CENTER FOR COMPUTER RESEARCH IN MUSIC AND ACOUSTICS AUGUST 1987

Department of Music
Report No. STAN-M-42

TOWARD A THEORY OF TIMBRE

by

Yee On Lo

Research sponsored by

The System Development Foundation

CCRMA
DEPARTMENT OF MUSIC
Stanford University
Stanford, California 94305

TOWARD

A

THEORY OF TIMBRE

by

Yee On Lo

1987

ABSTRACT

If musical listening amounts to discovering structures or meaningful patterns of acoustic events, then, for a piece of music to make sense, it must be perceptually organizable. Unfortunately, the acoustic events mentioned are primarily oriented towards pitch and rhythm. The structure-discovering concept has not been applied to the perception of timbres and their relationships so as to achieve musically successful composition of timbral patterns. The problem is due in part to the multidimensional nature of timbre and to the complexity of the acoustic elements that affect its perception.

This thesis is the formulation of an absolute description of timbre that permits an arbitrarily fine description of timbre in terms of its acoustic features in a hierarchical fashion. The description or definition provides a coherent approach to analysis and synthesis of musical timbre. Musicality is shown to be correlated to organizability of the acoustic pattern of vibration. The hierarchical organization can be formalized into either an ordinal or metric perceptual importance tree of acoustic features. These trees permit systematic organizations of timbre patterns in terms of the timbres' internal dynamics. In other words, the relationships previously obscured by the complex or multidimensional nature of individual timbres are now dynamically projected and therefore clearly illuminated into subtrees of appropriate classes.

The description is justified perceptually by casting the auditory system as an active observer that performs pattern recognition on the space-time mechanical response pattern of the inner ear. Such pattern recognition activity is in fact made possible by the highly redundant response characteristics of the basilar membrane—a fact strongly supported by physiological data. The transformation from the passive response behavior of a system of mechanical resonators into a system of features important to the perception of timbre can be modelled on Minsky's *Society of Mind* idea applied to the agents and agencies in the auditory processor. The perceptual model developed here bridges the gap between the low-level signal processing in sound analysis/synthesis and the higher level perception of musical timbre necessary for successful utilization of timbres and thus serves broader compositional purposes than are currently possible.

CENTER FOR COMPUTER RESEARCH IN MUSIC AND ACOUSTICS AUGUST 1987

Department of Music
Report No. STAN-M-42

TOWARD A THEORY OF TIMBRE

by

Yee On Lo

If musical listening amounts to discovering structures or meaningful patterns of acoustic events, then, for a piece of music to make sense, it must be perceptually organizable. Unfortunately, the acoustic events mentioned are primarily oriented towards pitch and rhythm. The structure-discovering concept has not been applied to the perception of timbres and their relationships so as to achieve musically successful composition of timbral patterns. The problem is due in part to the multi-dimensional nature of timbre and to the complexity of the acoustic elements that affect its perception.

This thesis is the formulation of an absolute description of timbre that permits an arbitrarily fine description of timbre in terms of its acoustic features in a hierarchical fashion. The description or definition provides a coherent approach to analysis and synthesis of musical timbre. Musicality is shown to be correlated to the organizability of the acoustic pattern of vibration. The hierarchical organization can be formalized into either an ordinal or metric perceptual importance tree of acoustic features. These trees permit systematic organizations of timbre patterns in terms of the timbres' internal dynamic. In other words, the relationships previously obscured by the complex or multidimensional nature of individual timbres are now dynamically projected and therefore clearly illuminated into subtrees of appropriate classes.

The description is justified perceptually by casting the auditory system as an active observer that performs pattern recognition on the space-time mechanical response pattern of the inner ear. Such pattern recognition activity is in fact made possible by the highly redundant response characteristics of the basilar membrane -- a fact strongly supported by physiological data. The transformation from the passive response behavior of a system of mechanical resonators into a system of features important to the perception of timbre can be modelled on Minsky's *Society of Mind* idea applied to the agents and agencies in the auditory processor. The perceptual model developed here bridges the gap between the low-level signal processing in sound analysis/synthesis and higher level perception of musical timbre necessary for successful utilization of timbres and thus serves broader compositional purposes than are currently possible.

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

This research was supported (in part) by the System Development Foundation under Grant SDF #345. The views and conclusions contained in this document are those of the author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of Stanford University or of the sponsoring foundation.

ACKNOWLEDGMENTS

In the last few years, i have had the priviledge of working at the excellent facility of CCRMA and would like to thank its director, Professor John Chowning, and the staff there. The author has benefitted from discussions with and help from practically everyone at CCRMA. In particular, Jeff Borish helped me get started at CCRMA. Julius Smith gave me all kinds of suggestions on signal processing. Bob Shannon shared with me some of his insights in speech and audio, and gave me encouragement. Bill Schottstaedt came to my rescue a few times when i got mired in errors. Xavier Serra generously supplied his Marimba analysis data. Servio Marin introduced to me the work of Pierre Schaeffer. Matt Fields, Doug Keislar and Stanislaw Kropowicz were always helpful; they provided me critical comments and valuable insights into many aspects of my research on timbre.

