 99 to 103) indicates that these are "attack shapers." A positive weight for the second basis will in general increase the attack rate and the peak onset-amplitude reached, as well as *subtract* harmonic amplitude from the later portion of the sound. A negative weight for the second basis alters the first basis shape contribution to produce a slower, less intense attack that builds slowly (the negative portion of the second basis is *added* to the contribution of the first basis shape).

The role played by the second basis appears to correspond to the "attack bite" interpretation of Wessel's second timbre space dimension [77] (see *Wessel* on page 39). It may be possible to compute a simple measure of the attack character of a sound using the second basis weights, although more attack information is also spread over the higher order bases (see below).

Third and Higher Order Bases

It is more difficult to interpret third and higher order bases. They appear to perform some of the same function as the second basis—refining the attack portion of a sound. Note that for higher order bases the first peak is reached at progressively earlier time positions.

The higher order bases are also capable of adding and subtracting peaks and valleys at later positions in the sound.

5.2 Envelope Approximations

5.2.1 Clarinet Sound

Figures 5.14 to 5.18 (pages 104 to 108) illustrate the envelope reconstruction for a clarinet sound with 1, 2, 3, 4, and 5 basis vectors. The bases used for these reconstructions were the ones derived from a combination of piano, tenor saxophone, trombone, and clarinet sounds (figure 5.8 on page 101). The clarinet sound was selected for illustration since the original envelope curves exhibit a significant degree of macro and microstructure variation. In addition, the bases used in the reconstruction are the most general ones available.

5.2.2 Macro and Microstructure

No formal criterion is used to distinguish macro and microstructure variation in amplitude envelopes for the following discussion. In general, macrostructure variation occurs over a longer time span than microstructure variation.

Macrostructure

Comparing the 5 figures (pages 104 to 108) reveals that the macrostructure approximations improve when more bases are included in the reconstruction. The 1-basis approximations (using the weights derived from the analysis) capture the overall amplitude for each harmonic averaged over the time period of the envelopes. Each additional basis adds progressively more macroscopic features. This is particularly apparent for harmonics 5, 7, and 11 to 16 in figures 5.14 to 5.18.

Microstructure

The figures reveal that microstructure variation is *not* captured in *any* of the reconstructions. The 5 bases envelope approximations (page 108) are still smooth (although more of the macro variation is included)[6].

A large number of bases would likely be required to approach the detailed microstructure of the original curves[7]. Eliminating the microstructure variation may in fact be desirable since some research indicates that it is not perceptually significant [27, 48, 49, 50, 56] (see *Line Segment Envelope Approximations* on page 18).

Local Features

One of the problems illustrated by the clarinet envelopes is that it is not always clear whether a shape feature should be classified as part of the microstructure (and omitted from the reconstruction) or included as part of the macrostructure. This is the same problem encountered by Strawn [69] (see page 19) in attempting to automate the extraction of line segment approximations to envelopes.

Progressively more local features ("blips", non-monotonic decay, etc.) tend to be included with each additional basis for the PCA envelope approximations. An example of a local feature that is *not* included by the PCA reconstruction (and perhaps should be) is illustrated by harmonic 9 in the 5 bases approximation on page 108. The original envelope curve has a "bump" at approximately a third of the way into the envelope. This bump is not captured by the 5 bases approximation but has instead been smoothed into the overall macrostructure of the envelope curve. Harmonic 14 (page 108) also has a bump at the midpoint envelope position that is missing in the reconstruction.

Since only 5 bases were retained from the PCA, it is not possible to assess the effect of additional bases on the reconstruction of these particular features. Presumably they would be included by the next few bases. Perhaps a more relevant question is: *To*

[6] The first few PCA basis vectors tend to average out any random fluctuations in the original envelope curves. The first basis vector is based on a simple average (over all envelopes) of the amplitude values at each time position. The second basis vector is also based on an average of all envelope time positions, after the influence of the first basis vector has been removed...and so on.

[7] Basis n oscillates $n - 1$ times between positive and negative regions (see figures 5.4 to 5.13, pages 99 to 103). As n increases, the oscillations will be capable of adding more detailed fluctuation to the reconstructed envelopes.

what degree is the omission of these apparent features perceptually significant?[8] This question will be addressed in *Aural Evaluation* on page 84.

5.2.3 Effect of Higher Order Bases

Attack Refinement

The interpretation of the higher order bases as refining the attack portion of the envelopes (page 80) is supported by the reconstructed envelopes of harmonics 11 and 13. The inclusion of the 4th basis (page 107) captures the "blip" at the start of harmonic 11 (which was missing in the 3 bases reconstruction on page 106) and the inclusion of the 5th basis (page 108) introduces the onset blip for harmonic 13.

Non-monotonic Variation

The 4th and 5th bases also appear to include non-monotonic variation in the later portions of the envelopes that is missing from the 3-bases approximations (compare harmonics 5, 11, 14, and 15, on pages 106, 107 and 108)[9].

5.2.4 General versus Instrument Specific Bases

Figure 5.19 (page 109) illustrates the 5 basis reconstruction of the same clarinet sound using the clarinet analysis bases. Comparing figure 5.19 to figure 5.18 (page 108) reveals very little difference between the reconstructed curves. This seems to indicate that a general basis set is feasible for a wide range of instruments.

[8] PCA bases reconstruction (with *all* the bases available) would be a useful tool in assessing the perceptual impact of a variety of local features and degrees of envelope data reduction.

[9] Fletcher's study of the quality of piano tones [20] revealed that piano harmonics can increase in amplitude during the decay and that eliminating this non-monotonic behaviour in resynthesis was not perceptually noticeable. Harmonic 4 for the piano in figure 5.25 (page 115) exhibits this decay pattern.

Figures 5.20 and 5.21 (pages 110 and 111) illustrate the 3 and 5 bases reconstruction of a clarinet sound using the guitar-clarinet analysis bases.

5.2.5 Other Instruments

Figures 5.22 to 5.28 (pages 112 to 118) illustrate envelope reconstruction for guitar, tenor saxophone, piano, trombone, flute, and a short trombone sound.

The flute reconstruction (page 117) is noteworthy since the variance accounted for by the first few bases (see Table 5.1 on page 75) is very high (98.3% for the first basis, 98.9% for the second, etc.). As can be seen in figure 5.27 (page 117), there is very little difference between any of the envelope curves reconstructed with more than 2 bases.

5.2.6 Algorithmic Microstructure

The original envelopes of the wind instruments tend to have considerable microstructure ("jagged" fluctuations) while the piano and guitar do not. These jagged fluctuations appear to be somewhat random, hence, it may be possible to add such amplitude fluctuations (to the PCA reconstructions) in an algorithmic manner, if they are found to have perceptual significance.

5.3 Aural Evaluation

The previous section compared the original harmonic amplitude envelopes with the envelopes reconstructed from PCA basis vectors and weights by visual inspection of the envelope shapes. A better test of the effectiveness of the reconstruction is to compare the sounds produced by both groups of envelopes. Ideally the differences would be assessed by a discrimination measure similar to the one employed by Grey and Moorer [27] and Charbonneau [10]. This testing was not carried out. In lieu of a more rigorous assessment the following subjective evaluation is presented.

5.3.1 Sound Categories for Comparison

The following sound groups were compared for timbre similarity and quality.

- *A*. The original digitized versions of the musical instrument sounds (sampled at 8 bit resolution with a sampling rate of 28,185 samples per second).
- *B*. The sounds resynthesized with *all* the original amplitude envelope information (spaced at 36 ms intervals). Constant integral harmonic frequencies were used[10].
- *C*. The sounds resynthesized with from 2 to 5 basis vectors and weights (derived from the PCA). Constant integral harmonic frequencies were used.
- *D*. The sounds resynthesized with 1 basis vector and weight (per harmonic). Constant integral harmonic frequencies were used.

The sounds resynthesized with 1 basis vector and weight (group *D*) are placed in a separate category since they were easily distinguished from sounds resynthesized with 2 or more bases (group *C*).

5.3.2 Sampled Sound versus Resynthesized Sound

In all cases the most obvious difference in sound quality was between the original digitized sounds in group *A* and *any* of the resynthesized versions (except the 1-basis approximations in group *D*). All the resynthesized sounds (groups *B*, *C*, and *D*) tended to lack subtle attack characteristics of the original digitized sounds. A certain "liveliness" was also missing in many of the resynthesized sounds, most likely due to the fixed harmonic frequencies or the 20 harmonic limit to the resynthesized sounds.

However the timbral quality of the resynthesized sounds in groups *B* and *C* were, for the most part, acceptable—particularly if the sounds were not compared directly to

[10] Slightly inharmonic frequency values were available from the analysis but were not used due to the occasional presence of beating between the harmonics.

the original digitized versions in group A. The exception to this was the piano sounds. Piano sound quality was quite poor for groups B, C, and D, and it is questionable whether the sounds are recognizable as originating from a piano[11]. Since the timbre quality was poor for group B, it is not reasonable to expect an increase in the number of basis vectors, in group C, to improve the sound quality.

5.3.3 Sounds Resynthesized with 1 Basis

Instrument Identification

The sounds produced by 1-basis approximations in group D were surprisingly characteristic of the instrument producing the sound (see the interpretation of the first basis on page 78). Envelopes reconstructed with 1 basis are illustrated in figure 5.14 on page 104 and figure 5.28 on page 118.

Instrument Class and General Bases

The major exception to 1-basis instrument identifiability was for the guitar and piano sounds that were resynthesized with a set of bases derived from a combination of instruments that included the energy-sustained wind instruments (see *Two Instrument Classes* on page 78).

The first basis vector from these analyses (see figure 5.8 on page 101 and figure 5.11 on page 102) does not capture the overall decay characteristics of the guitar and piano sounds. It was difficult to identify the instrument producing the sound due to the lack of characteristic decay. Resynthesis with more than 1 basis did produce acceptable sound for the guitar which indicates that the decay information has moved into higher-order bases. The guitar sounds reconstructed with 1 basis, using the *guitar* analysis

[11] The poor resynthesized piano sound is probably due to *a)* inharmonicity caused by radically different string thicknesses and lengths, *b)* multiple strings per note—resulting in subtle beating effects (good piano tuners slightly mistune the unisons), *c)* complex resonances of the large soundboard, and *d)* complex sympathetic vibrations induced by a common bridge for several hundred strings.

bases (figure 5.10 on page 102) were identifiable as guitar sounds[12]. Resynthesis of piano sounds with more than 1 basis produced a sound that was similar to the sounds resulting from resynthesis with all the envelope information (poor).

Sound Quality

All sounds resynthesized with a 1-basis approximation (group D) were noticeably different from the same sounds resynthesized with all the envelope information (group B) as well as those sounds resynthesized with more than 1 basis (group C). The 1-basis sound was lifeless and unchanging, as would be expected given that all harmonic envelopes had the same "shape." The trombone and guitar sounds (with the guitar specific bases) produced by 1-basis approximations fared better than the other instruments.

5.3.4 Sounds Resynthesized with 2 or More Bases

The envelopes produced by 2, 3, 4, and 5 bases approximations capture progressively more of the macrostructure of the original envelopes. This is illustrated in figures 5.15 to 5.18 (pages 105 to 108) and discussed on page 81. However, the *perceptual* difference between 2 to 5 bases approximations (within group C) is not as pronounced as a visual inspection of the envelopes would suggest.

Guitar and Piano

For the guitar and piano, the difference between 2 to 5 basis approximations (within group C) was minimal or non-existent. In addition, a comparison of group C sounds to the group B sounds (all the envelope information included) revealed little or no difference.

Figures 5.22 and 5.23 (pages 112 and 113) illustrate the envelope curves for 2 and

[12] Higher guitar notes that died out rapidly tended to sustain too long with a 1-basis approximation. Two or more bases approximations captured the early decay.

5 bases guitar sounds. Figure 5.25 (page 115) illustrates the envelope curves for 2, 3, and 5 bases piano sounds.

Wind Instruments

The wind instruments did exhibit a noticeable difference between group *C* sounds and group *B* sounds. The group *B* sounds "fluctuated" in a manner that suggested the player's interaction with the instrument. The sound was uneven, probably as a result of the 'jaggedness" of the envelopes (see figure 5.18 page 108). Increasing the number of bases to 5 in group *C* had only a slight effect on the sound and did not capture the "unevenness" in the group *B* versions. Except for this unevenness the timbre quality was almost identical between group *B* and *C* sounds.

5.3.5 Perception of Microstructure

Given the perceptual differences between the group *B* and *C* wind instrument sounds, a tentative conclusion is that the microstructure envelope variation *does* have a perceptible impact—not so much on the timbre as on the sense of a player interacting with the instrument. The sounds in group *C* sounded "perfect" and to some extent not as interesting as the uneven sounds in group *B*. As discussed on page 84, it may be possible to replicate this effect algorithmically.

Several previous studies have concluded that microstructure envelope variation is *not* perceptually significant [27, 48, 49, 50, 56]. The sounds included in these studies were short (less than half a second) and most of the sounds included in this study were considerably longer (2.7 seconds). It may be the case that microstructure variation is significantly more noticeable in longer sounds.

5.4 Creating New Sounds with PCA Envelopes

The PCA basis vector representation of harmonic amplitude envelopes stores envelope information in a parsimonious, standardized format. The representation also allows the expression of additional sounds, not included in the original analysis, in terms of weighted basis vectors (see *Representing Envelopes not in the Original Analysis* on page 73).

The original sounds can be replicated (with some degree of approximation) by reconstructing the envelopes with equation 4.6 on page 72 and resynthesizing the sound with formula 4.3 on page 65.

5.4.1 Low Level Envelope Manipulation

It is also possible to *alter* the sound by manipulating the weights associated with each basis and harmonic. Figure 5.1 on page 90 illustrates a user interface for the manipulation of harmonic amplitude envelopes.

This form of envelope manipulation is suitable for low-level control of sound reconstruction. The major deficiency with this scheme (figure 5.1) is that the envelopes are manipulated individually. From the perspective of a sound designer, it would be advantageous to have access to higher-level control mechanisms that take into consideration the harmonic envelopes as a whole.

The simplest method of increasing high-level control would be to manipulate the harmonic envelopes as a group, for example, using the basis weight sliders in figure 5.1 to alter *all* the envelopes in unison, or just the odd harmonics, etc. This method is too coarse for producing realistic acoustic instrument sounds. However, it would be useful for roughing in a sound before fine-tuning the envelopes with the interface illustrated in figure 5.1.

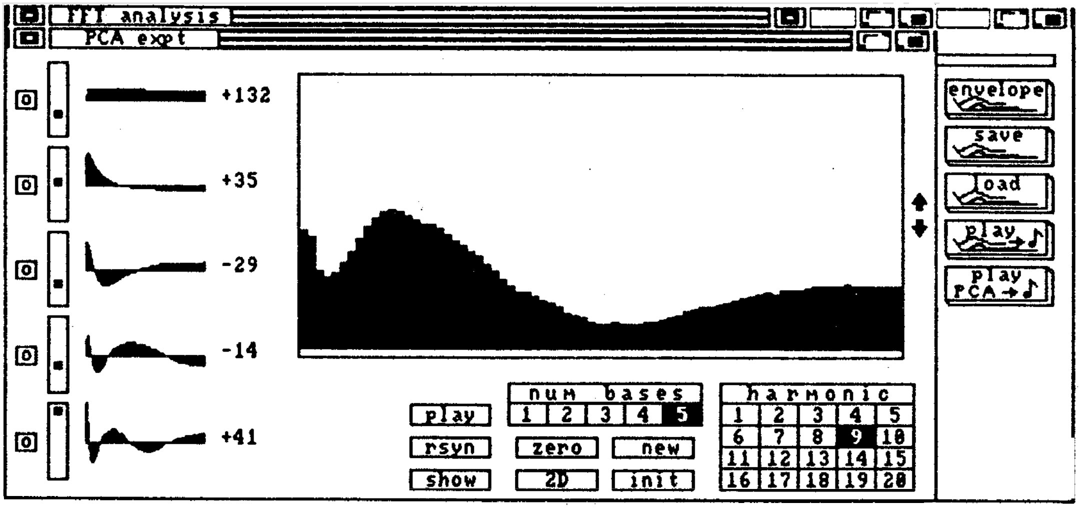


Figure 5.1: A user interface for constructing harmonic amplitude envelopes by manipulating PCA basis weights. The 5 vertical sliders on the left alter the weights for the basis vectors displayed immediately to their right. The envelope shape is altered continuously as the slider is moved. From 1 to 5 basis vectors can be used to reconstruct the amplitude envelope of any harmonic. The envelopes for a particular sound (for example a trombone F♯ note) can be selected as a starting point. The effect of the envelope alterations can be assessed by resynthesizing the sound.

5.5 Higher Level Control Mechanisms

The advantage of the PCA basis representation of envelopes is that all envelopes have a common underlying representation (weighted basis vectors). This property can be exploited in the generation of higher-level control schemes. The following sections suggest some possible strategies for the development of higher-level control mechanisms. They have not been implemented.

5.5.1 2-Dimensional Weight Graphs

Figure 5.2, on page 91 graphs the first two basis weights of a clarinet sound. Since additional bases add little to the perceptual qualities of the sound (see page 88), the information displayed in figure 5.2 captures a significant amount of the relevant

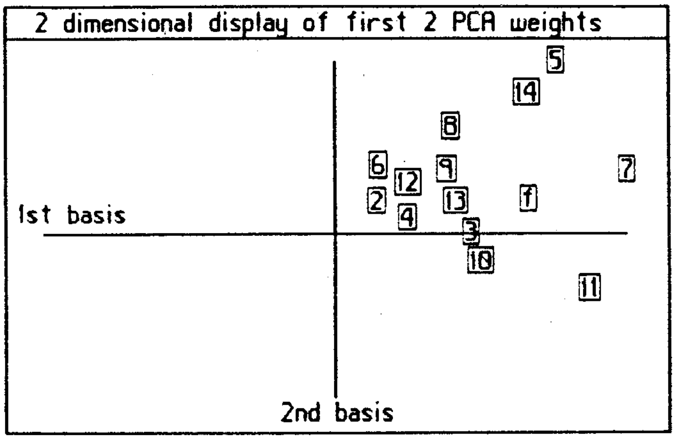


Figure 5.2: A 2-dimensional graph of the first two basis weights for the clarinet A (220 Hz) note depicted in figure 5.15 (page 105). Only the first 14 harmonics are displayed (harmonics 15 to 20 overlap the first 14). Different scales are used for the 2 axes.

envelope information for the sound.

Even though all the envelope information for the sound is displayed on one graph, it is difficult to interpret it since the harmonics are scattered over the graph in no particular order. Manipulating the first two basis weights in this format (by setting envelope points in a 2-dimensional grid) is not intuitive and not likely to be useful as a means of higher-level control. Futhermore, the placement of harmonic points in this 2-dimensional space is *unconstrained* and it would be difficult to implement meaningful 2-dimensional constraints on the placement of harmonics points.

Since the first two bases represent different aspects of envelope "shape"—and have different interpretations (see *Interpretation of Bases* on page 78)—it may be more useful to display weight information on 1-dimensional graphs.

5.5.2 1-Dimensional Weight Graphs

First Basis Weights

Figures 5.29 to 5.31 (pages 119 to 121) display the first basis weights (for all harmonics) for three different sounds. Figures 5.29 and 5.30 are directly comparable since the underlying bases are the same.