I thank Professors John Pierce, Al Bregman and Richard Felciano for having contributed useful comments and suggestions regarding either the contents or presentation (written or oral) of my thesis. I am grateful to my thesis reading and oral committee members for their valuable time. In particular, Professors Dorothy Huntington, Bill Poser, and Earl Schubert have brought to my attention a number of mistakes and provided suggestions for important improvements, although i am responsible for any remaining errors.

I am indebted to my past teachers, among whom i would like to mention in particular Professors Harry Ore, Robert H. Walker, Fred Lerdahl, Aram Thomasian, John B. Anderson, John Chowning, and Earl Schubert. But it was Professor Schubert who had made this thesis possible. He did this by teaching me Psychoacoustics and showing me the complexity of nature, by insisting that i listen to my own ears who do the real listening and do not lie, by urging me to make good use of the digital computer, by having my interest at heart and giving me great freedom to pursue a subject of my own choosing, by being always available for discussion and helpful, by frequent encouragement and occasional reproof, and by having financially supported me throughout most of my years at Stanford.

I would also like to acknowledge the long standing encouragement of my former physics advisor Professor Robert H. Walker of the University of Houston, my father and mother, and my friend and colleague Daniel Hitt of the mathematics department of the University of California at Santa Cruz. In particular, Dan has sacrificed innumerable hours type-setting this thesis, — in addition to providing me with the opportunity to discuss with him most of the ideas in this thesis as soon as they sprouted in my head so that they had a chance to grow. Of course, any error in this thesis is my sole responsibility.

Finally, i am most grateful to my family for their emotional, moral and even financial support during the critical time of my last two year's work when i needed it to live.

Table of Contents

Chapter I: Goals and Historical Review

Chapter II: Theory

Chapter III: Analysis, Synthesis, and Tests of Theory

Chapter I: Goals and Historical Review

1.0 Overview.

This thesis attempts to develop a more appropriate treatment for timbre, especially timbre with significant dynamic characteristics. This treatment is in the form of a theory, in the sense of a framework and a formal description. The framework includes a perceptual model, a formal and a dynamic description of timbre, a hierarchical model of feature composition, a model of timbre interpolation consistent with the dynamic description, and an approach to analysis/synthesis of timbre in terms of acoustic distinctive features. These pieces together form what we consider as a necessary framework for a more adequate description of timbre in general. We will show how this is the case through discussion in the next three chapters. We will show many examples and cite appropriate parts of the literature to support our arguments. Examples are necessary for an experimental field as psychoacoustics, even though our aim is the development of the theory.

We will first articulate our goals in 1.1. We will then discuss Helmholtz's contribution to our knowledge of timbre in fair detail, partly because most of the discussion pertains to how his mechanical model may be applied to the study of transient perception as well as the stationary signals he developed it for. (The equations and mathematics developed are to complement Helmholtz's own treatment for the more comprehensive view taken.) We also discuss it because of what he did and did not do. His strengths and weaknesses provide us with an example from which we can learn and which we can adapt to our own advantage.

We will then discuss Winckel's work. In borrowing his prophetic idea of sound perception as a communication problem, we will discuss in detail some of the source and receiver characteristics and what they mean for timbre perception. The interpolated nature of physical changes will be applied repeatedly later.

Grey's idea of timbre research as analysis and synthesis of timbres in terms of distinctive features, as simplification of complexity in acoustic data, as exploration of timbre relationships and structure of timbre space, and as interpolation of timbres will be the focus of and inspiration for our new approach to the treatment of timbre.

The idea of organization, relationship, and hierarchical perception of Erickson will be seen as the central thrust of our theory.

Schaeffer's keen observation into the dynamic nature of timbre, his laws on timbre perception, and his linguist's view of timbre research have greatly helped our effort to put together a coherent treatment.

Beauchamp, Charbonneau, Chowning, Ehresman, Morrill, Risset, Wessel, and many others have contributed to the outlook of this thesis, especially to the idea of pattern recognition (Charbonneau), the idea of distinctive acoustic features (Risset, Chowning), the trade between algorithmic and data complexity (Beauchamp), the dynamic character of timbre relationships (Morrill, Chowning), and approaches to the study of timbre relationships in general (Ehresman and Wessel).

In chapter II, we start with a gradual development of the fundamental notion of timbre so that the current concept of timbre will be clear by the time we start developing the principal part of the theory. We feel this is necessary because of the complex nature of timbre and because of the confusing state of our knowledge about timbre.

The dynamic character of sound and the ear's ability to follow it provide timbre with a dynamical character as well. This is certainly true of natural timbres (i.e., timbres arising in a natural acoustic environment, as opposed to electronically synthesized ones). However, synthetic timbres turn out to be less interesting unless a certain "natural" quality is somehow added. Thus we interpret timbre in a broad sense to be a function of time, and to include all the nuances necessary for the perceptual identical resynthesis of it. Therefore, for our purposes, where synthesis is an integral part of the ultimate goal of analysis, it serves no useful purpose to separate the aspects of the perceived sound known by linguists and phoneticists as vowel quality from the notion of timbre when speech vowels are concerned. It is equally unnatural, for our purpose, to divide a speech sound, or a continuous group of speech sounds into phonemic units, assigning timbral description to individual phonemes and ignoring the transitions between them. Similarly, the timbral quality induced by changes in acoustic features that also induce changes in pitch or loudness must be part of the notion of timbre if we hope to perceptually *duplicate* the timbre. Therefore timbre will also be function of the acoustic parameters that control pitch and loudness of the sound as well.