Given the interpretation of the first basis (page 78), these graphs can be taken as a rough approximation to the "steady-state" spectrum for the sounds illustrated. The graphs illustrate the overall relationships between the harmonics of a sound (see *Interharmonic Relationships* on page 20). The characteristic clarinet emphasis on odd harmonics is readily seen in figure 5.29.

Second Basis Weights

Figures 5.32 to 5.34 (pages 122 to 124) display the second basis weights (for all harmonics) for the same three sounds illustrated in figures 5.29 to 5.31.

Given the interpretation of the second basis (page 78) these graphs can be taken as a rough approximation to the degree of "attack bite" for the sounds illustrated. Note that the weights can be positive or negative. Negative weights decrease the "attack bite", result in lower rise times, and increase the harmonic amplitude later in the sound.

Manipulating the Graphs

Manipulating the harmonic amplitude envelopes for a sound would appear to be somewhat more intuitive when the weight information (for one basis) for all envelopes is displayed on the same graph. All of the envelope weights can be compared directly and the sets of weights for the first two bases have interpretations that translate into simple perceptual effects (see *Interpretation of Bases* on page 78, and Wessel's work on *Timbre Space* [77], described on page 39).

Similar graphs for the higher order bases would probably not be as useful.

An implementation might include a graphical interface that allows a user to interactively construct the weight curves (over all harmonics) for each basis (as in figures 5.29 to 5.34 on pages 119 to 124).

Constraining the Selection of Weights

One problem with the interface suggested above is that the form that the weight curves can take would still not be constrained in a useful manner[13]. A possible solution is to incorporate limits on the range of values that the weights can assume, in order to stay within the timbre range of a particular instrument. For example, a maximum and minimum value for the weight of each harmonic could be plotted on a graph similar to figure 5.29 (page 119). Connecting adjacent (harmonic) maximum weights, and adjacent minimum weights, and highlighting the enclosed regions, would give a sound designer an overview of a reasonable set of weights to use. The maximum and minimum weight values would be taken from the actual weights of analyzed sounds for a particular instrument—ranging over note register, intensity, etc. Figure 5.35 on page 125 shows a hypothetical example.

A simple scatter plot of weights (for each harmonic) could supply additional information as to the distribution of analyzed weights for a particular range of sounds.

Color Coded Constraints

This application is also well suited to the visual display of information using color. For example, shading of the acceptable regions for a set of weights might color the average weight path in a darker shade with progressively lighter shading towards the maxima and minima extremes. Shading intensity could also be varied (using standard

[13] Experience with constructing envelopes with the low-level interface illustrated in figure 5.1 (page 90), and then listening to the resulting sound, reveals that the bases are quite general and that a wide variety of sounds can be constructed. The characteristic timbre of a particular instrument can be quickly lost when manipulating the basis weights of the envelopes.

deviations as units) to reflect the distribution of the weight values. A particular sound (e.g. a G♯ clarinet sound) could be plotted over top of the distribution to act as a reference point.

Interpolation between instrument timbre could be facilitated by color coding acceptable ranges for different instruments, all displayed on the same graph. Assessing the impact of different parameters on the basis weights (such as note intensity, picking position on a guitar string, etc.) could also benefit from a color coded display.

Range of Sounds

A preliminary investigation reveals that the range of sounds included for display on the same graph would have to carefully selected. For example, the first basis weights of the fundamental envelopes of the 24 clarinet sounds included in the piano-sax-clarinet-trombone analysis (see Table 5.1 on page 75) have a mean of 199, a standard deviation of 121, a maximum of 508, and a minimum of 77. This imposes very little constraint on the range of first basis weights for fundamental envelopes.

However these 24 sounds range over 2 octaves on the clarinet. The variation in fundamental weights is much more constrained when the sounds are grouped by octave. The equivalent statistics for sounds limited to the 12 notes of the clarinet's first octave are; a mean of 120, a standard deviation of 28, a maximum of 167, and a minimum of 77.

This clearly illustrates both the degree to which the envelopes change for different registers of the same instrument and the problems faced in attempting to characterize an instrument's timbre (as a whole) in terms of the harmonic amplitude envelopes.

A variety of other timbre control and manipulation schemes are no doubt possible with a bases representation of envelopes. The underlying representation and construction mechanisms could be made apparent or hidden from the user. Low-level control of the envelopes (as in figure 5.1 on page 90) would still be available for fine-tuning the sound.

5.6 Guitar Attack Algorithm

Capturing the subtleties of the attack portion of a sound for *any* of the instruments included in this study proved difficult. There was simply much more going on at the start of a sound than could be captured with a Fourier analysis, given the frequency/time resolution tradeoff (see *Frequency Resolution* on page 56).

The resynthesized guitar attack, in particular, lacked the onset "bite" of the original digitized sound. This was most prevalent for notes in the upper registers of the instrument. Adding random frequency components at the start of the sound did not replicate the "low frequency clunk" characteristic of the original onset transients.

5.6.1 Air and Top Resonances

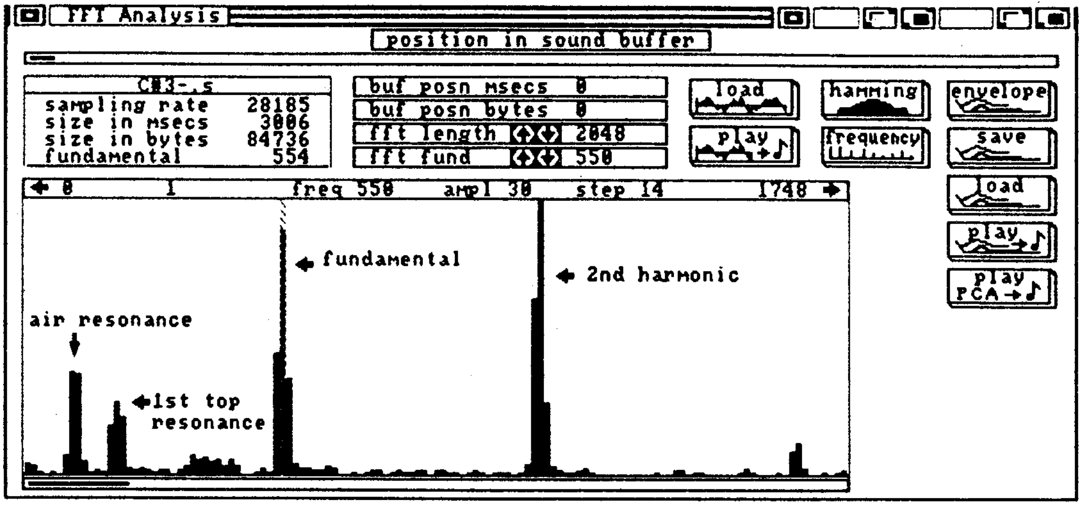
An examination of the frequency domain display at the beginning of guitar sounds revealed inharmonic components of fixed frequency. Since the fixed frequencies were in the range of the air resonance and major top resonance of the acoustic guitar[14], further experiments were performed to get accurate frequency readings for these components.

A large-window FFT (32,768 samples with a frequency resolution Δf of .86 Hz) revealed three fixed inharmonic frequency components[15] present in many guitar sounds of widely different fundamental frequencies. The inharmonic frequencies were 103 Hz, 195 Hz, and 385 Hz.

Figure 5.3 on page 96 shows a frequency display that includes the air and first top resonance of a guitar sound. The second top resonance (385 Hz) is between the first top resonance and the fundamental but has very low amplitude in this example.

[14] This information was obtained from previous experience as a luthier.

[15] These inharmonic components were discernible but of very low amplitude since the inharmonic onset transients tended to die out quickly and more than 1 second of sound was involved in this frequency analysis.



FFT Analysis

C#3-.s
sampling rate 28185
size in msecs 3006
size in bytes 84736
fundamental 554

buf posn msecs 0
buf posn bytes 0
fft length 2048
fft fund 550

0 1 freq 550 ampl 30 step 14 1748

Figure 5.3: A frequency domain display of a C♯ (554 Hz) guitar sound illustrating the inharmonic air resonance and first top resonance. These resonances typically appear at the onset of a guitar sound and quickly decay. The frequency analysis was performed over the first 72 ms of the sound.

5.6.2 Verifying the Air Resonance

The first inharmonic component (103 Hz) was verified as the air resonance of the instrument by humming into the instrument to activate the air resonance (the instrument "comes alive" at the air resonance frequency) and digitizing the resulting sound. A frequency analysis (with a small Δf) of the humming sound revealed a pronounced peak at 103 Hz. It was not possible to verify the other two frequencies as top resonances (there are usually several[16]), however, the first major top resonance for the guitar is typically in the 175–220 Hz range.

5.6.3 Resonance Characteristics

The envelope extraction algorithm (see *Inharmonic Partials* on page 61) was expanded to extract and store the three resonances. Graphing the resonance amplitudes over time revealed a roughly exponential decay for these components, tapering to zero

[16] See Hutchins [31] for a graphical display of the top resonances (eigenmodes) of violin plates.

amplitude in approximately 100–300 ms. The amplitude values of the resonances at the onset of the sound varied considerably—anywhere from 10–60% of the initial fundamental amplitude. It was not possible to distinguish resonance components from harmonic components for sounds with harmonic frequencies close to the resonance frequencies.

5.6.4 Resonance Hypothesis

These inharmonic frequency components were unexpected since they are not part of the harmonic structure of the string vibrations. They are somewhat different from formant frequencies since they do not merely reinforce certain frequency ranges but actively contribute to the sound, irrespective of the note being played on the instrument. Since the resonances appear to die out exponentially, a tentative hypothesis as to how they are produced is the following:

When the string is first struck, the motion (and resulting frequency content) is chaotic. This broadband energy manages to excite the air and top resonances of the instrument via energy transferred to the instrument by the bridge connecting the strings to the top. The string vibrations quickly stabilize into simple harmonic motion, leaving the air and top inharmonic vibrations to die out naturally, at a damping rate determined by the physical characteristics of the instrument. In other words, the instrument body behaves like a *bell* at the onset of the sound (with considerably higher damping than a metal bell).

5.6.5 Resynthesizing the Resonance Components

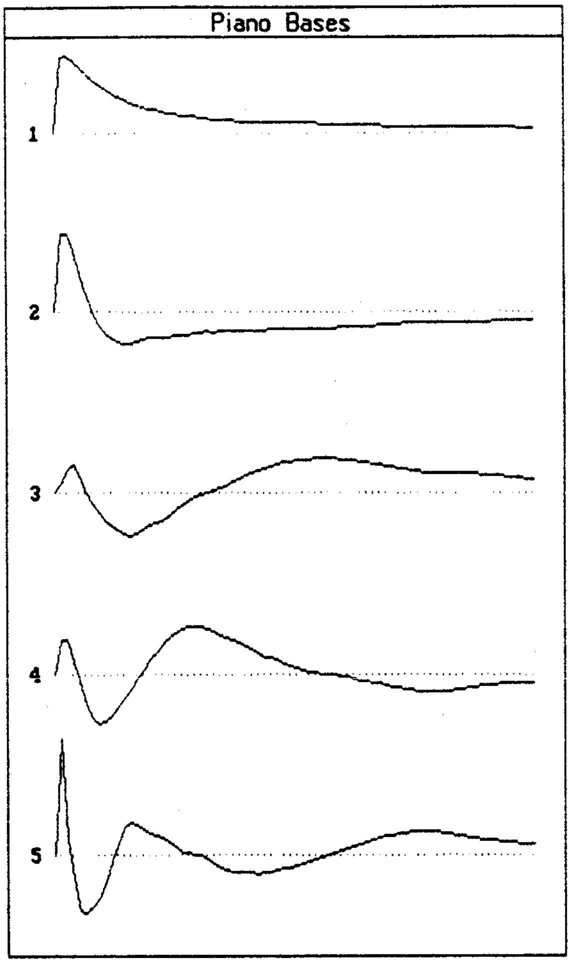
Attempts were made to resynthesize these attack resonance components. The first method simply used the resonance frequency information extracted from the analysis by adding the resonance amplitude envelopes (at the resonance frequencies) to the harmonic components.

The second method reconstructed this information as a separate module added on top of the harmonic resynthesis—using no analysis information—by specifying the initial amplitude of the air resonance (as a percentage of the initial fundamental amplitude) and the decay time. The top resonance amplitudes were scaled down from the stronger air resonance component. The exponential decay was approximated with three straight line segments. A reasonable setting for the intial amplitude of the air resonance was in the 10–30% range, with a decay time of 150–200 ms.

The resynthesized sounds captured to some degree the flavor of the attack in the digitized versions. Although the resynthesized attack sounds were not as full or as rich as the original, the results do appear to support the hypothesis that a significant portion of the attack qualities in guitar sounds is produced by a bell-like decay of the instrument resonances.

The initial amplitude setting and decay time were critical for blending the inharmonic resonance components into the overall sound. Too high an amplitude setting or too long a decay resulted in an unnatural, distracting quality in the attack portion of the sound.

The algorithm could benefit from the inclusion of additional top resonance components, however, it was difficult to discern distinct top resonances beyond the second one (385 Hz). Higher top resonances are most likely weak but numerous. It may be possible to simply choose arbitrary inharmonic values for these resonances.



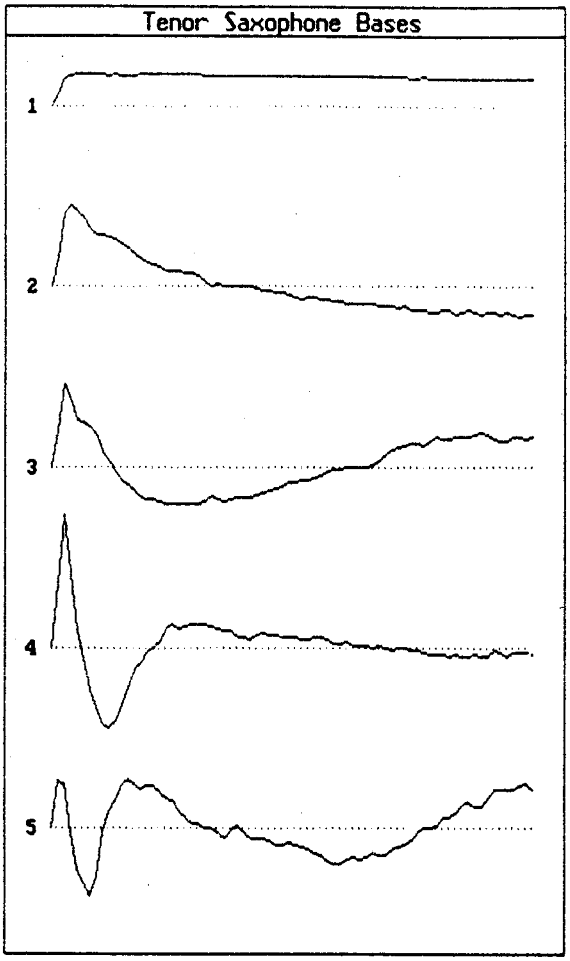
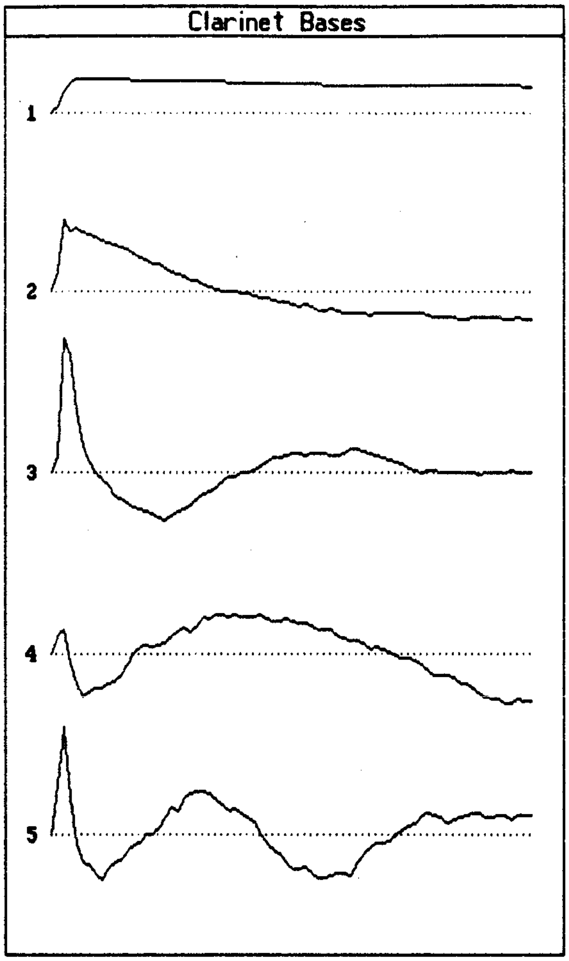


Figure 5.4: (left) **Principal component bases for piano**. The first 5 principal component bases for the harmonic amplitude envelopes of 49 piano sounds (75 envelope points—2.7 seconds of sound).

Figure 5.5: (right) **Principal component bases for tenor saxophone**. The first 5 principal component bases for the harmonic amplitude envelopes of 24 tenor saxophone sounds (75 envelope points—2.7 seconds of sound with no decay).



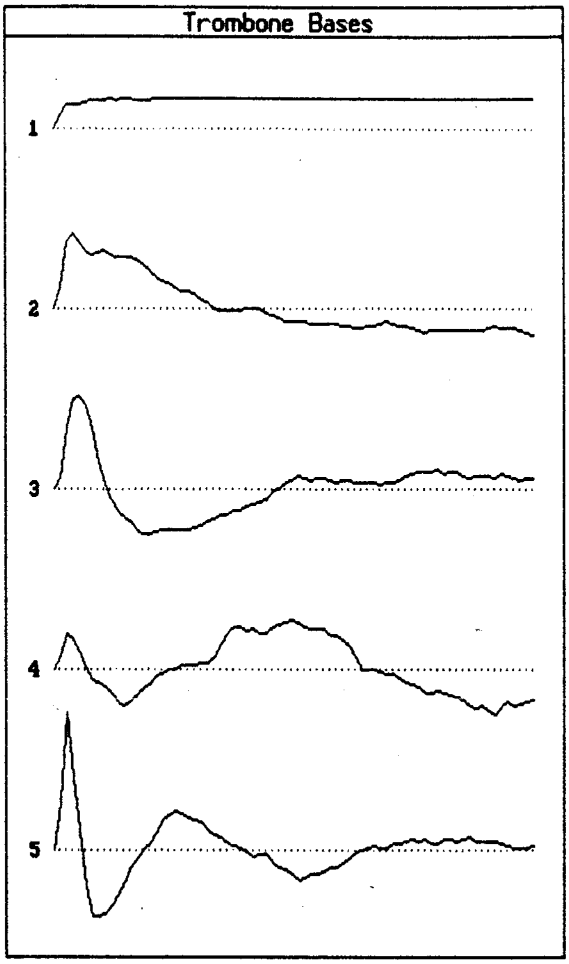
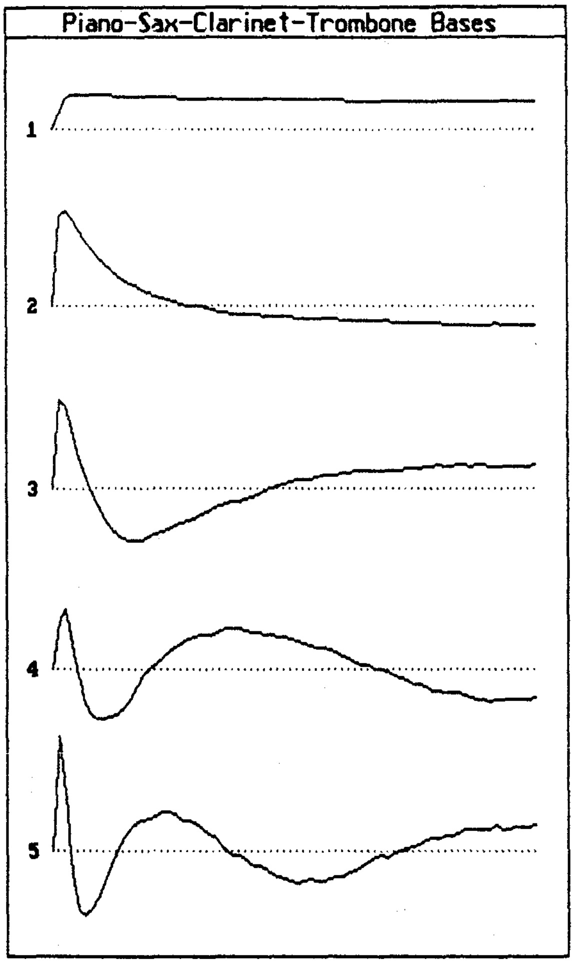


Figure 5.6: (left) **Principal component bases for clarinet**. The first 5 principal component bases for the harmonic amplitude envelopes of 24 clarinet sounds (75 envelope points—2.7 seconds of sound with no decay).

Figure 5.7: (right) **Principal component bases for trombone**. The first 5 principal component bases for the harmonic amplitude envelopes of 24 trombone sounds (75 envelope points—2.7 seconds of sound with no decay).



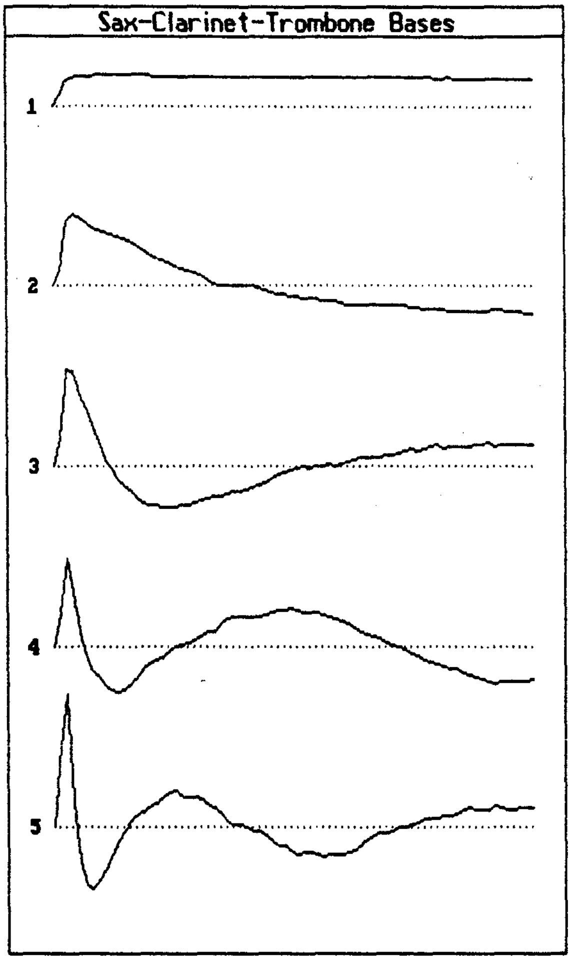
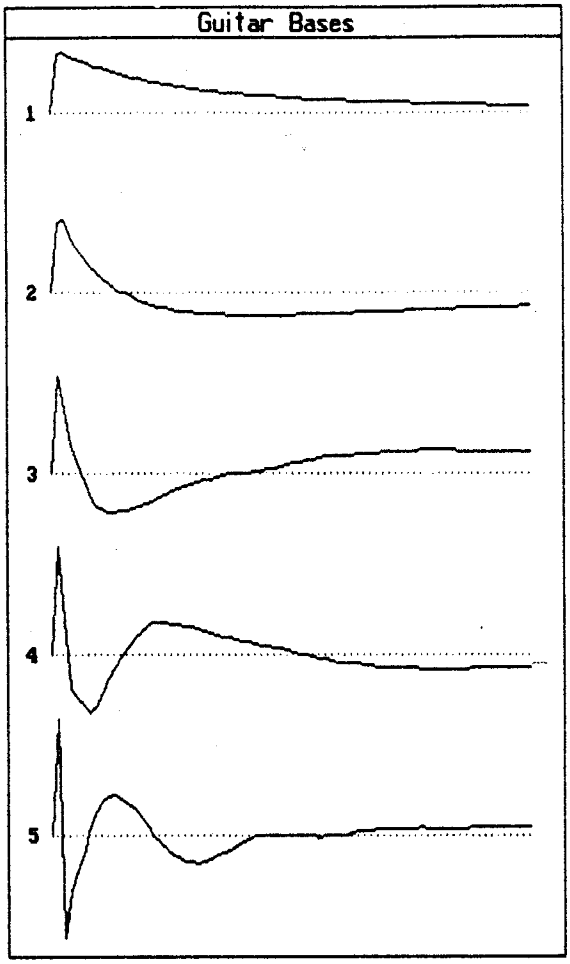


Figure 5.8: (left) **Principal component bases for piano-sax-clarinet-trombone**. The first 5 principal component bases for the harmonic amplitude envelopes of 24 sounds each of tenor saxophone, clarinet, and trombone (with no decay), as well as 49 piano sounds (75 envelope points—2.7 seconds of sound).

Figure 5.9: (right) **Principal component bases for sax-clarinet-trombone**. The first 5 principal component bases for the harmonic amplitude envelopes of 24 sounds each of tenor saxophone, clarinet, and trombone (75 envelope points—2.7 seconds of sound with no decay).



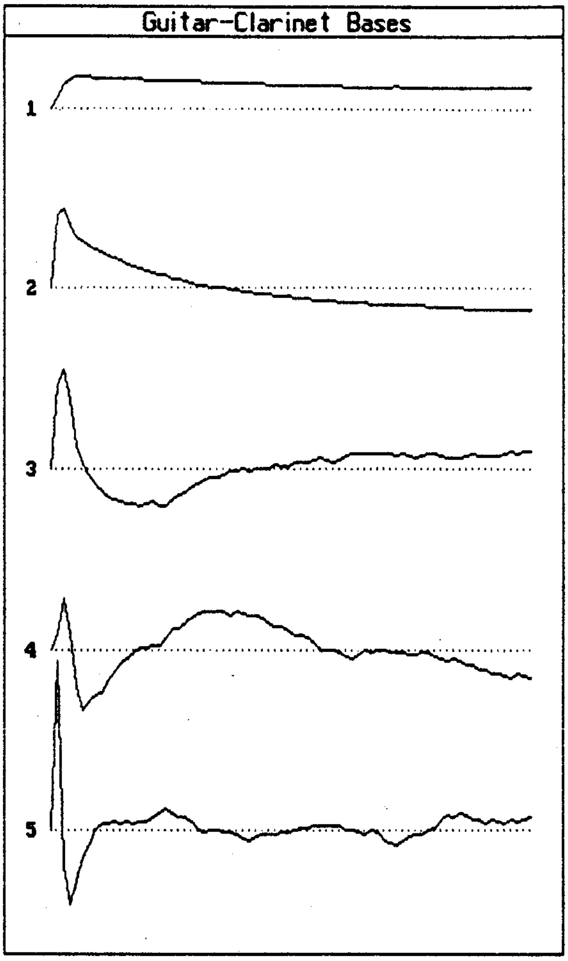


Figure 5.10: (left) **Principal component bases for acoustic guitar**. The first 5 principal component bases for the harmonic amplitude envelopes of 63 acoustic guitar sounds (75 envelope points—2.7 seconds of sound).

Figure 5.11: (right) **Principal component bases for guitar-clarinet**. The first 5 principal component bases for the harmonic amplitude envelopes of 15 clarinet sounds (with no decay), and 15 acoustic guitar sounds (75 envelope points—2.7 seconds of sound).

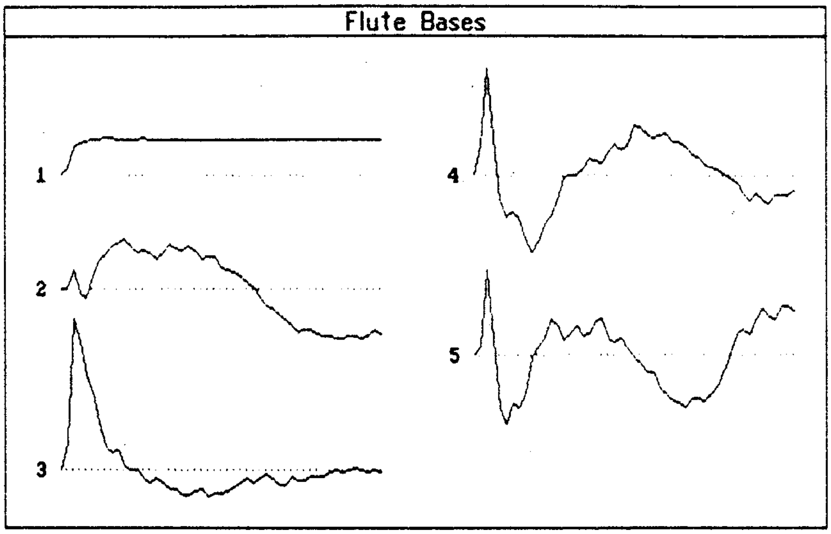


Figure 5.12: **Principal component bases for flute.** The first 5 principal component bases for the harmonic amplitude envelopes of 24 flute sounds (50 envelope points—1.8 seconds of sound with no decay).

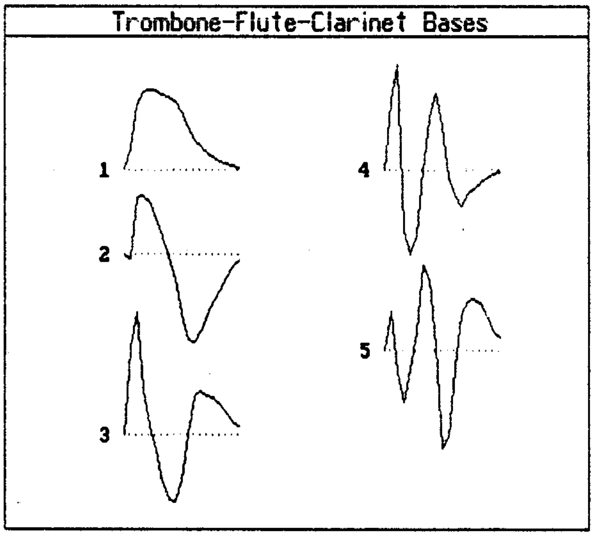


Figure 5.13: **Principal component bases for trombone-clarinet-flute (short sounds).** The first 5 principal component bases for the harmonic amplitude envelopes of 24 sounds each of trombone, clarinet, and flute (18 envelope points—.65 seconds of sound with natural decay).

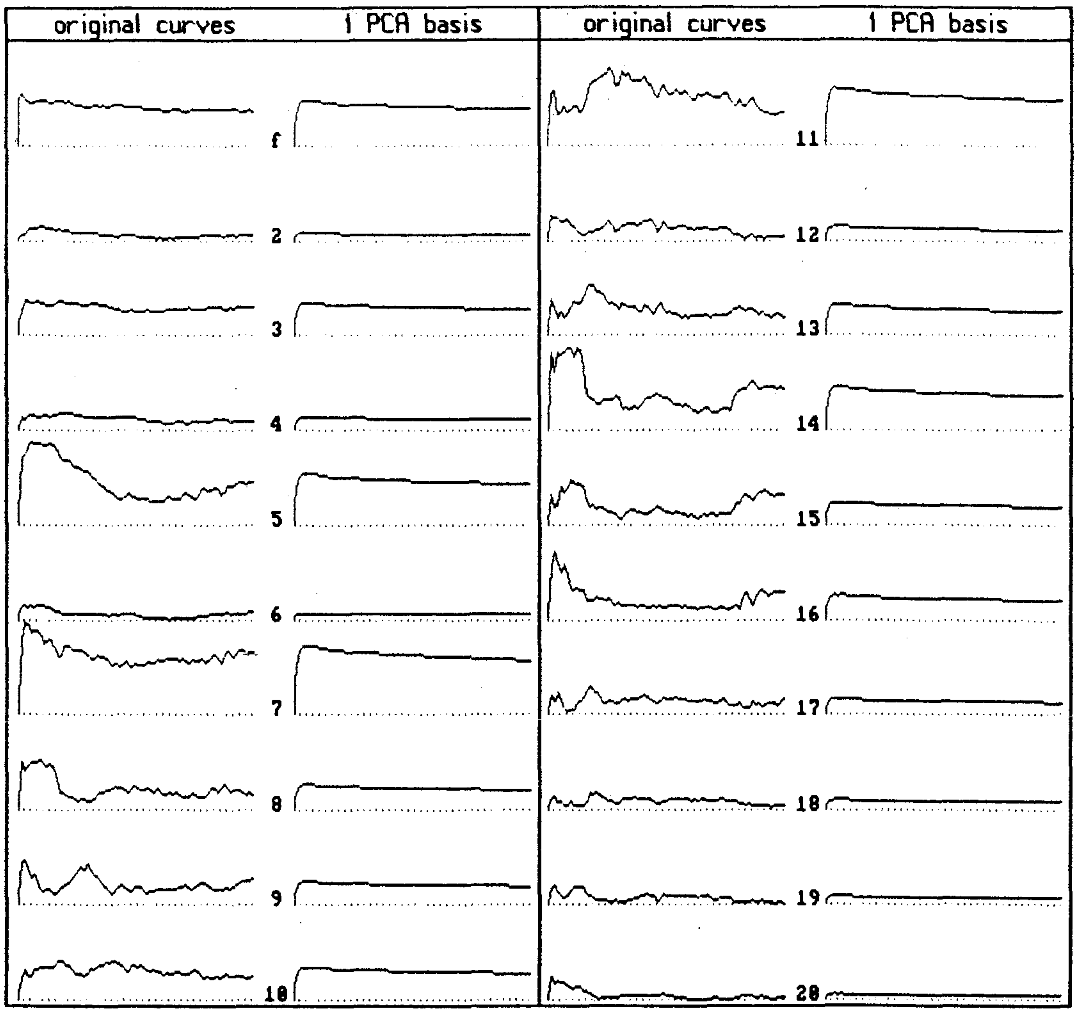


Figure 5.14: **Envelope reconstruction of a clarinet A note with 1 PCA basis.** The original envelopes for the first 20 harmonics of an A (220 Hz) clarinet sound are on the left half of the above graphs. The envelopes reconstructed with 1 PCA basis vector and weight for the same sound are on the right. The basis used was from the piano-sax-trombone-clarinet PCA analysis (figure 5.8 on page 101). The time span is 2.7 seconds.

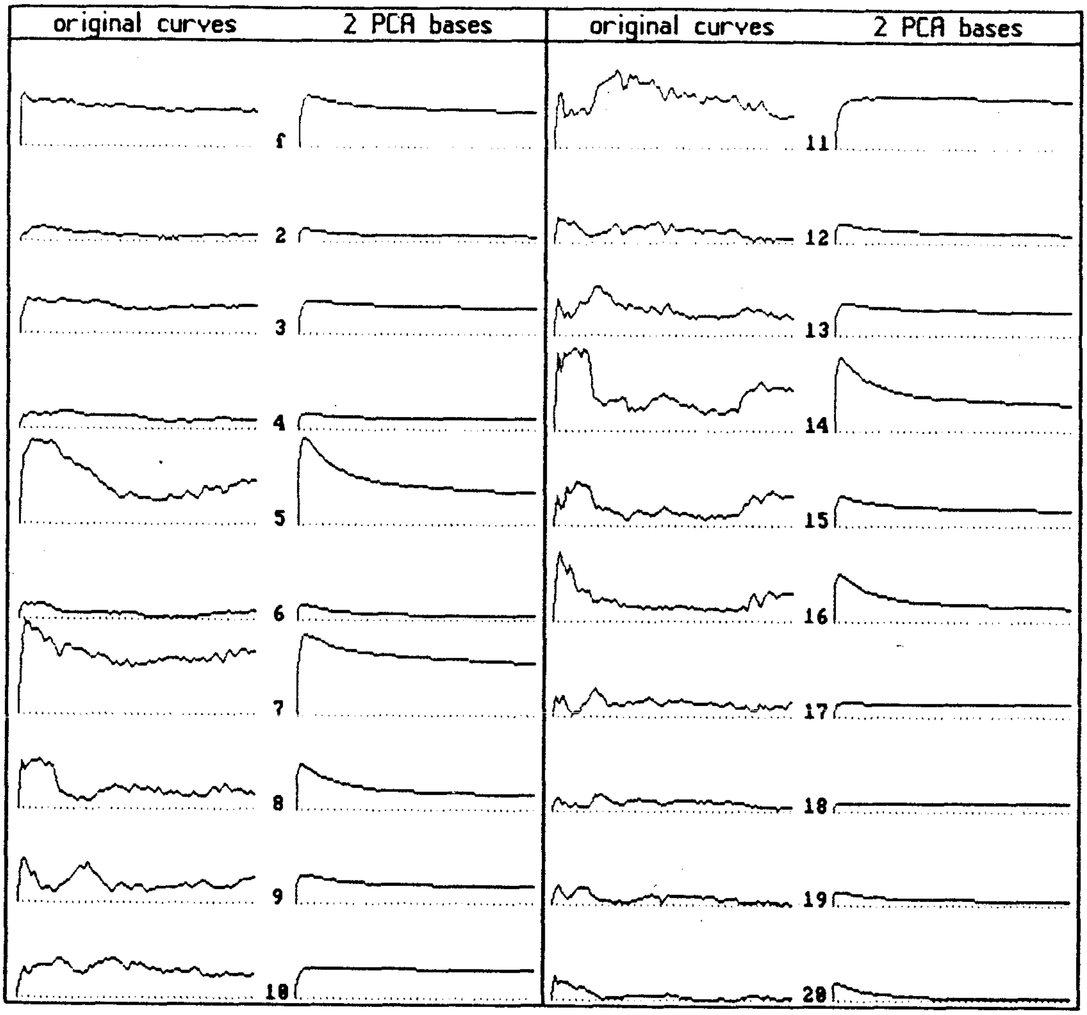


Figure 5.15: **Envelope reconstruction of a clarinet A note with 2 PCA bases.** The original envelopes for the first 20 harmonics of an A (220 Hz) clarinet sound are on the left half of the above graphs. The envelopes reconstructed with 2 PCA basis vectors and weights for the same sound are on the right. The bases used are from the piano-sax-trombone-clarinet analysis (figure 5.8 on page 101). The time span is 2.7 seconds.