After the development of the notion of timbre, we survey the fundamental properties of Helmholtz's mechanical model of timbre perception. We observe that

in order to achieve Grey's goal of developing an analysis and synthesis approach for timbre in terms of distinctive features, we must introduce an active observer role to the mechanical model in analogy to Minsky's agent idea in his treatment of the *Society of Mind.* When the problem becomes a pattern recognition one, we then argue for a set of fundamental acoustic features that are necessary to the analysis and synthesis of timbre in the perceptual sense. In other words, we will find out what acoustic features, i.e., in what form the acoustic waves, contribute in the most *direct (natural)* ways to the perception of timbre. As a result, we derive an absolute and dynamic description of timbre in terms of hierarchical composition of acoustic features.

Throughout, our purpose is to correlate what we hear with what is in the signal. Therefore, we take a system view, and by the word *ear* we refer to the entire auditory system, including both the mechanical and higher processing parts. The mechanical part interfaces directly with the acoustic waveform. The higher auditory processing part performs active organization on the acoustic information received at the mechanical level. We will be specific when we are talking about a specific organ, such as the cilia, the basilar membrane, or the cochlea.

In chapter III, the analysis and synthesis of timbre in the perceptual sense is then translated into algorithms performed on a digital computer. Their performances are then analyzed, some aspects of the theory tested, and experiments for other aspects of the theory are proposed.

We have consciously attempted to imitate the methodology of Helmholtz, and figure 1.0 shows two block diagrams that summarize the similarities and differences between our approach and the master's. The central issue was: What is the correlation between a timbre and its physical data? Helmholtz's answer was Ohm's acoustic law. He justified this answer by studying the question of how we might perceive timbre. He answered this question by means of the mechanical resonator model of the inner ear. With the central issue settled, he then described timbre relationships in terms of similar characteristics in their Fourier magnitude spectra.

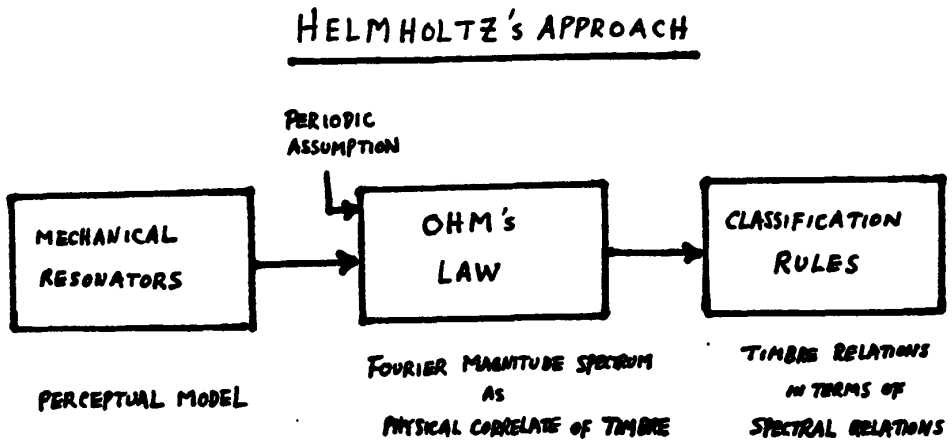


figure 1.0(a)

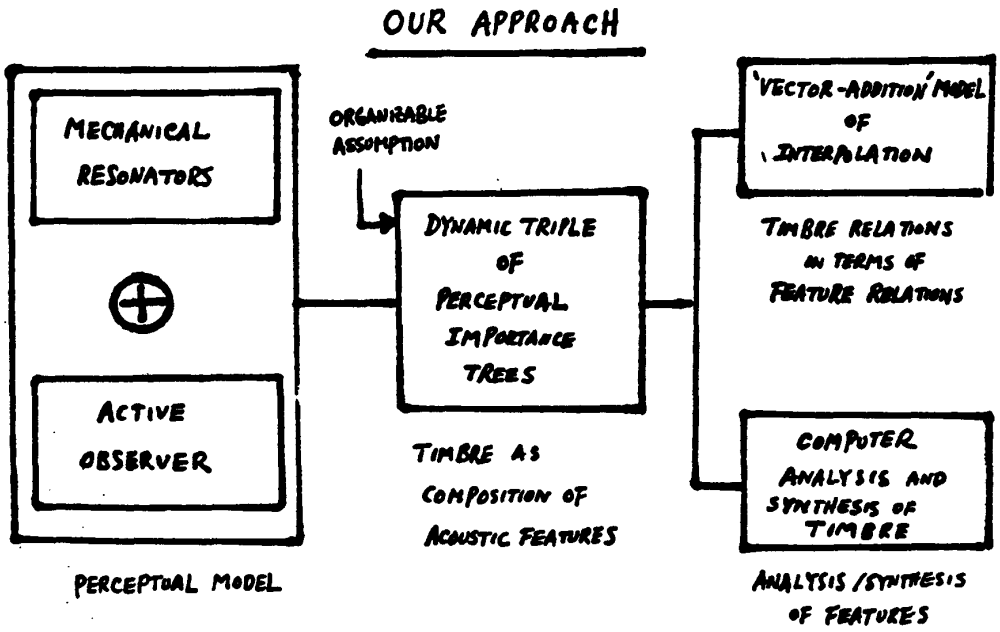


figure 1.0(b)