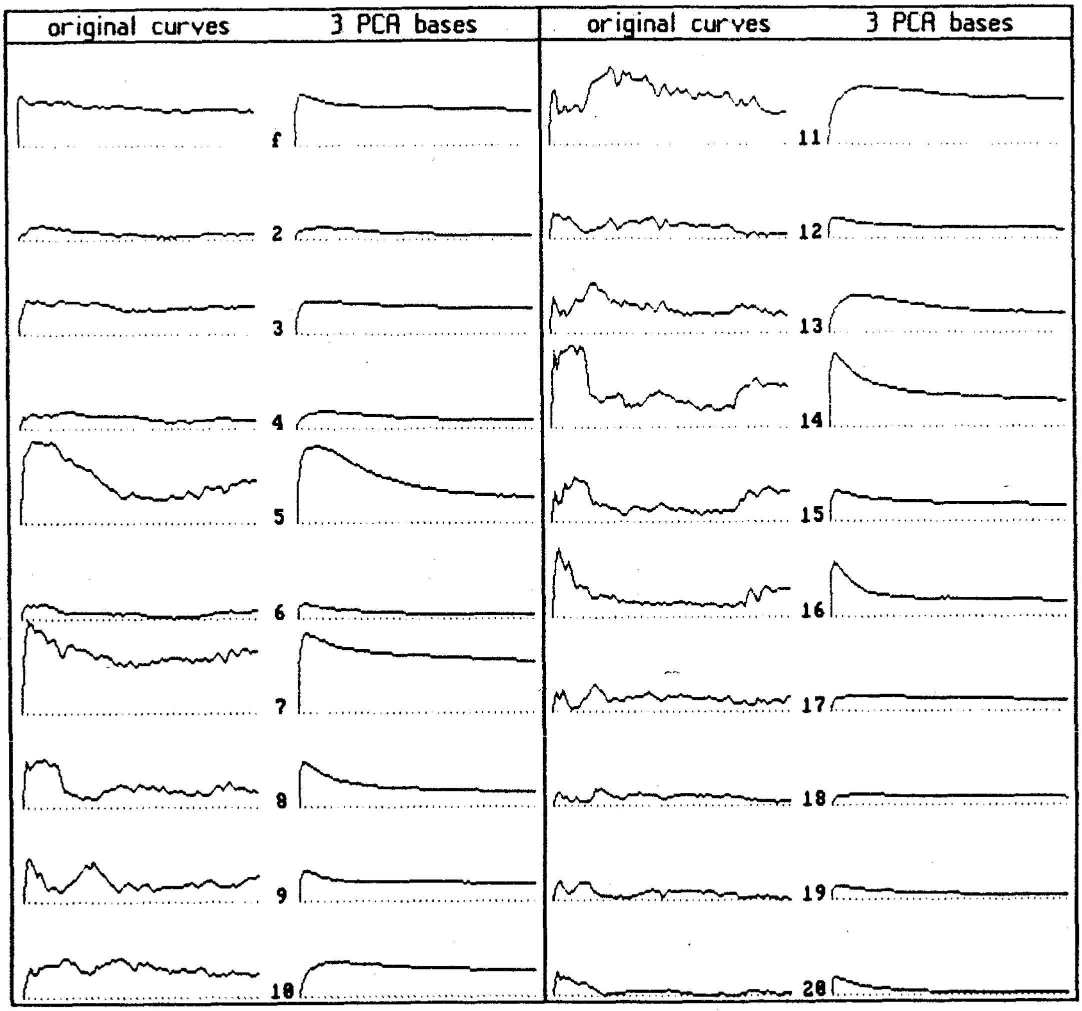


Figure 5.16: **Envelope reconstruction of a clarinet A note with 3 PCA bases.** The original envelopes for the first 20 harmonics of an A (220 Hz) clarinet sound are on the left half of the above graphs. The envelopes reconstructed with 3 PCA basis vectors and weights for the same sound are on the right. The bases used are from the piano-sax-trombone-clarinet analysis (figure 5.8 on page 101). The time span is 2.7 seconds.

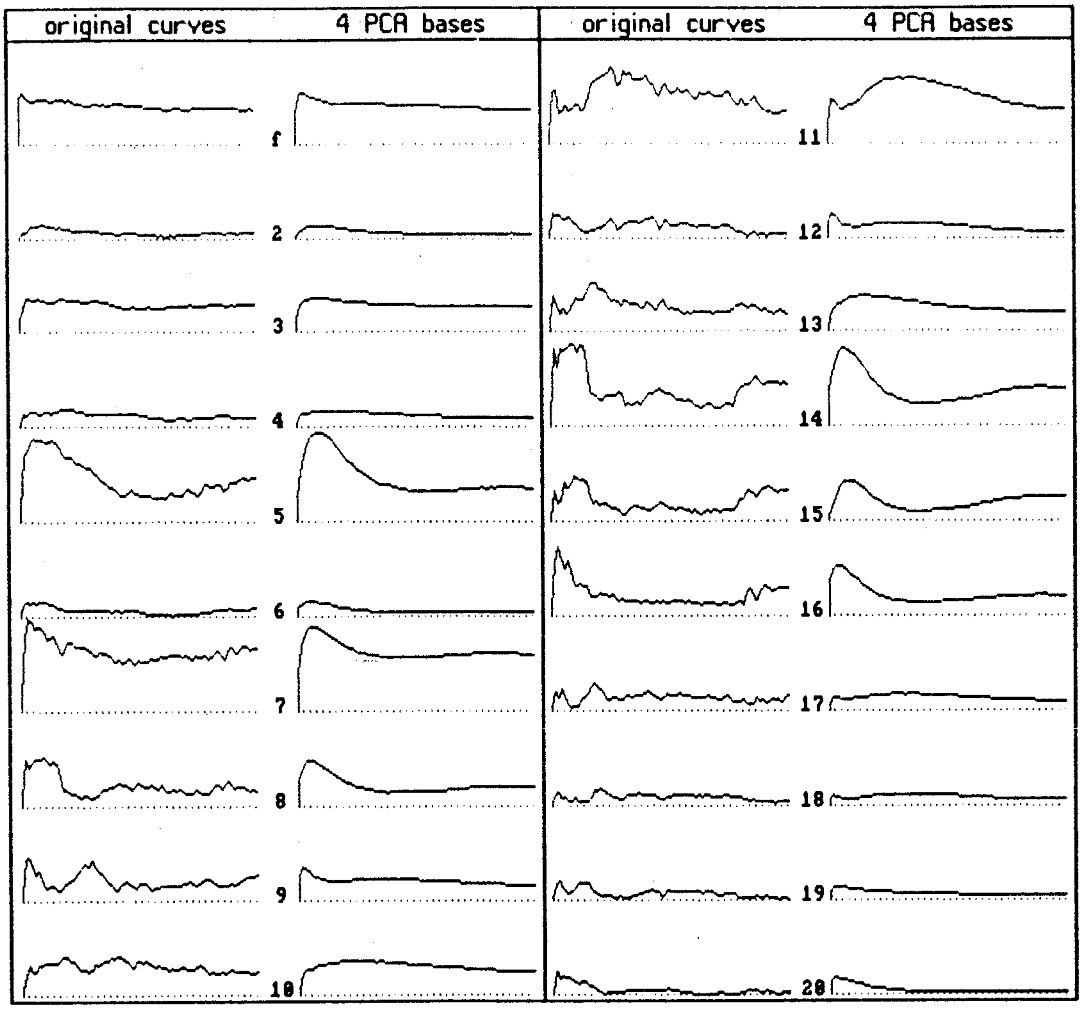


Figure 5.17: **Envelope reconstruction of a clarinet A note with 4 PCA bases.** The original envelopes for the first 20 harmonics of an A (220 Hz) clarinet sound are on the left half of the above graphs. The envelopes reconstructed with 4 PCA basis vectors and weights for the same sound are on the right. The bases used are from the piano-sax-trombone-clarinet analysis (figure 5.8 on page 101). The time span is 2.7 seconds.

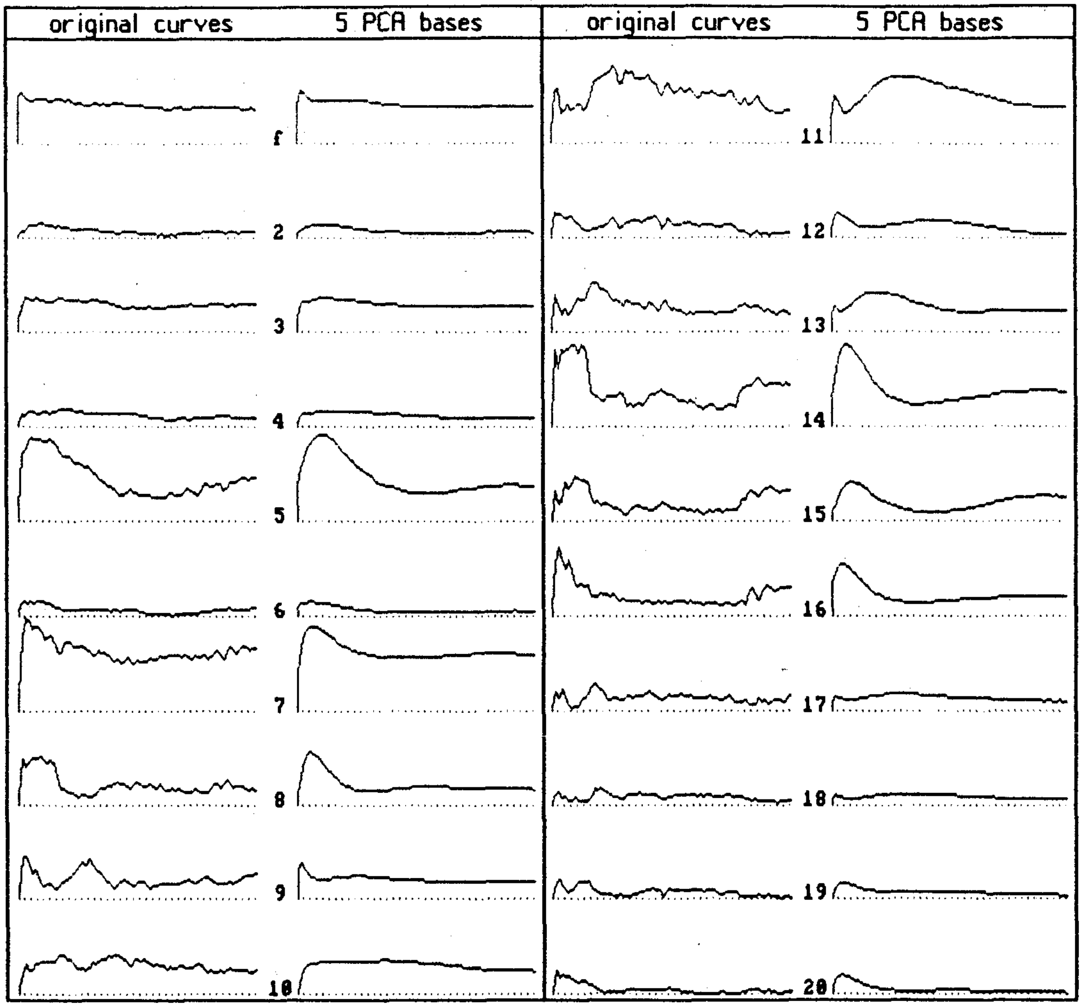


Figure 5.18: **Envelope reconstruction of a clarinet A note with 5 PCA bases.** The original envelopes for the first 20 harmonics of an A (220 Hz) clarinet sound are on the left half of the above graphs. The envelopes reconstructed with 5 PCA basis vectors and weights for the same sound are on the right. The bases used are from the piano-sax-trombone-clarinet analysis (figure 5.8 on page 101). The time span is 2.7 seconds.

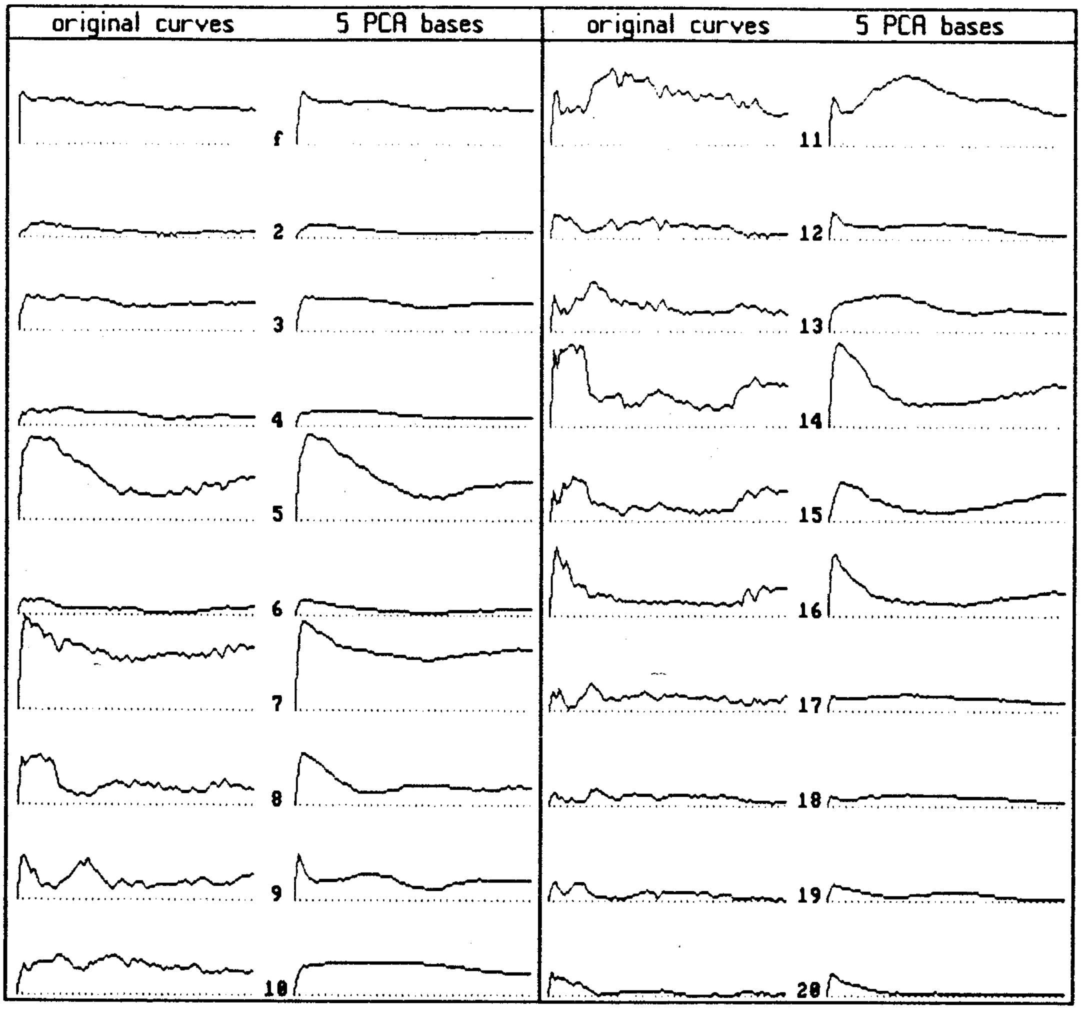


Figure 5.19: **Envelope reconstruction of a clarinet A note with 5 PCA bases from the clarinet analysis**. The original envelopes for the first 20 harmonics of an A (220 Hz) clarinet sound are on the left half of the above graphs. The envelopes reconstructed with 5 PCA basis vectors and weights for the same sound are on the right. The bases used are from the clarinet analysis (figure 5.6 on page 100). The time span is 2.7 seconds. The envelopes reconstructed here can be compared to the envelopes reconstructed with a more general bases set in figure 5.18 on page 108.

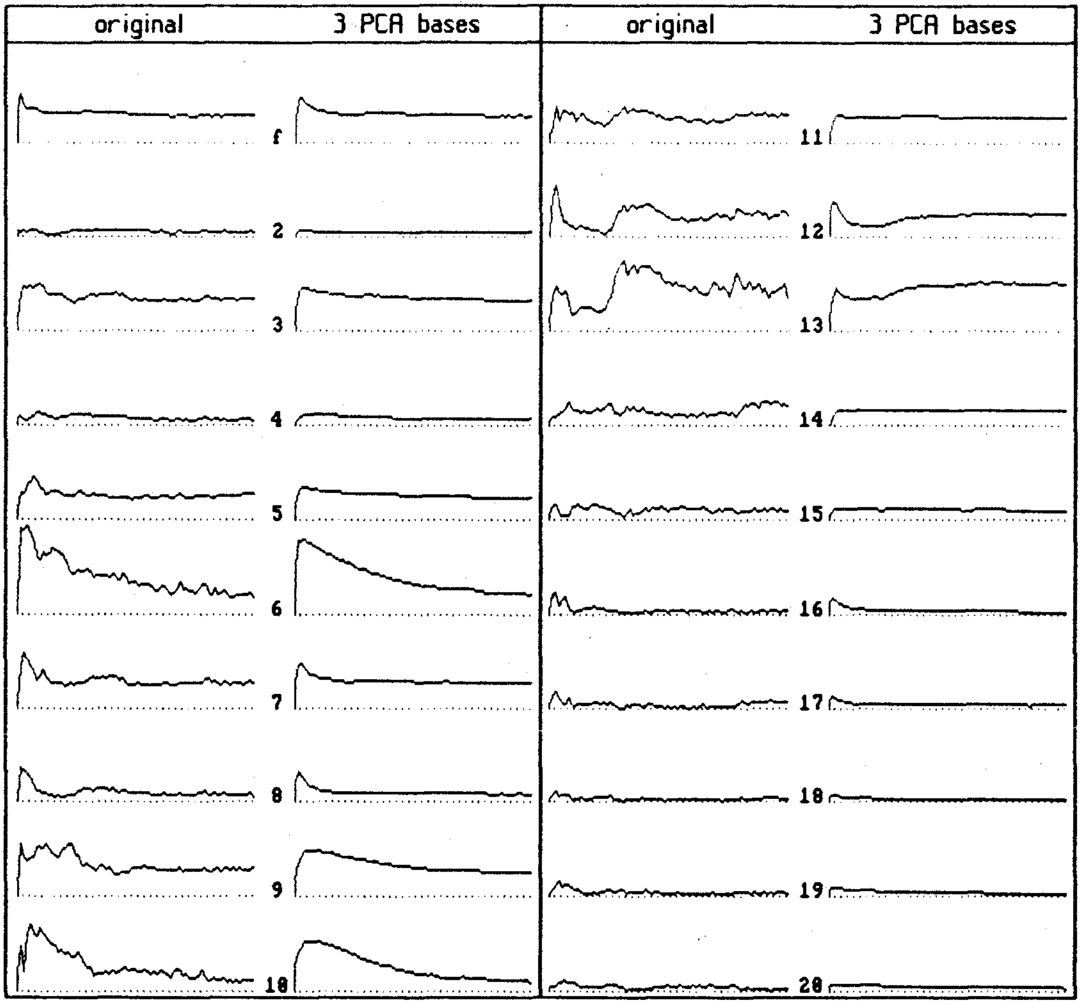


Figure 5.20: **Envelope reconstruction of a clarinet B note with 3 PCA bases from the guitar-clarinet analysis**. The original envelopes for the first 20 harmonics of a clarinet B (247 Hz) sound are on the left half of the above graphs. The envelopes reconstructed with 3 PCA basis vectors and weights for the same sound are on the right. The bases used are from the guitar-clarinet analysis (figure 5.11 on page 102). The time span is 2.7 seconds.

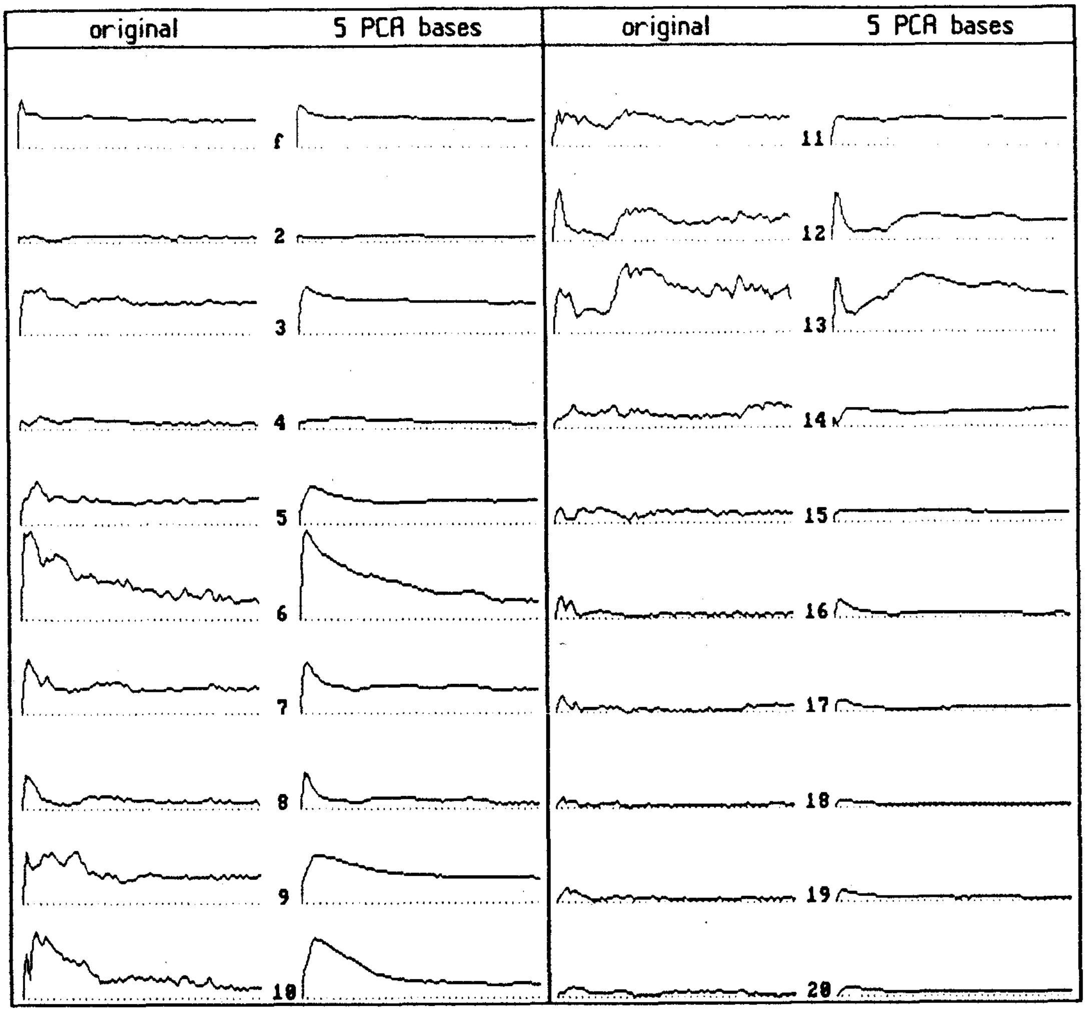


Figure 5.21: **Envelope reconstruction of a clarinet B note with 5 PCA bases from the guitar-clarinet analysis**. The original envelopes for the first 20 harmonics of a clarinet B (247 Hz) sound are on the left half of the above graphs. The envelopes reconstructed with 5 PCA basis vectors and weights for the same sound are on the right. The bases used are from the guitar-clarinet analysis (figure 5.11 on page 102). The time span is 2.7 seconds.

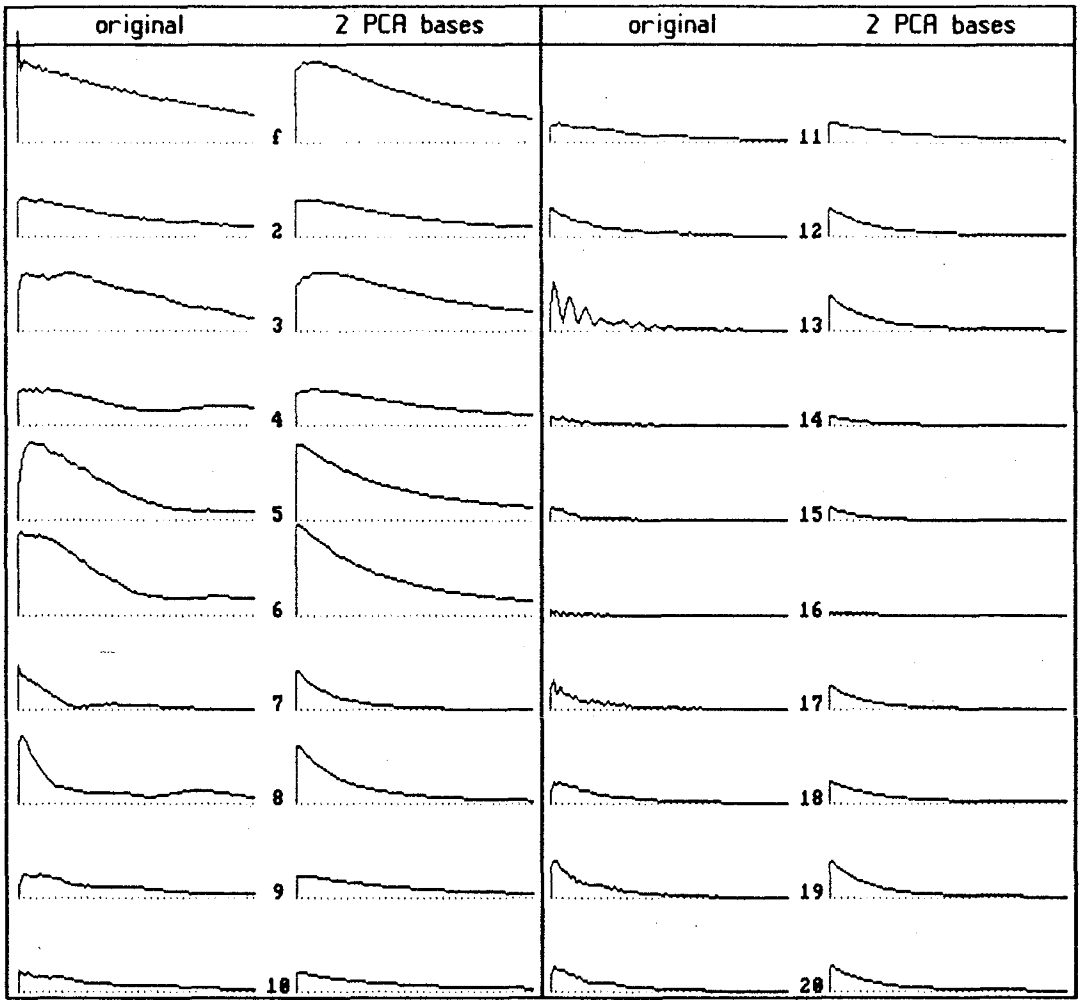


Figure 5.22: **Envelope reconstruction of a guitar D note with 2 PCA bases.** The original envelopes for the first 20 harmonics of a guitar D (147 Hz) sound are on the left half of the above graphs. The envelopes reconstructed with 2 PCA basis vectors and weights for the same sound are on the right. The bases used are from the guitar analysis (figure 5.10 on page 102). The time span is 2.7 seconds.

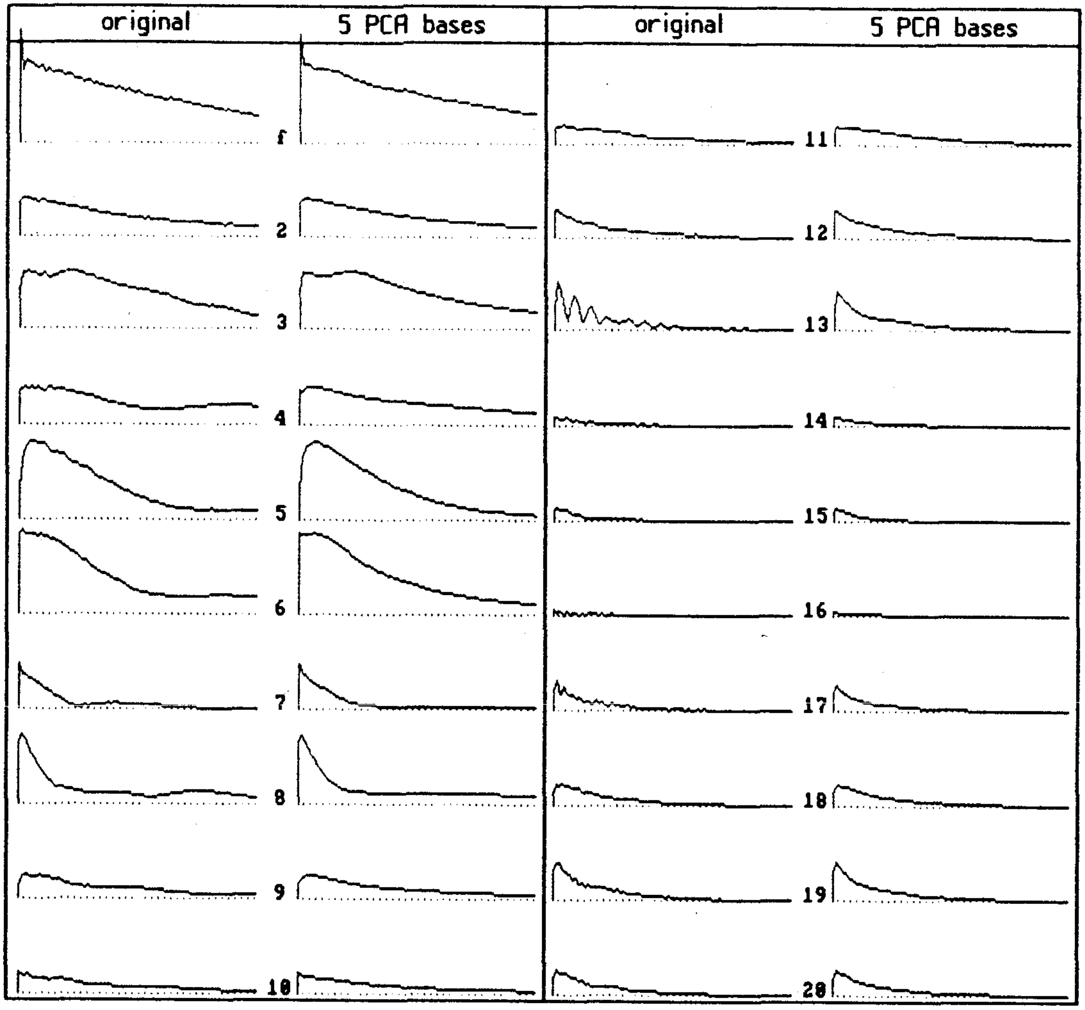


Figure 5.23: **Envelope reconstruction of a guitar D note with 5 PCA bases.** The original envelopes for the first 20 harmonics of a guitar D (147 Hz) sound are on the left half of the above graphs. The envelopes reconstructed with 5 PCA basis vectors and weights for the same sound are on the right. The bases used are from the guitar analysis (figure 5.10 on page 102). The time span is 2.7 seconds.

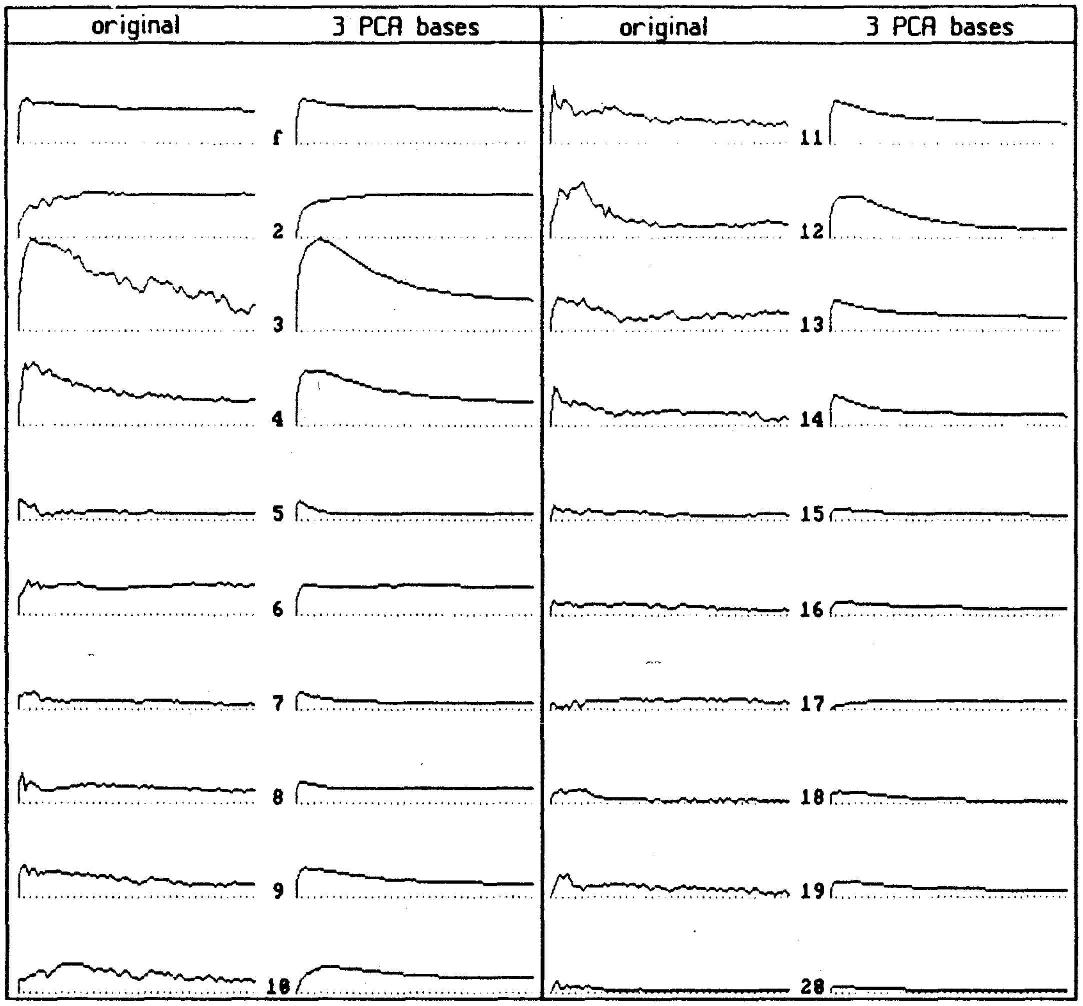


Figure 5.24: **Envelope reconstruction of a saxophone G note with 3 PCA bases.** The original envelopes for the first 20 harmonics of a tenor saxophone G (200 Hz) sound are on the left half of the above graphs. The envelopes reconstructed with 3 PCA basis vectors and weights for the same sound are on the right. The bases used are from the sax-clarinet-trombone analysis (figure 5.9 on page 101). The time span is 2.7 seconds.

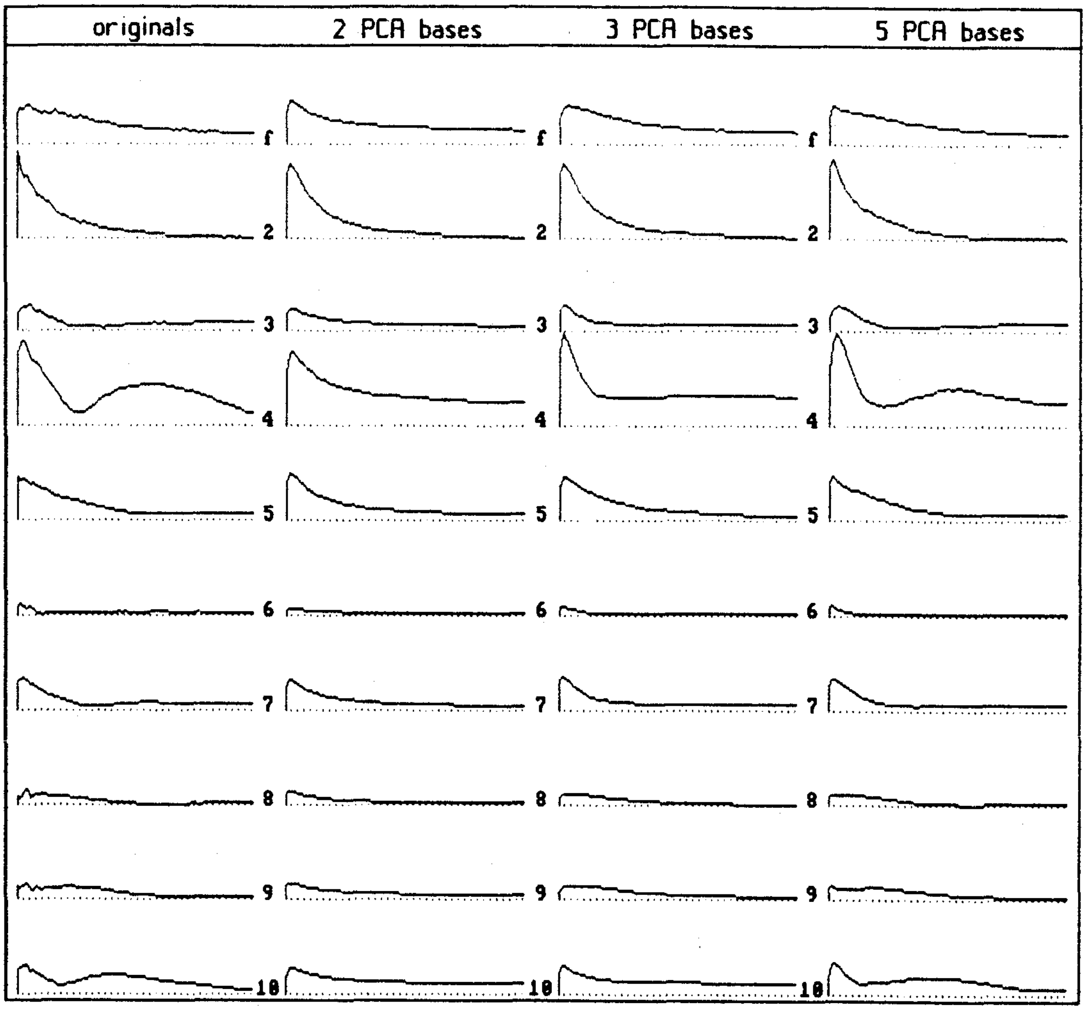


Figure 5.25: **Envelope reconstruction of a piano A note with 2, 3, and 5 PCA bases**. The original envelopes for the first 10 harmonics of a piano A (110 Hz) sound are on the left of the above graph. The envelopes reconstructed with 2, 3, and 5 PCA basis vectors and weights for the same sound are on the right. The bases used are from the piano-sax-clarinet-trombone analysis (figure 5.8 on page 101). The time span is 2.7 seconds.