1.1 Statement of Purpose.

In timbre studies, we want to ultimately delineate the relationship among timbres as Mach, the famed nineteenth century psychophysicist, might have suggested (see [Yougrau and Mandelstam, 1968]). These relationships in the form of a structure should permit us to generate sounds with corresponding perceptual relationships. But in order to do that we need to know the relationships between the perceptual data and the physical data so that we can control the generation of timbres. However, given the multidimensional nature of timbre, or, the complexity of all the waveforms possible, it is crucial to try to understand the process of timbral perception in order to discover the relationships between the physical data and the perceptual data. This is particularly true in view of the fact that in dealing with musical timbre, including speech timbre, one is in the curious position of dealing with a control space of huge dimensionality on the one hand, and a perceptual space of apparently much smaller dimensionality, on the other. This apparently paradoxical view, although it cannot be either experimentally or theoretically proven at the present state of our knowledge, is nevertheless widely held among psychoacousticians and seems to make a lot of sense based on a large amount of evidence scattered across the literature of hearing and speech, linguistics, music theory, and cognitive psychology.

There are only relatively few degrees of freedom in speech production and there is only a (small) finite set of phonemes in any human language. In languages, there are only a small number of ways words can be strung together to make meaningful sentences. Similarly, in music, the listener often finds chaos in composition when too much freedom is taken in combinations of acoustic events that generate the music. Even composition that is very structured from an intellectual point of view can be perceived as chaotic when the complexity is too great; this phenomenon seems to stem at least partly from the fact that our auditory channel has a very finite capacity. The total information rate the ear can handle at any moment is quite limited. If this is true on a macro-time scale also, we should expect the same on a micro-time scale as well, since we are dealing with the same physical communication channel in either case.

But if a musical composition involves higher level information processing, then the observation must be even more true for micro-temporal acoustic events, since

there are fewer resources available to discover structures in them. So it is reasonable to believe that musical timbre is distinguishable from noise based on the actual or irreducible rate of information the sonic event contains.

In other words, musical timbre corresponds to acoustic events where there is a high degree of redundancy embedded in the wave fluctuation, whereas noise has very little redundancy and therefore over time, the ear is loaded with more information than it can take. To put it in another language, we may speculate that musical timbre corresponds to acoustic events in which the ear can organize and make sense of the waveform, but noise corresponds to acoustic events in which the ear cannot. Of course, this is the basis of Helmholtz's theory of timbre [Helmholtz, 1877], in which he considers musical timbre to arise from periodic acoustic signals where redundancy greatly reduces the control space dimensionality of timbres under study.

But Helmholtz's timbre in its strict sense can be realized only by electronic means (because non-electronic devices, including tuning forks, are incapable of producing exactly periodic signals), and such sounds are lifeless and boring, at least compared with natural signals. (Of course the timbres Helmholtz actually studied were not exactly periodic.)

We are interested in natural timbres, which exhibit strong transient characteristics. But the mere fact of transience does *not* therefore imply that such waveforms are *not* organizable; and we will show how, by doing analysis and synthesis on a strongly transient marimba tone, as well as on other natural timbres, such waveforms can in fact be organized.

Now, one might argue that the distinction between musicality and noise is not absolute. This is true, and in fact, the contextual nature of musicality versus noise can be articulated with the same language that involves the notions of redundancy, structure, and organizability. We will in fact explore this issue in detail in this thesis.

So, while the physical space of all acoustic waveforms possible is indeed large in dimensionality, the perceptual space may be much smaller. The ear is often said to prefer data reduction, but it seems that the ear will do so only when there is redundancy, when there is structure, or when the ear can perform organization on the signal. However, the notion of organization seems to go further. Although we

can model a certain class of damped periodic signals with a few resonators, say n, defined by their characteristic frequencies and damping constants, thus getting a dimensionality of $2n$, there is experimental evidence that the ear seems at least to a first approximation to ignore certain dimensions in the perceptual space, if the mapping from the control space to the perceptual space is one-to-one (see [Plomp, 1970]).

At the same time, though, our experience shows that small variations in the sample values of a digitized waveform can change the timbral quality upon careful listening. Thus it is reasonable to expect that acoustic features are important to perception of timbre, and are *hierarchically* organized where lesser features surface only upon careful comparative listening. We believe that a rigorous (quantitative) exploration of this idea is possible and will treat it in terms of what we call perceptual importance trees in chapters II and III.

The goal of this thesis is to formulate a dynamic theory of timbre based on what we currently believe to be a probable model of timbre perception such that

(1) timbre will be described in a way consistent with the dynamic character of the sound waves that enter the ear;

(2) a universal language will be found that will enable us to describe a diverse collection of timbres; and

(3) a timbral operating environment will emerge as a result of (1) and (2) that will provide more precise control over the timbres we want to generate and allow us to generate them efficiently.