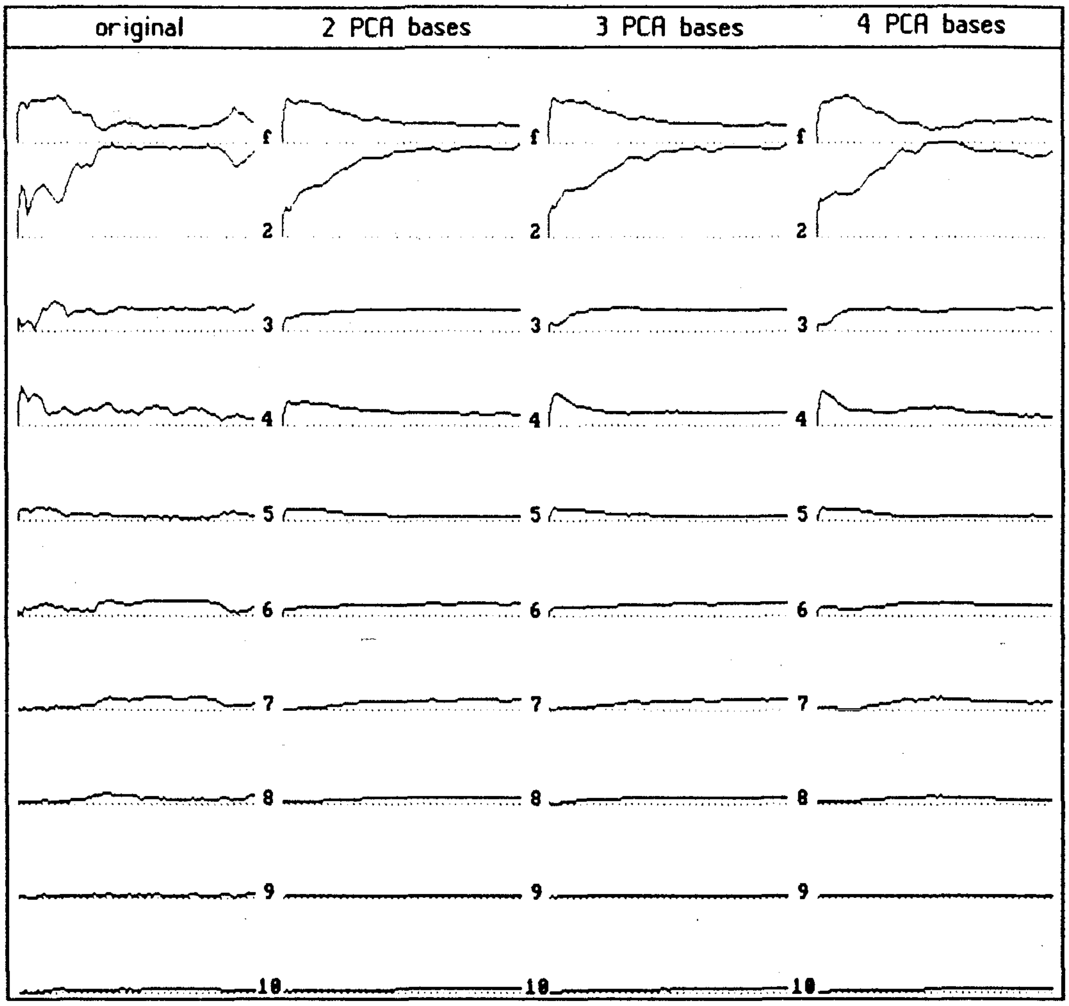


Figure 5.26: **Envelope reconstruction of a trombone D♯ note with 2, 3, and 4 PCA bases**. The original envelopes for the first 10 harmonics of a trombone D♯ (311 Hz) sound are on the left of the above graph. The envelopes reconstructed with 2, 3, and 4 PCA basis vectors and weights for the same sound are on the right. The bases used are from the trombone analysis (figure 5.7 on page 100). The time span is 2.7 seconds.

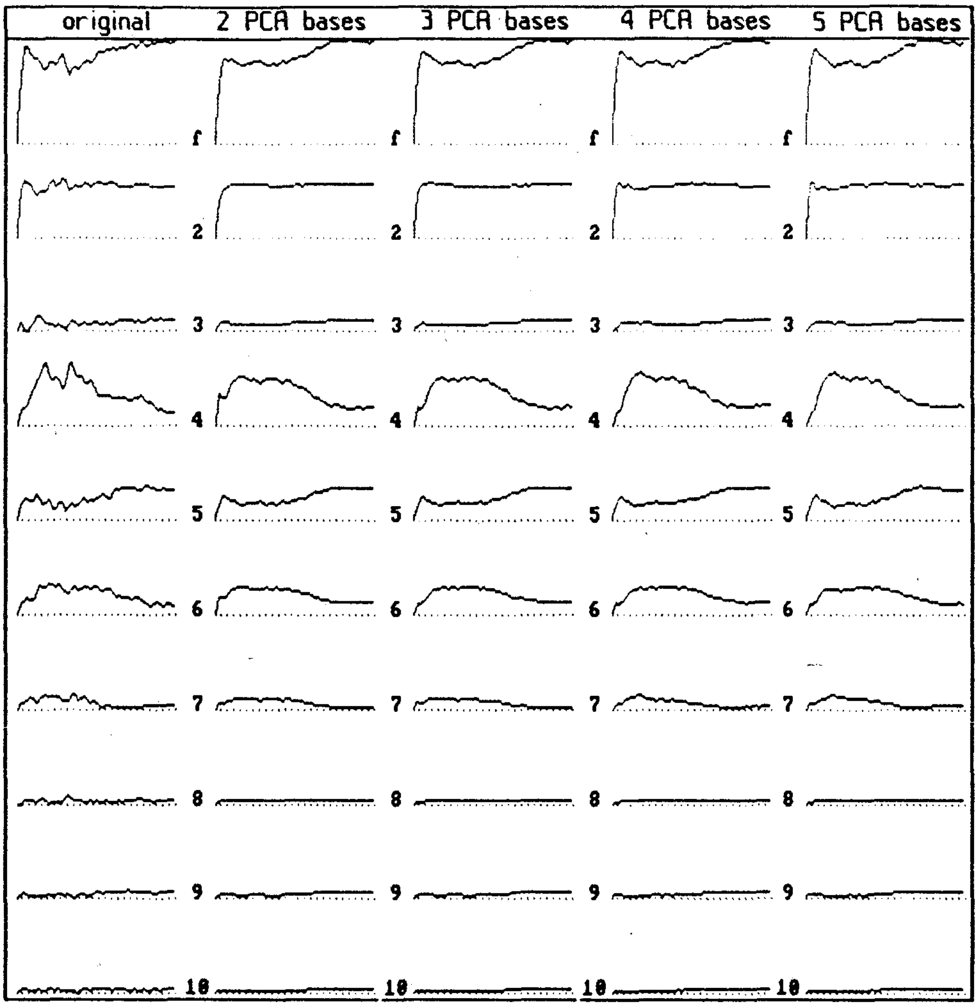


Figure 5.27: **Envelope reconstruction of a flute G note with 2, 3, 4, and 5 PCA bases.** The original envelopes for the first 10 harmonics of a flute G (392 Hz) sound are on the left of the above graph. The envelopes reconstructed with 2, 3, 4, and 5 PCA basis vectors and weights for the same sound are on the right. The bases used are from the flute analysis (figure 5.12 on page 103). The time span is 1.8 seconds.

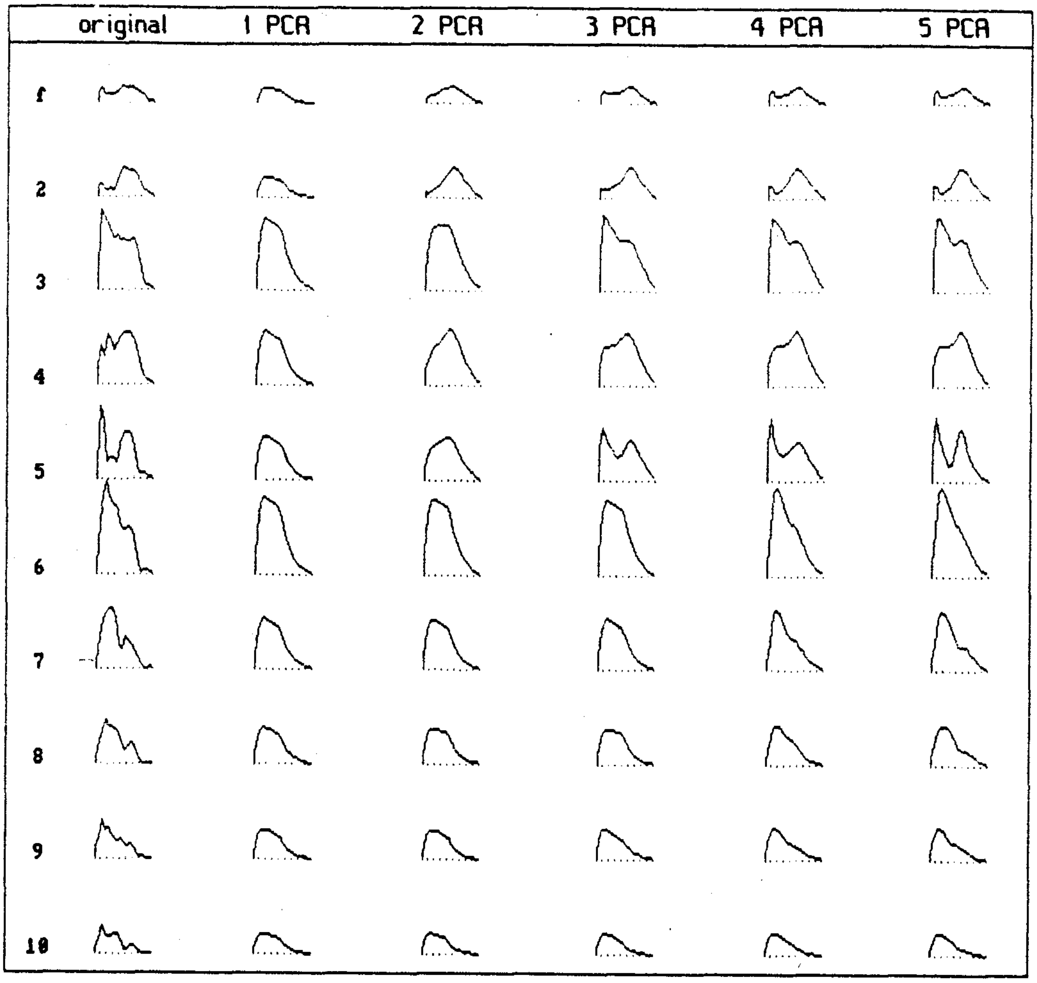


Figure 5.28: **Envelope reconstruction of a (short) trombone B note with 1, 2, 3, 4, and 5 PCA bases.** The original envelopes for the first 10 harmonics of a short trombone B (123 Hz) sound are on the left of the above graph. The envelopes reconstructed with 1, 2, 3, 4, and 5 PCA basis vectors and weights for the same sound are on the right. The bases used are from the trombone-clarinet-flute (short sound) analysis (figure 5.13 on page 103). The time span is .65 seconds and the sound decays naturally.

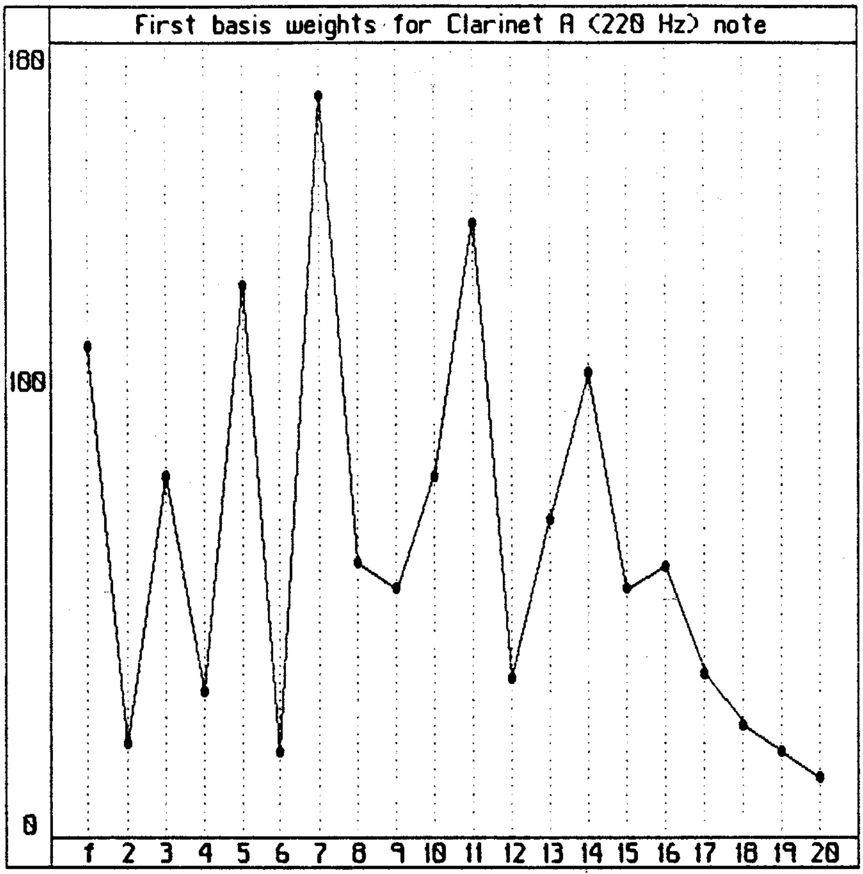


Figure 5.29: **First basis weights for a clarinet A (220 Hz) note.** The envelope curves resulting from these first basis weights are illustrated in figure 5.14 on page 104. The bases used are from the piano-sax-clarinet-trombone analysis (figure 5.8 on page 101). Note that the first basis vector is weighted heavily for the odd harmonics—a characteristic of clarinet sounds.

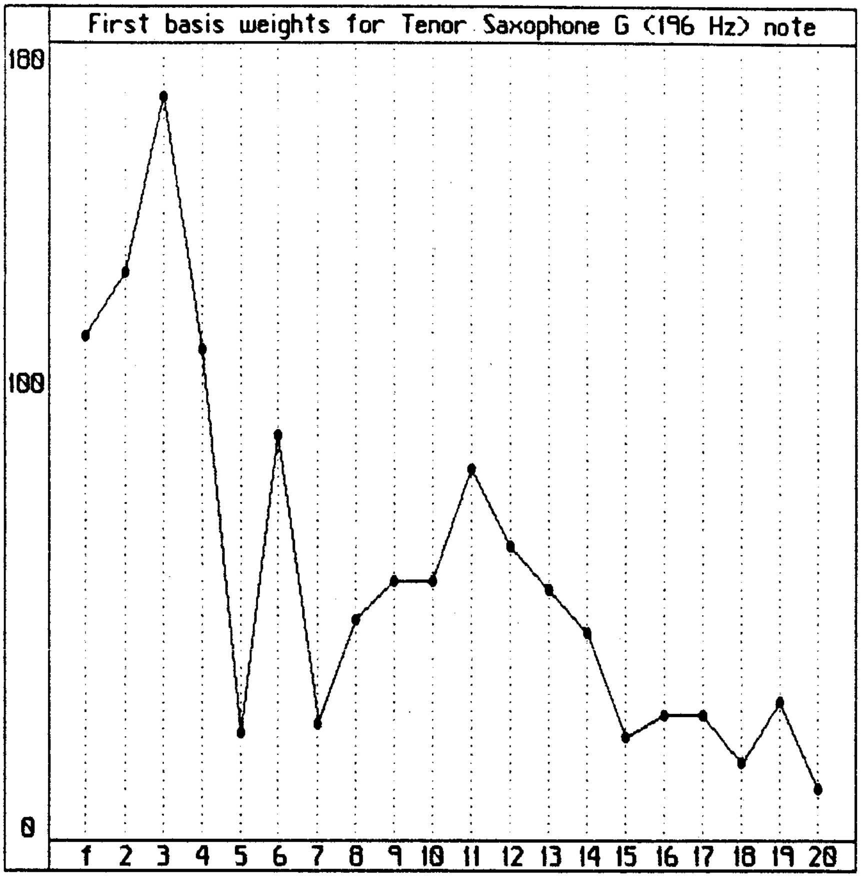


Figure 5.30: **First basis weights for a tenor saxophone (196 Hz) G note.** The bases used are from the piano-sax-clarinet-trombone analysis (figure 5.8 on page 101).

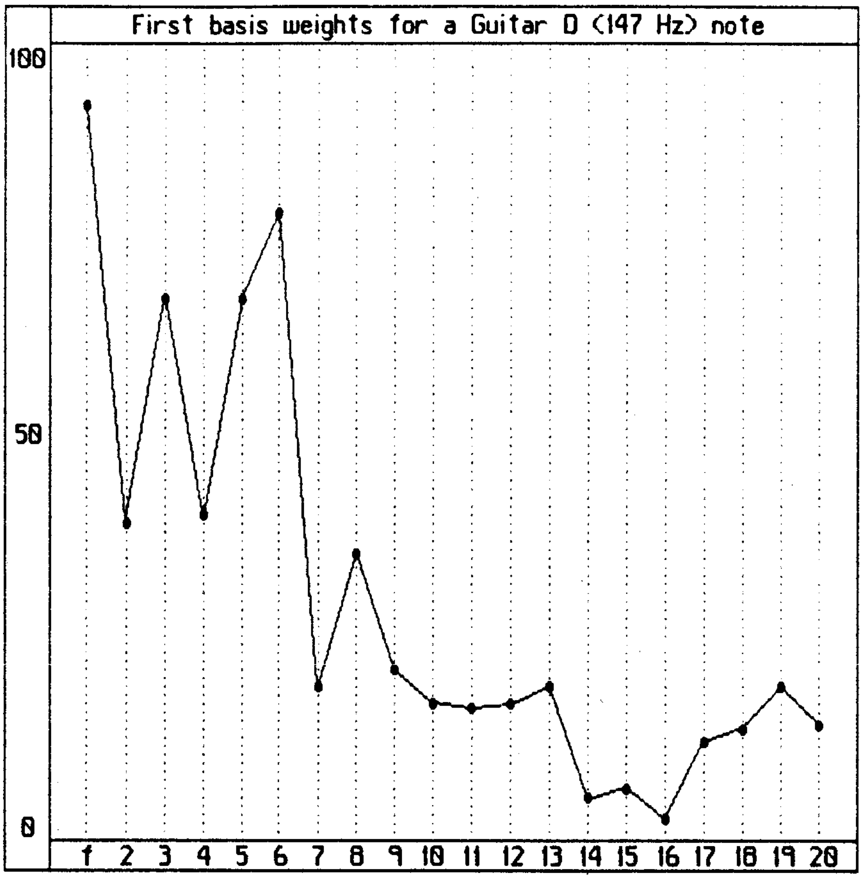


Figure 5.31: **First basis weights for a guitar D (147 Hz) note.** The envelope curves resulting from these first basis weights (as well as the second basis weights) are illustrated in figure 5.22 on page 112. The bases used are from the guitar analysis (figure 5.10 on page 102).