The fundamental issue here is the relationship of the precisely observable acoustic waveform of a sound event to the less precisely or reliably observable but nevertheless observable percept of timbre—a typical situation where the human response is concerned. Perceptually, a sound event can be characterized by its length, loudness, pitch, and timbre. For a sound event, the perceived length is not trivially related to its objective counterpart. In fact, it is a function of the dynamics or microdynamics of the sound. For instance, a ten millisecond long woodblock sound does not sound proportionately shorter than a 700 millisecond (70 times as long) marimba tone or an /a/ voice sound. In fact, the /a/ sound mentioned has an octave drop and it seems longer than a marimba tone of the same physical duration. However it is clear from the definition that perceptual length of a sound event is a

scalar.

Loudness is also a function of the dynamics or microdynamics of a sound event, and it is usually thought of as a scalar also. But outside the laboratory environment, loudness of a sound event is generally a dynamic quantity, i.e., something that can change with time. Similarly pitch also is a dynamic quantity and has been recognized as such as evidenced by such terms as "pitch trajectory", "pitch bend", "pitch glide", etc.

This is of course a consequence of the fact that sound as an event is fundamentally a process, i.e., an occurrence in progress driven by the arrow of time, and the ear is capable of following it (with its short-time auditory windows). So why is timbre not described in terms consistent with the dynamic character of the sound wave? To put it another way, is there any reason why timbre should not be described in terms of a timbre trajectory similar to pitch? And since timbre is a function of changes in the acoustic waveform, do we hear a timbre trajectory, if we listen carefully? From experience, some small modification in the waveform of a given sound brings about some small but noticeable change in timbre (as in timbre interpolation). Do we hear the modification in the waveform that evolves gradually over time? These questions provide us a point of departure in our attempt to really understand the complex behavior of timbre in terms of the dynamics of the waveform.

Concerning a timbral language, we should remark at the outset that from the standpoint of evolution, one can assert that there is no need for such a language, by arguing that the function of timbre (from the receiver's point of view) is merely to determine what the source is or what it is doing. For example, the ear cares only as far as to discover whether the sound signifies the presence of danger, food source, or simply noise. When sound impinges upon the ear drum, the receiver would say,

"Is it a predator?"

"Does it sound like it is hungry?"

"Is it prey?"

"Does it sound like it is so relaxed that I could make a catch?"

"Is it just noise?"

"Let's find out!"

So one might argue that all we need is a discrete dictionary of timbre that serves

the need of our survival, and maybe this is exactly what we have. Therefore, there is really no need for a relational description of timbre, and if we try, we may fail to discover any. This observation may turn out to be the truth, but at this point, we don't know. But we have a need, namely that of (musical) timbre composition, and the seemingly unlimited potential for timbre generation and manipulation by digital means is waiting with impatience. The more scientific minds would also like to discover the structure, if any, of timbre space. And at the same time, we have good evidence from research in speech, especially continuous speech, for the innate potential of the ear to understand relationships of timbre. One should not forget that the ear has evolved for millions of years from the time when it existed only as a fin on a fish's belly, and when it knew only about predator and prey and what they were doing. In its primitive form, language exists among all higher animals, including birds. (The aggressive calls are known to biologists to exhibit certain specific patterns of transition in spectral energy among most species which make such calls. Similarly, the warning calls, the social calls, and the mating calls possess other specific patterns of spectral transition.)

In order to make a spoken language as sophisticated as ours to work, there is good reason to believe that our ears have been highly trained to recognize timbral relationships much more complicated than "What is the source doing?" It is well known that perception of continuous speech relies strongly on transitions, especially between consonants and vowels. The more irregular the vibration pattern of the consonant is, the more strongly the ear relies on the transitions. In other words, some form of timbre interpolation already exists in the perception of continuous speech. Therefore, although the universe of timbre is larger than that of speech timbre alone, there are good reasons to believe some sort of innate language in timbre already exists in the ear and is waiting to be articulated.

As it turns out, the first and second goals that we stated at the beginning of this chapter are quite inter-related: The diversity of timbres that occur in nature calls for a language that can describe most of them without awkwardness. Also, unless we have some way to bring together very diverse timbres in the sense of filling the gaps among them, *viz.*, timbre interpolation, we really don't understand the character of individual timbres very well. And at the same time, unless we have a general description of individual timbres in terms of the properties of their

physical stimuli, we cannot realize timbre interpolation. We need a description of the relationship between a timbre and its physical stimulus. We need a description of the relationships among timbres. Above all, we need a description that will unify both of these. We shall see that a dynamic theory of timbre could be the basis for achieving these goals together.

But then of course such a theory must necessarily provide the tools, the control required to generate these timbres that fill the timbre space. The tools must be precise and at the same time easily manipulable so that we can move around easily among known timbres, and discover the structure of the space bounded by them. Naturally, these tools, organized in a research environment, ought to be equally important for artistic creation, even in view of the abundance of digital synthesis techniques available. The success of the frequency modulation (FM) sound synthesis technique, pioneered and developed by Chowning [Chowning, 1973], for example, should not obscure the fact that timbre space remains much larger than the collection of timbres existing techniques can produce. Part of this gap is due to limitation in any analysis technique (if one exists at all) for a particular synthesis algorithm chosen.

As it turns out, for the important class of natural timbre where transients are important perceptual characteristics, Fourier analysis and its time-dependent versions are not as powerful as they initially promised to be based on mathematical considerations alone. The range of acoustic waveforms in nature is too wide to be described by any particular set of special functions. The ear, which has successfully adapted itself to its environment through the ages of evolution, is too versatile to limit its capability to short-time Fourier analysis (STFA) or judgments based only on some mean-squared criterion. (See [Rabiner and Shafer, 1978] for a discussion of STFA.) In fact, STFA tends to smear data in both the time and frequency domains (see [Claasen and Mecklenbräuker, 1980]). The windows of stationarity for acoustic transients of many natural timbres often conflict with the time window for sufficient frequency resolution. But more importantly, if the ear actively organizes the acoustic signal it receives to discover structure, and separate redundancy from features (these are complementary processes for successful and intelligent communication), SFTA and other existing sound analysis techniques certainly do not provide this level of analysis. In other words, perceptually important acoustic feature extraction

is not part of these existing methods. They involve what we will call the *passive* observer approach to analysis. For example, if one takes the analysis data from a SFTA consisting of phase and magnitude of frequency channels for a natural timbre, say that of a saxophone, and gradually varies the phase or increases the magnitude of one of the harmonics, there will be an abrupt appearance of a separate component in the timbre. In other words, there is a certain perceptual discontinuity in response to continuous variation of the analysis data. Perceptual discontinuity in response to analysis data variation also happens in other existing techniques, such as linear prediction coding (see [Moorer, 1979]).

In general, we remain unable to recreate many natural timbres, much less interpolate among them, using Fourier or other existing techniques. Lerdahl, a music theorist and a composer in his own right, pointed out very recently (1985) that the much cited dream of Arnold Schoenberg, timbral composition, remains unfulfilled. In fact Schoenberg was very interested in integrating naturally spoken speech timbres (apparently recognizing their richness and expressiveness) into music that includes the traditional acoustical instruments. Does the so-far unsuccessful effort in using timbre as a structural element in musical composition have something to do with the intrinsic sensory difference between timbre in audition and color in vision (note that color has been a successful structural element in painting)? Or is it because musicians have not had a timbre operating environment as conveniently empirical as the painter's palette. It is therefore a major part of our goal to find such a composer's timbral palette to see whether timbral compositions, compositions for which timbre is the principal structural/functional element, are realizable.

We have briefly explained our goal, but in order to really appreciate the merits of having a dynamic theory of timbre we must first review what has been done to provide a proper context for such a purpose. Although we believe that the value of the theory lies in its ability to support a timbral operating environment briefly described above i.e., the theory can be properly assessed only in terms of what it can do in the future, (and we will get into the detail some more after we develop the theory in later chapters), its present existence (in a form to be brought forth later in this thesis) must however be justified by more than its purpose (as it relates itself to the future). And the only way it can justify its present existence is through its relationship with the past, i.e., the state of our knowledge, and not necessarily

only from the relatively young field of psychoacoustics. So we will begin with other people's work which has had a strong influence on this one.

1.2 The Mechanical Model of Timbre Perception of Helmholtz.

Helmholtz was the first to clarify the mechanical basis of timbre perception. Although much has been written about his contribution to the study of tone perrception in the steady-state, the mechanical role in perceiving acoustic transients has never been discussed in clear detail. We believe that despite the fact that Helmholtz chooses to confine his study to only periodic tones perceived in the steady-state and understandably so in view of the technology then available, a good understanding of the mechanical behavior of the ear is important to understanding the percepts of acoustic transients and hence timbre in general. Therefore, our discussion will be a little more detailed.

1.2.1 Noise Versus Musical Timbre.

In his *Sensations of Tone*, the first order of business is to make a distinction between noise and musical timbre. He describes noise as acoustic vibrations with irregular form. By contrast, musical timbres correspond to vibrations with highly regular form. Specifically, he observes that sustained tones of pleasing quality are pitched—which he regards as an indication of periodic repetition in the air vibration. While perfectly periodic tones are a mathematical idealization, approximately periodic tones are about the only thing that the technology of his day can adequately handle, so it is not unreasonable for him to restrict his attention to them. But unfortunately he goes too far in stating that musical timbres are of periodic form, as we shall see later. For example, he would not consider the timbre of a wood block sound musical because it is not sufficiently pitched. In fact, for Helmholtz, this sound has no periodic structure in the vibration form and is too short to be easily observed or analyzed in his time.

1.2.2 Existence of a Fourier Series Representation for the Acoustic Waveform.

Although the possibility of expressing a periodic vibrating form as a Fourier series, i.e., as a sum of sines and cosines of multiples of a certain fundamental frequency, was reported by Daniel Bernoulli in the eighteenth century (before Fourier came along), the issues of existence and convergence remained under fierce debate until this century. The difficulties lie with the number and behavior of discontinuities in the function. Consider, to take a concrete case, the tangent function. It

is a periodic function but its discontinuities at $\frac{\pi}{2}$ and $\frac{3\pi}{2}$ have the same order of growth as does the discontinuity of $1/z$ near $z = 0$ so that the integrals defining its Fourier coefficients don't make sense in the ordinary interpretation. Another case is the impulse train, which has discontinuous derivatives (as does any discontinuous function) and its Fourier coefficients do not even go to zero (so their sum cannot possibly converge in the usual sense). Still another case is a periodic function obtained from a periodic extension of truncated white noise, which has infinitely many jump discontinuities. And two periodic functions f and g differing only at one point in each period must have the same Fourier series if it exists for them. Then the question is how can one recover the functions from the Fourier series? That is, which, if either, is the "right" one? In fact, Euler, the great Swiss mathematician of Bernoulli's time, recognized this problem and received the latter's report with skepticism. The existence issue was finally settled with the formulation of Lebesgue's theory of measure.