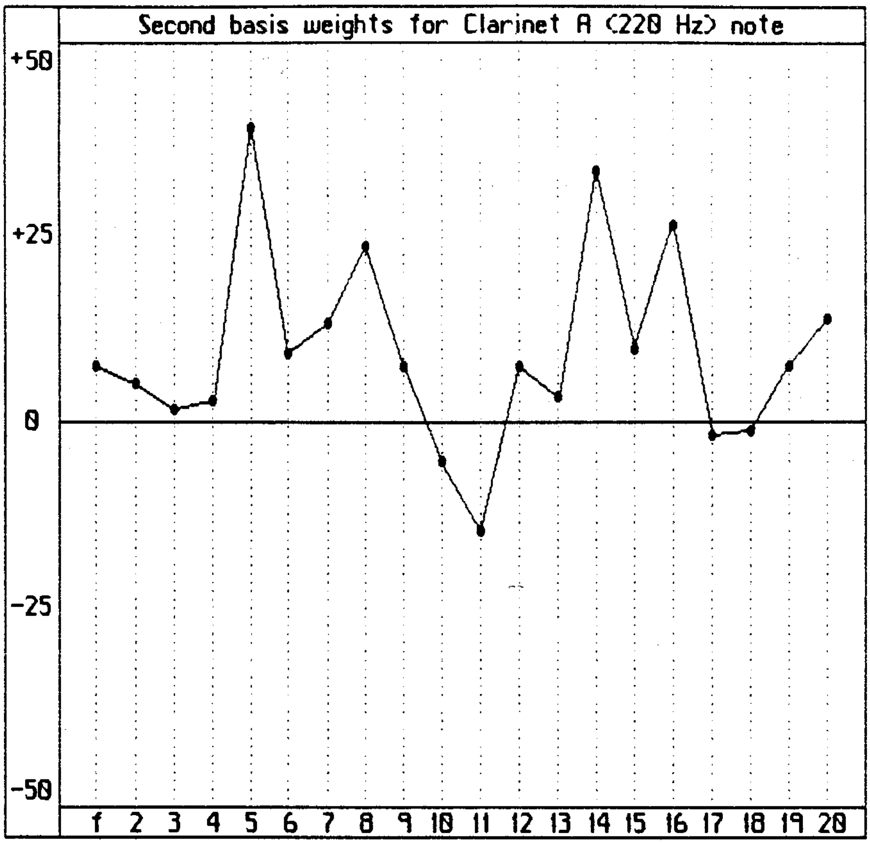


Figure 5.32: **Second basis weights for a clarinet A (220 Hz) note.** The envelope curves resulting from these second basis weights (as well as the first basis weights) are illustrated in figure 5.15 on page 105. The bases used are from the piano-sax-clarinet-trombone analysis (figure 5.8 on page 101). Note the low "attack" values for harmonics 10 and 11. This is reflected in the envelope curves in figure 5.15.

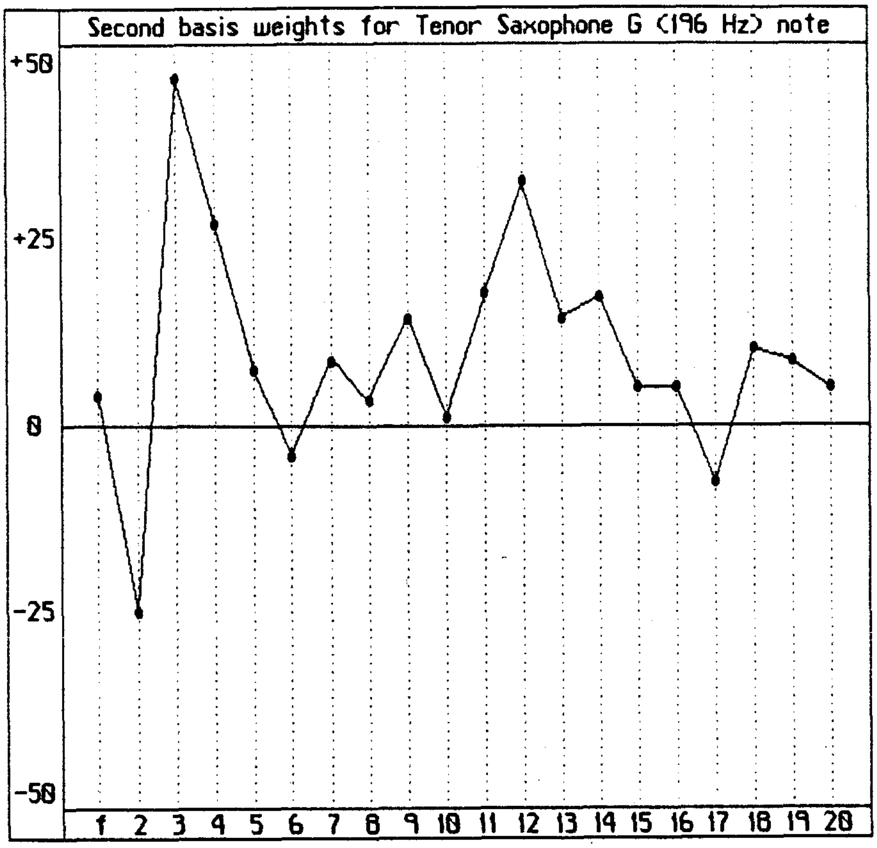


Figure 5.33: **Second basis weights for a tenor saxophone (196 Hz) G note.** The bases used are from the piano-sax-clarinet-trombone analysis (figure 5.8 on page 101).

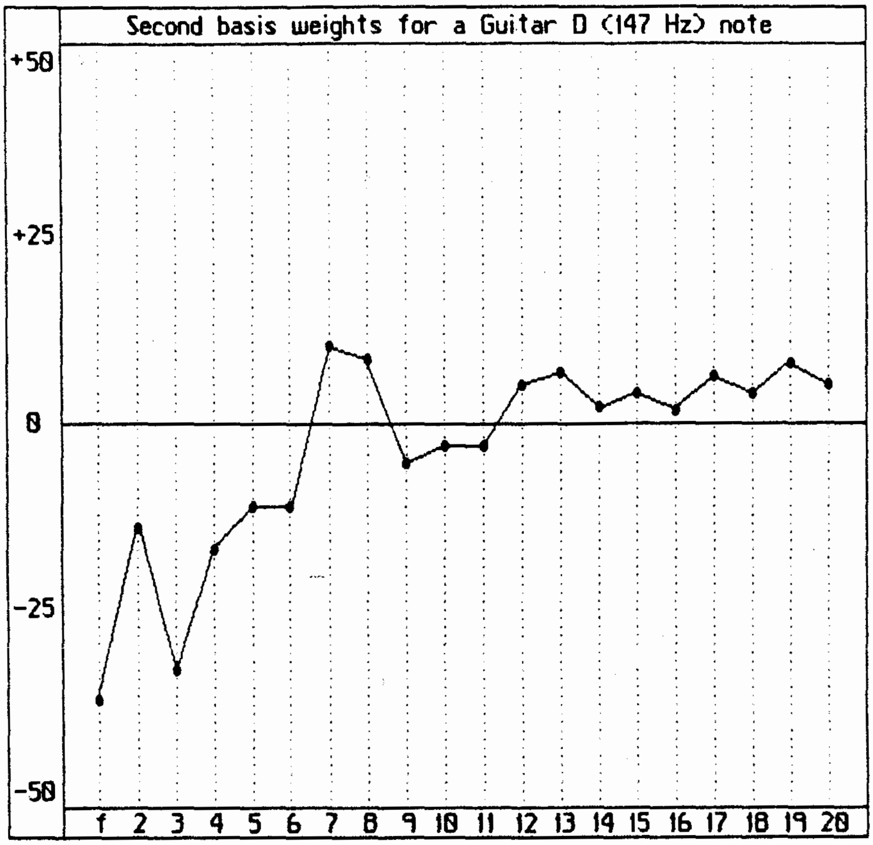


Figure 5.34: **Second basis weights for a guitar D (147 Hz) note.** The envelope curves resulting from these second basis weights (as well as the first basis weights) are illustrated in figure 5.22 on page 112. The bases used are from the guitar analysis (figure 5.10 on page 102).

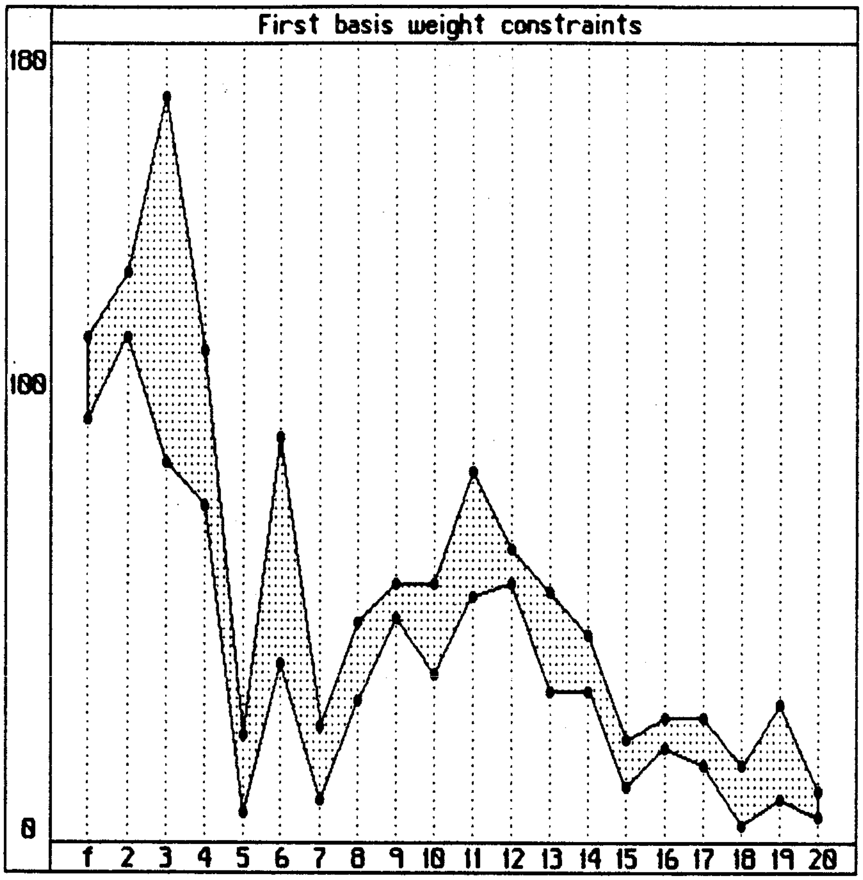


Figure 5.35: **Constraints on the weights for a range of sounds.** This graph shows the weight constraints (first basis) for a hypothetical example. The user would select a weight path within the bounds illustrated, in order to construct sounds with closely related timbral qualities.

Chapter 6

Conclusions

The following section summarizes the previous chapter, *Analysis Results and Interpretation*.

6.1 Summary of Results

A principal component analysis (PCA) of the harmonic amplitude envelopes of 280 musical instrument sounds was quite successful at reducing the dimensionality of the envelope curves. Two principal components accounted for approximately 97% of the variance of the envelope curves and five principal components accounted for approximately 99% of the variance. Table 5.1 on page 75 summarizes the variance accounted for in the various instrument groups analyzed.

A PCA bases representation of harmonic amplitude envelopes is proposed as an alternate to the approximation of envelopes with line segments (see page 18).

6.1.1 Advantages of Bases Representation

A PCA bases representation has the following advantages:

- Smoothly varying envelopes curves can be reconstructed quickly (typically several hundred multiplications and additions, see equation 4.6 on page 72) from a set of bases (common to all envelopes) and a small number of scalar weights for each envelope curve.

- The extraction of bases and weights is automated. Heuristic methods of line segment approximation or complex extraction algorithms as proposed by Strawn [69] (see page 19) are not required.

- Envelope curves can be significantly reduced in dimension and represented by as few as two parameters, without the loss of perceptually important features of the envelopes. Reduction to one dimension retains some of the qualities of instrument timbre, useful for instrument identification.

- Gradated levels of approximation to the original envelopes are available by selecting the number of basis vectors to use in the envelope reconstruction. Specific local features ("blips", non-monotonic variation, etc.) can be captured by increasing the number of bases in the reconstruction.

- Envelope "shapes" are represented in a standardized form, over all envelopes (common bases). This facilitates the comparison of harmonic amplitude envelopes and provides an alternate strategy for grouping harmonics (see Charbonneau [10] discussed on page 23, and Kleczkowski [34] discussed on page 24).

- The common bases representation allows grouping of envelopes—over all the harmonics of a sound—to produce higher level control mechanisms.

- The bases representation allows both high and low level control mechanisms to use the same underlying representation.

- The first few bases have a simple perceptual interpretation that may be useful in developing intuitive control strategies.

- The representation can incorporate new envelopes (and sounds) not included in the original analysis.

- The mathematical technique used to extract the bases automatically smooths out random fluctuation in envelopes.
- The orthogonal bases are *data dependent* and hence maximize data reduction of the envelope curves.
- The PCA data collection and analysis is completely separate from a resynthesis implementation using the bases representation. The basis vectors and weights are in the form of ASCII floating point data and can be implemented on any system, using any sample resolution or sampling rate[1].

6.1.2 Interpretation of Bases

A tentative interpretation of the basis vectors is the following:

- The first basis vector captures the overall amplitude of harmonics averaged over the duration of the sound. The first basis weights, considered as a group (over all harmonics of a sound), yield a measure of the *spectral energy distribution* of a sound.
- The second basis vector captures temporal qualities of the attack/decay characteristics of harmonics.
- Third and higher order bases refine the attack/decay characteristics of harmonics as well as include local envelope features throughout the duration of the sound.

An examination of the first basis vectors over the different instrument groups included in the analyses (see Table 5.1 on page 75) indicates that the instrument sounds can be grouped into two major classes:

[1] Resynthesizing with a higher sampling rate would simply require computing more sample values between envelope points.

- Sounds whose overall energy (wind instruments) is sustained throughout the sound[2].

- Sounds with a natural decay (plucked or struck stringed instruments—the guitar and piano in this study).

Mixing instruments from the two classes—in one PCA—works surprisingly well. The major disadvantage is that decay information (for the instruments with natural decay) moves into higher order bases. Hence, 1-basis approximations for these sounds lose the character of the instrument. For two or more bases approximations it appears reasonable to include *any* harmonic instrument in the analysis group. The range of instruments included in a PCA analysis can be limited for optimal accuracy for a particular instrument (or instruments) or a wide assortment can be included for generality. Once the envelope curves are available, any particular grouping of instrument sounds can be selected for PCA analysis. The choice of bases to use for resynthesis can then be left up to the user for different applications. In some cases it may be desirable to separate instruments into the two classes mentioned above. In other cases, grouping of instruments over the two classes may be more appropriate.

The early decay of harmonics (or the sound as a whole) does not pose a problem when 2 or more bases are used in the reconstruction of envelopes. In these cases the second basis "subtracts" amplitude components from the first basis contribution at later positions in the sound.

6.1.3 Aural Evaluation

The major aural discrepancy was between the original digitized versions of instrument sounds and *any* of the resynthesized versions. Detailed attack information was noticeably absent from the resynthesized sounds. Resynthesized sounds also lacked the "liveliness" of the digitized sounds. This is most likely due to the constant frequency

[2] Including the decay portion of wind instrument sounds in the PCA analyses would not likely alter the two-class interpretation since wind instrument sounds die out quickly when wind energy is removed from the instrument.

harmonics, the 20 harmonic analysis limit, and the lack of inharmonic components in the resynthesized versions. The resynthesized piano sounds were particularly poor.

Sounds reconstructed with 1-basis approximations to the envelopes were clearly inferior to the sounds resynthesized with all the original envelope information. However, instruments were still recognizable with 1-basis approximations (with a few exceptions, see page 86).

On the other hand, there was little perceptual difference (with some exceptions) between sounds resynthesized using all the envelope information extracted from the Fourier analyses, and sounds resynthesized with 2 or more PCA bases approximations. The principal discrepancy in timbre quality was with the wind instrument sounds. These sounds resynthesized with PCA bases approximations lacked some of the "uneveness" present in sounds resynthesized with all the envelope information. An inspection of the envelopes of wind instrument sounds reveals considerable microstructure. A tentative hypothesis is that envelope microstructure conveys the influence of player induced idiosyncrasies on the sound. It may be possible to replicate this effect algorithmically rather than extracting microstructure variations from a sound specific analysis.

6.1.4 Guitar Attack

Inharmonic components of fixed frequency were found to be present at the onset of guitar sounds. Analysis revealed the major inharmonic component to be the air resonance of the instrument body. It is suggested that the other fixed inharmonic components result from resonances of the soundboard, although it was not possible to verify this. All the inharmonic components decayed at a roughly exponential rate from the onset of the sound.

It is hypothesized that a significant portion of the characteristic sound produced at the onset of guitar sounds is a result of a bell-like decay of these instrument resonances.

An algorithm was developed to replicate this effect, using no sound specific information (other than the empirically determined resonances of the instrument). A subjective assessment of the sounds produced, with and without the resonance components, tends to support the hypothesis.

6.2 Suggested Analysis Improvements

The two major deficiencies of the sound analysis performed for this thesis are the 8 bit sample resolution and the lack of frequency fluctuation information.

While frequency fluctuations were not the focus of data reduction, their availability for resynthesis would allow a better aural assessment of the success of the bases approximation of harmonic envelope curves.

Additive resynthesis (see equation 2.1 on page 16) can be modularized into three separate (independent[3]) components: *1)* harmonic amplitude changes over time, *2)* frequency fluctuations, and *3)* attack reconstruction[4] (optional). To fully assess the effectiveness of resynthesis methods applied to *one* of the components (harmonic amplitude envelopes in this case) the other components (attack characteristics and frequency fluctuations) should replicate as closely as possible the effects they are trying to emulate. Any discrepancies between resynthesized sounds and the original digitized versions can then be attributed to the resynthesis module being manipulated.

For this study, the perceptual similarity between sounds resynthesized with PCA bases envelopes and sounds resynthesized with all the envelope information, supports the efficacy of the representation. Even stronger support would result if the PCA bases approximated sounds and the original digitized versions were perceptually indistinguishable. This was not the case. Extracting the original frequency fluctuations from the sounds and using them in the resynthesis might reduce the discrepancy between the digitized and resynthesized versions and allow the effects of amplitude envelope manipulation to be assessed independently of the other two components.

[3] While frequency fluctuations and spectral changes over time are not necessarily independent physical processes, it may be advantageous to manipulate them independently in resynthesis—particularly if frequency fluctuations can be generated algorithmically.

[4] This module would add instrument specific attack information (possibly inharmonic) *not* captured by the other two modules.

While the 8 bit sample resolution is less of an issue in analysis as opposed to resynthesis, analyzing sounds with a 16 bit sample resolution would allow *dynamics* (intensity variations) to be included in the sounds analyzed.

The following improvements are suggested for any future analysis of principal components of amplitude envelopes:

- Extract frequency fluctuations from the samples and include them in all resynthesized versions of the original digitized sounds.
- Digitize sounds at 16 bit resolution and at a higher sampling rate (e.g. 44.1 kHz).
- Include more harmonics in the analysis (e.g. 40–100). A higher sampling rate would allow this.
- Include intensity variations in the instrument sounds—and any other influences that are of interest.
- Include more instruments and instrument families in the sounds analyzed.
- Use a standardized recording environment or an anechoic chamber.
- Analyze longer sounds and include the natural decay of all instruments (this will involve carefully timing wind instrument sound duration during recording).
- Concentrate envelope points in the critical onset of the sound. Envelope (analysis) points need not be equally spaced and wider spacing would likely be acceptable later in the sound.

6.3 Potential Applications and Future Research

There are two principal applications of the PCA bases representation of harmonic amplitude envelopes. The first is as a research tool for studying the psychoacoustics of timbre, the second is as a practical method of envelope manipulation in musical sound synthesis.

6.3.1 Psychacoustic Research on Timbre

Relevance of Local Envelope Features

One of the issues in timbre research is the perceptual relevance of local features in envelope curves [10, 34, 64, 69]. Given that data reduction of the signal information in harmonic sound aids in establishing psychophysical relationships, the question is: *To what degree can the envelope information be simplified without losing perceptually important information?*

A PCA representation of the envelopes with *all* the bases available may help to answer this question. Psychological testing of subject reactions to sounds resynthesized with increasing numbers of bases could establish the point at which envelope curve information is redundant.