But what do these oddly behaving mathematical functions have to do with perception of tones? The point is when we talk about air vibration, we must recognize that the air molecules execute random motion on an individual basis. Fortunately, the pressure wave generated by the excitation of an acoustic object exists as an average of these random behaviors, and as such is continuous and even smooth in most cases. Therefore there is no question about either existence or convergence (of the Fourier series) in the class of musical tones that Helmholtz is interested in, but in general we have to bear in mind that not every arbitrary periodic function has a Fourier series—and if it does, it might not converge. This is an important point in view of all the possible periodic waveforms the digital computer can generate.

1.2.3 The Ear as a "Fourier Analyzer".

1.2.3.1 Physiological Basis.

Helmholtz observed that physiologically it is possible for the ear through the basilar membrane of the cochlea to behave as a bank of mechanical resonaters, each with its characteristic frequency, damping, and bandwidth. (Note that the last two are not independent parameters.)

1.2.3.2 A Mechanical Filter Model.

Helmholtz sets up the necessary mathematical apparatus to model the behavior

of a linear mechanical resonator, first as a simple pendulum and then as a membrane. As it turns out, the membrane response is very similar to that of a bank of mechanical resonators each behaving like a simple pendulum, under appropriate assumptions on the stretching properties of the basilar membrane in the axial and radial directions.

1.2.3.3 Sympathetic Vibration.

Helmholtz then shows how an "arbitrary" periodic motion containing a particular Fourier component can induce the mechanical resonator which he describes to vibrate strongly, no matter how weak that component is, provided that sufficient time is available to build up that sinusoidal motion and that its frequency matches the characteristic frequency of the resonator through a mechanism of what is called "sympathetic resonance." He points out that one can readily experience this kind of behavior with ordinary (macroscopic) mechanical devices at a rate much slower than the audio frequencies so that we can monitor the sinusoidal motion visually.

1.2.3.4 Training and Experimental Verification.

Finally, Helmholtz shows that the human ear can indeed be trained to pick out these Fourier components to an extent limited by the mechanical properties of the membrane (for example, the rule-of-thumb constant-Q behavior of a filter bank implies the harmonic partials will tend to be less and less resolved as their indices increase). The training essentially involves first directing the listener's auditory attention to the frequency location of a particular Fourier component contained in the periodic vibration by presenting an independent sinusoidal stimulus of that frequency before the complex vibration is presented. Notice that the analytical task involves additional context not provided in normal listening.

1.2.3.5 The Ear as a Fourier Analyzer?

The ability to pick out Fourier components however does not automatically imply that the ear is a Fourier analyzer in the sense of recovering the coefficients of the signal's Fourier series even within the context of Helmholtz's resonator model. In general, all it means is that the ear detects energy at that particular frequency of the Fourier component. (According to his model, each resonator has its own characteristic impulse response.) In the special case of a periodic signal, the steady-state response is sinusoidal, but the gain and phase will be different as a result of the resonator's characteristic response. Therefore the strength distribution of the

response of the resonator bank (even in its steady state) is not the magnitude spectrum of the signal's Fourier series. The shape of the frequency response curve supports this fact. This fact is independent of whether the signal is stationary or nonstationary.

1.2.4 A Perceptual Consequence of Helmholtz's Analytical Model.

The purpose of Helmholtz's *Sensations of Tones* is to discover the laws that govern the causal relationship between the sounds as they independently exist and the images the ear presents to the brain. A consequence of Helmholtz's analytic model of audition is that he can bypass the detailed operation of the ear and directly correlate the auditory images with the physical data. We will see how in the following subsections.

1.2.4.1 Causality—The Relation between the Sense Data and the Physical Data.

Contrary to the phenomenalistic theory of knowledge of Mach, which asserts that "a scientific theory can do no more than describe systematically the simple sense data and the relations between them and that physical 'reality' can never be causally explained," [Yourgrau and Mandelstam, 1968, p. 169] Helmholtz approaches his subject by attempting to discover first the physical basis of tone perception. That is, he tries to find the analytic process of tone perception that forms the *basic* cause of the sense reality. Then, he looks for the relationship between the sense data and their physical correlates. We will call this the *secondary* or *derived* cause of the sense reality.

1.2.4.2 Ohm's Acoustic Law as a Consequence of the Analytic Process of Audition.

The *basic* cause, in the form of the mechanical response of a resonator bank, allows Helmholtz to go back to the Fourier series of the periodically varied stimulus to find clues for the patterns of organization. The universal order he finds evidence for is known as Ohm's acoustic law. This law states that the timbre of a periodic tone is essentially a function of the distribution of the magnitudes of its Fourier components. As a consequence, the notions of a spectral envelope and formant regions have long been recognized as a perceptually important organizing element in the classification of vowel sounds and many acoustic instrument tones. Ohm's law and these notions can thus be considered to be the *general* form of the derived

cause of the sense reality.