Such a study might also address the question raised in this thesis as to the perceptual *role* played by the microstructure of envelope variation.

Timbre Space

Given the tentative interpretations assigned to the first few bases, and the correspondence of the interpretations to the physical parameters of Wessel's timbre space [77] (see page 39), it would be useful to assess the similarity judgements (using multidimensional scaling, see page 32) of sounds reconstructed with the first two bases.

The ordering of the sounds in a 2-dimensional timbre space could be correlated with various (scalar) measures extracted from the *set of weights* (over all harmonics of a sound) of the first two bases[5]. A high correlation would indicate that useful

[5] It may be necessary to include higher order bases in the derivation of a scalar "attack bite" measure.

parameters have been found for resynthesizing sounds with simple perceptual parameters. These measures would be suitable for the high-level manipulation of parameters in timbre space. In addition the full power of PCA bases approximated resynthesis would be available (at a lower level) to fine-tune the sounds within the constraints imposed by the simple scalar measures.

Assessing the Effects of Various Parameters on Timbre

Since the envelopes all have the same underlying representation (weighted bases), the effect of parameters such as note register, note intensity, and various instrument specific parameters (picking position on a guitar string, etc.) can be assessed both graphically and statistically. This may be useful in exploring the phenomenon of *timbral constancy* of instruments across widely varying harmonic envelope shapes.

Automated Instrument Identification

The standardized envelope representation may be useful in developing pattern matching criteria for instrument identification (vowel recognition using principal component analysis is discussed by Plomp, Pols, and van de Geer [52], see page 30). This would be most useful in the machine recognition of polyphonic music (see Moorer [45]).

Spectral Changes over Time

Principal component analysis could also be applied to the harmonic *spectrum* curves at different time points in the sounds. The evolution of the sound could then be plotted in terms of the bases weights as they change over time. This would provide an alternate (perhaps complementary) perspective on the timbre of sound.

Cognitive Strategies

Principal component data reduction has been useful in computational color vision for developing image analysis and resynthesis algorithms. While these methods do not imply that the same strategies are employed in human perception, they do provide existence proofs for data reduction on the same scale as is hypothesized to exist in human perception. Principal component analysis might play a similar role in computational audition, particularly with respect to the problem of timbral constancy.

6.3.2 Musical Synthesis Applications

A modular approach to synthesis of acoustic musical instrument sounds has some advantages, the primary one being that each module can be manipulated (and studied) independently.

Three modules are proposed:

- *1.* Harmonic amplitude envelope generation.
- *2.* Harmonic frequency fluctation generation.
- *3.* Generating attack characteristics *not* accounted for by the first two modules.

Frequency Fluctuation Generation

The work of Charbonneau [10] and Kleczkowski [34] (see pages 28 and 24) indicates that considerable data reduction of harmonic frequency fluctuations is possible. The *form* that this data reduction takes is different from the data reduction of harmonic envelopes. Frequency fluctuations also appear to be suitable for manipulation *independently* of harmonic amplitude envelopes. Further data reduction of frequency fluctuations may be possible by generating the fluctuations algorithmically (as originally proposed by Risset and Mathews [56]). Different instruments may require different treatment.

Generating Attack Characteristic

Attack characteristics may also require an instrument specific treatment, given the ear's high discriminability for onset transients. The guitar attack algorithm discussed on page 95 is an example. Further work on this algorithm might concentrate on verifying the soundboard resonance components and expanding the number of soundboard resonances—either through analysis or choosing arbitrary resonance frequencies (soundboard resonances will in general vary from instrument to instrument). Attack algorithms might also be developed for other instruments.

Harmonic Amplitude Envelope Generation

A PCA bases representation of harmonic amplitude envelopes—as developed in this thesis—is proposed as an alternative to line segment approximations. Some of the advantages are listed on page 126 and some suggested refinements of the analysis method are outlined on page 132. Futher empirical work might focus on methods of adding microstructure variations to bases reconstructed envelopes. This may require instrument (or instrument family) specific algorithms.

Additional work is also required on ways to alter the *duration* of the resynthesized sounds, since the PCA bases extend over a fixed time frame. There are several strategies that could be employed. The first is to analyze sounds of different duration[6] and use a different set of bases for each time frame.

It may also be possible to simply reduce the volume level of the sound (as a whole) to emulate the natural decay of instrument sound. This may be appropriate for instruments whose intensity level can be maintained by the player (the wind instruments for example). A study by Saldanha and Corso indicates that the decay characteristics of sound have little perceptual significance [61] (see footnote on page 25).

Sustaining sound *beyond* the bases time frame, by splicing in "steady state" sound,

[6] Sounds produced by acoustic guitars and the piano have a natural decay, largely independent of player influence. For this class of instrument sound, the analysis time frame should be long enough to include the longest sustaining note or at least a significant portion of it.

is possible but not likely to produce realistic instrument sound. This may however be required for very long sustaining sounds. A better strategy may be to analyze sound with a long time frame and shorten the sound by reducing the intensity level to zero as discussed above.

Some suggested uses for the bases representation of envelopes in a musical context are the following:

- Cataloging a large number of instrument sounds over a wide variety of conditions (instrument, note register, intensity, etc.). The parsimonious representation and automated extraction of basis weights makes this "brute force" method feasible. A large number of subtle influences on timbre can be easily captured and stored in a database.
- The representation may be useful for the development of hierarchical control structures with intuitive control parameters (see *Higher Level Control Mechanisms* on page 90).
- The representation may be appropriate as the underlying envelope manipulation strategy in Wessel's timbre space [77] (see page 39).
- Interpolation between instrument sounds. The common bases representation facilitates the alteration of envelopes in a smooth, predictable fashion.
- Altering the timbre of sounds while staying within specified timbral boundaries. This could be implemented with a graphical display of constraints (see *Constraining the Selection of Weights* on page 93).

The bases representation of harmonic amplitude envelopes is well suited to graphically oriented resynthesis applications as well as a useful tool for timbre research. Realistic musical instrument envelopes can be reconstructed easily and quickly. The problem of real-time (additive) resynthesis could be solved by implementing the method on a computer with a hardware interface to a Digital Signal Processor.

Bibliography

[1] Daniel Arfib. Digital synthesis of complex spectra by means of multiplication of nonlinear distorted sine waves. *The Journal of the Audio Engineering Society*, 27(10):757–768, October 1979.

[2] John Backus. *The Acoustical Foundations of Music.* W. W. Norton, New York, 1969.

[3] James W. Beauchamp. Analysis and synthesis of cornet tones using nonlinear interharmonic relationships. *The Journal of the Audio Engineering Society*, 23(10):778–795, December 1975.

[4] James W. Beauchamp. Synthesis by spectral amplitude and "brightness" matching of analyzed musical instrument tones. *The Journal of the Audio Engineering Society*, 30(6):396–406, June 1982.

[5] Arthur H. Benade. *Fundamentals of Musical Acoustics.* Oxford University Press, London, 1976.

[6] Spencer Bennett and David Bowers. *An Introduction to Multivariate Techniques for Social and Behavioural Sciences.* John Wiley & Sons, New York, 1976.

[7] Kenneth W. Berger. Some factors in the recognition of timbre. *The Journal of the Acoustical Society of America*, 36(10):1888–1891, October 1964.

[8] E. G. Boring. *History of sensation and perception in experimental psychology.* Appleton-Century Crofts, New York, 1942.

[9] Richard C. Cabot, Michael G. Mino, Douglas A. Dorans, Ira S. Tackel, and Henry E. Breed. Detection of phase shifts in harmonically related tones. *The Journal of the Audio Engineering Society*, 24(7):568–571, September 1976.

[10] Gérard R. Charbonneau. Timbre and the perceptual effects of three types of data reduction. In Curtis Roads, editor, *The Music Machine: Selected Readings from Computer Music Journal*, pages 521–530. The MIT Press, Cambridge, Massachusetts, 1989.

[11] John M. Chowning. The synthesis of complex audio spectra by means of frequency modulation. *The Journal of the Audio Engineering Society*, 21(7):526–534, September 1973.

[12] William W. Cooley and Paul R. Lohnes. *Multivariate Data Analysis.* John Wiley & Sons, New York, 1971.

[13] Peter Dallos. Biophysics of the cochlea. In Edward C. Carterette and Morton P. Friedman, editors, *Handbook of Perception*, volume IV Hearing, chapter 4, pages 125–162. Academic Press, New York, 1978.

[14] P. B. Denes and E. N. Pinson. *The Speech Chain.* Anchor, 1973.

[15] Diana Deutsch, editor. *The Psychology of Music.* Series in Cognition and Perception. Academic Press, New York, 1982.

[16] Charles Dodge and Thomas A. Jerse. *Computer Music: Synthesis, Composition, and Performance.* Schirmer Books, Macmillan Inc., New York, 1985.

[17] Robert Erickson. New music and psychology. In Diana Deutsch, editor, *The Psychology of Music*, Series in Cognition and Perception, chapter 18, pages 517–536. Academic Press, New York, 1982.

[18] George A. Ferguson. *Statistical Analysis in Psychology and Education.* McGraw-Hill, Toronto, fifth edition, 1981.

[19] Harvey Fletcher, E. Donnell Blackman, and O. N. Geertsen. Quality of violin, viola, cello, and bass viol tones. *The Journal of the Acoustical Society of America*, 37:851–863, 1965.

[20] Harvey Fletcher, E. Donnell Blackman, and Richard Stratton. Quality of piano tones. *The Journal of the Acoustical Society of America*, 34(6):749–761, June 1962.

[21] Daniel J. Freed and William L. Martens. Deriving psychophysical relations for timbre. In Paul Berg, editor, *ICMC 86 Proceedings*, pages 393–405, San Francisco, 1986. International Computer Music Conference, Computer Music Association. Conference held in The Hague, Netherlands.

[22] J. J. Gibson. *The senses considered as perceptual systems.* Houghton, Boston, 1966.

[23] John M. Grey. *An Exploration of Musical Timbre.* PhD thesis, Stanford University, 1975. Available as Report No. Stan-M-2 from Stanford University Dept. of Music, Stanford, CA.

[24] John M. Grey. Multidimensional perceptual scaling of musical timbres. *The Journal of the Acoustical Society of America*, 61(5):1270–1277, May 1977.

[25] John M. Grey. Timbre discrimination in musical patterns. *The Journal of the Acoustical Society of America*, 64(2):467–472, August 1978.

[26] John M. Grey and John W. Gordon. Perceptual effects of spectral modifications on musical timbres. *The Journal of the Acoustical Society of America*, 63(5):1493–1500, May 1978.

[27] John M. Grey and James A. Moorer. Perceptual evaluations of synthesized musical instrument tones. *The Journal of the Acoustical Society of America*, 62(2):454–462, August 1977.

[28] Cyril M. Harris and Mark R. Weiss. Pitch extraction by computer processing of high-resolution fourier analysis data. *The Journal of the Acoustical Society of America*, 35:339–343, March 1963.

[29] Lejaren Hiller and Pierre Ruiz. Synthesizing musical sounds by solving the wave equation for vibrating objects: Part I. *The Journal of the Audio Engineering Society*, 19(6):462–470, June 1971.

[30] Lejaren Hiller and Pierre Ruiz. Synthesizing musical sounds by solving the wave equation for vibrating objects: Part II. *The Journal of the Audio Engineering Society*, 19(7):542–551, July/August 1971.

[31] Carleen Maley Hutchins. The acoustics of violin plates. *Scientific American*, 245(4):170–186, October 1981.

[32] David A. Jaffe and Julius O. Smith. Extensions of the Karplus-Strong plucked-string algorithm. *Computer Music Journal*, 7(2):56–69, Summer 1983.

[33] Kevin Karplus and Alex Strong. Digital synthesis of plucked-string and drum timbres. *Computer Music Journal*, 7(2):43–55, Summer 1983.

[34] Piotr Kleczkowski. Group additive synthesis. *Computer Music Journal*, 13(1):12–20, Spring 1989.

[35] Marc Le Brun. Digital waveshaping synthesis. *The Journal of the Audio Engineering Society*, 27(4):250–266, April 1979.

[36] Paul R. Lehman. Harmonic structure of the tone of the bassoon. *The Journal of the Acoustical Society of America*, 36(9):1649–1653, September 1964.

[37] William H. Lichte. Attributes of complex tones. *Journal of Experimental Psychology*, 28(6):455–480, June 1941.

[38] David A. Luce. *Physical Correlates of Nonpercussive Musical Instrument Tones.* PhD thesis, MIT, Cambridge, Massachusetts, 1963.

[39] David A. Luce. Dynamic spectrum changes of orchestral instruments. *The Journal of the Audio Engineering Society*, 23(7):565–568, September 1975.

[40] David A. Luce and Melville Clark Jr. Physical correlates of brass-instrument tones. *The Journal of the Acoustical Society of America*, 42(6):1232–1243, 1967.

[41] R.C. Mathes and R.L. Miller. Phase effects in monaural perception. *The Journal of the Acoustical Society of America*, 19(5):780–797, September 1947.

[42] M. V. Mathews, Joan E. Miller, and E. E. David Jr. Pitch synchronous analysis of voiced sounds. *The Journal of the Acoustical Society of America*, 33(2):179–186, February 1961.

[43] Stephen McAdams and Albert Bregman. Hearing musical streams. *Computer Music Journal*, 3(4):26–43,60, 1979.

[44] James R. Miller and Edward C. Carterette. Perceptual space for musical structures. *The Journal of the Acoustical Society of America*, 58(3):711–720, September 1975.

[45] James A. Moorer. *On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer.* PhD thesis, Stanford University, 1975. Available as Report No. Stan-M-3 from Stanford University Dept. of Music, Stanford, CA.

[46] James A. Moorer. Signal processing aspects of computer music—a survey. *Computer Music Journal*, 1(1):4–37, February 1977.

[47] James A. Moorer. The use of the phase vocoder in computer music applications. *The Journal of the Audio Engineering Society*, 26(1/2):42–45, January/February 1978.

[48] James A. Moorer, John M. Grey, and J. M. Snell. Lexicon of analyzed tones part 1: A violin tone. *Computer Music Journal*, 1(2):39–45, 1977.

[49] James A. Moorer, John M. Grey, and John Strawn. Lexicon of analyzed tones part 2: Clarinet and oboe tones. *Computer Music Journal*, 1(3):12–29, 1977.

[50] James A. Moorer, John M. Grey, and John Strawn. Lexicon of analyzed tones part 3: The trumpet. *Computer Music Journal*, 2(2):23–31, 1978.

[51] R. Plomp. Timbre as a multidimensional attribute of complex tones. In R. Plomp and G. F. Smoorenburg, editors, *Frequency Analysis and Periodicity Detection in Hearing*. A. W. Sijthoff, Leiden, 1970.

[52] R. Plomp, L. C. W. Pols, and J. P. van de Geer. Dimensional analysis of vowel spectra. *The Journal of the Acoustical Society of America*, 41(3):707–712, 1967.

[53] R. Plomp and H. J. M. Steeneken. Effect of phase on the timbre of complex tones. *The Journal of the Acoustical Society of America*, 46(2):409–421, 1969.

[54] Robert W. Ramirez. *The FFT: Fundamentals and Concepts*. Prentice-Hall, Englewood Cliffs, N.J., 1985.

[55] C. Radhakrishna Rao. The use and interpretation of principal component analysis in applied research. *Sankyā Series A*, 26:329–358, 1964.

[56] Jean-Claude Risset and Max V. Mathews. Analysis of musical-instrument tones. *Physics Today*, 22(2):23–30, February 1969.

[57] Jean-Claude Risset and David L. Wessel. Exploration of timbre by analysis and synthesis. In Diana Deutsch, editor, *The Psychology of Music*, Series in Cognition and Perception, chapter 2, pages 25–58. Academic Press, New York, 1982.

[58] Curtis Roads, editor. *The Music Machine: Selected Readings from Computer Music Journal*. The MIT Press, Cambridge, Massachusetts, 1989.

[59] Juan G. Roederer. *Introduction to the Physics and Psychophysics of Music*, volume 16 of *Heidelberg Science Library*. Springer-Verlag, New York, second edition, 1979.

[60] R. J. Rummel. *Applied Factor Analysis*. Northwestern University Press, Evanston Illinois, 1970.

[61] E. L. Saldanha and John F. Corso. Timbre cues and the identification of musical instruments. *The Journal of the Acoustical Society of America*, 36:2021–2026, 1964.

[62] Lawrence H. Sasaki and Kenneth C. Smith. A simple data reduction scheme for additive synthesis. *Computer Music Journal*, 4(1):22–24, Spring 1980.

[63] Richard A. Schaefer. Electronic musical tone production by nonlinear waveshaping. *The Journal of the Audio Engineering Society*, 18(4):413–417, August 1970.

[64] Keith W. Schindler. Dynamic timbre control for real-time digital synthesis. *Computer Music Journal*, 8(1):28–42, Spring 1984.

[65] J. F. Schouten, R. J. Ritsma, and B. Lopes Cardozo. Pitch of the residue. *The Journal of the Acoustical Society of America*, 34(8):1418–1424, September 1962.

[66] C. L. Searle. Speech perception from an auditory and visual viewpoint. *Canadian Journal of Psychology*, 36(3):402–419, 1982.

[67] A. W. Slawson. Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency. *The Journal of the Acoustical Society of America*, 43(1):87–101, 1968.

[68] Lawrence N. Solomon. Semantic approach to the perception of complex sounds. *The Journal of the Acoustical Society of America*, 30(5):421–425, May 1958.

[69] John Strawn. Approximation and syntactic analysis of amplitude and frequency functions for digital sound synthesis. In Curtis Roads, editor, *The Music Machine: Selected Readings from Computer Music Journal*, pages 671–692. The MIT Press, Cambridge, Massachusetts, 1989. reprinted from Computer Music Journal, Vol. 4, No. 3, Fall 1980.

[70] William Strong and Melville Clark Jr. Perturbations of synthetic orchestral wind-instrument tones. *The Journal of the Acoustical Society of America*, 41(2):277–285, 1967.

[71] William Strong and Melville Clark Jr. Synthesis of wind-instrument tones. *The Journal of the Acoustical Society of America*, 41(1):39–52, 1967.

[72] Barry Truax, editor. *Handbook for Acoustic Ecology*, volume 5 of *The Music of the Environment Series*. A.R.C. Publications, Vancouver B.C., 1978. The World Soundscape Project, R. Murray Schafer, series editor.

[73] G. von Bismarck. Sharpness as an attribute of the timbre of steady sounds. *Acustica*, 30:159–172, 1974.

[74] G. von Bismarck. Timbre of steady sounds: A factorial investigation of its verbal attributes. *Acustica*, 30:146–159, 1974.

[75] H. L. F. von Helmholtz. *On the sensations of tone*. Dover, New York, 1954. translated from the 1877 German original.

[76] Lage Wedin and Gunnar Goude. Dimension analysis of the perception of instrumental timbre. *Scandinavian Journal of Psychology*, 13:228–240, 1972.

[77] David L. Wessel. Timbre space as a musical control structure. *Computer Music Journal*, 3(2):45–52, 1979.

[78] Stephen A. Zahorian and Martin Rothenberg. Principal-components analysis for low-redundancy encoding of speech spectra. *The Journal of the Acoustical Society of America*, 69(3):832–845, March 1981.