1.2.4.3 Detailed Correlations between the Sense Data and Physical Data

In specific terms, Helmholtz restricts himself to mapping classes of physical data into subjective categories. In a certain way, the classification leads to a relation among sense data, i.e., data which belong to the same descriptive category. And he casts his classification into rules [Helmholtz, 1877, pp.118-119]:

> "1. *Simple Tones,* like those of tuning-forks applied to resonance chambers and wide stopped organ pipes, have a very soft, pleasant sound, free from all roughness, but wanting in power, and dull at low pitches.
>
> "2. *Musical Tones,* which are accompanied by a moderately loud series of the lower partial tones, up to about the sixth partial, are more harmonious and musical. Compared with simple tones they are rich and splendid, while they are at the same time perfectly sweet and soft if the higher upper partials are absent. To these belong the musical tones produced by the pianoforte, open organ pipes, the softer piano tones of the human voice and of the French horn. The last-named tones form the transition to musical tones with high upper partials; while the tones of flutes, and of pipes on the flue-stops of organs with a low pressure of wind, approach to simple tones.
>
> "3. If only the unevenly numbered partials are present (as in narrow stopped organ pipes, pianoforte strings struck in their middle points, and clarinets), the quality of tone is ***hollow***, and, when a large number of such upper partials are present, ***nasal***. When the prime tone predominates the quality of tone is ***rich***; but when the prime tone is not sufficiently superior in strength to the upper partials, the quality of tone is ***poor***. Thus the quality of tone in the wider open organ pipes is richer than that in the narrower; strings struck with pianoforte hammers give tones of a richer quality than when struck by a stick or plucked by the finger; the tones of reed pipes with suitable resonance chambers have a richer quality than those without resonance chambers.
>
> "4. When partial tones higher than the sixth or seventh are very distinct, the quality of tone is ***cutting*** and ***rough***. The reason for this will be seen hereafter to lie in the dissonances which they form with one

another. The degree of harshness may be very different. When their force is inconsiderable the higher upper partials do not essentially detract from the musical applicability of the compound tones; on the contrary, they are useful in giving character and expression to the music. The most important musical tones of this description are those of bowed instruments and of most reed pipes, oboe (hautbois), bassoon (fagotto), harmonium, and the human voice. The rough, braying tones of brass instruments are extremely penetrating, and hence are better adapted to give the impression of great power than similar tones of a softer quality. They are consequently little suitable for artistic music when used alone, but produce great effect in an orchestra. Why high dissonant upper partials should make a musical tone more penetrating will appear herafter."

1.2.4.4 Adequacy of the Classification Scheme.

We observe from these rules that Helmholtz's psychophysical approach to the specific derived cause of the sense data, namely the physical correlates of musical timbres in the form of *characteristics* in the magnitude spectra, is *roughly* the following: (1) to start with known steady-state musical timbres, (2) to discover *characteristics* in their magnitude spectra, (3) for a given characteristic, to group together timbres which are perceptually similar, and (4) to find a label (like sweet, rich, hollow, etc.) which describes the perceptual similarity.

This approach is obviously of a rather preliminary nature and thus incomplete. One of the questions is whether one can, from a modern viewpoint, improve one's knowledge enough about the relationship among steady-state musical timbres within the context of Helmholtz's classification approach to deduce a complete description of the timbres in the sense of being able to recover the magnitude spectrum from such a description. Also, Helmholtz claims that all musical timbres are periodic waveforms. Can we then deduce the converse statement that all periodic waveforms are "musical" timbres? To answer these questions, we proceed with the following discussion.

1.2.4.4.1 Recovery of Physical Data.

From a modern viewpoint, one of the most important questions to face a student of timbre is how to synthesize timbres. In other words, given a set of perceptual elements specifying a timbre, how can one recover the acoustic stimulus? Thus if we

attempt to develop Helmholtz's rules into a recovery system for timbres which are characterized by their Fourier magnitude spectra, then we believe that the following issues must be addressed.

1.2.4.4.1.1 Universality of Class Labels.

First, we must make sure that our classes are well-defined, or universal in the sense that experiments can be conducted meaningfully to lead to a consensus of assignments of data. That is, a sound which listener A describes as "sweet" and "hollow" should be described by B in the same way with high probability. One of the problems with terms like "sweet" or "rich" is that we don't know whether the identifications are completely dissociated from feelings and emotions. Different people of course have different feelings and emotions unrelated to the immediate output of the auditory processor. It is this issue that is the point of departure of Stumpf's later studies on timbres [Stumpf, 1926].

1.2.4.4.1.2 Completeness of the Class Structure.

Second, we must make sure we have enough classes to specify a timbre completely: if one can hear the effects of twenty partials, then based on dimensionality considerations we expect to need at least twenty classes (not including intersections of classes) or characteristics. As they stand, Helmholtz's classes, when viewed as dimensions, provide a size 2^n combinatorial lattice where n is the number of classes (the dimension) he introduces, which is about seven or eight.

1.2.4.4.1.3 Scaling within a Class.

Third, we must be able to tell how sounds within a given class are related: which are more "hollow" or "sweet"? The need should be obvious since there are many more timbres than the small number of classes (and intersections of classes) that we are talking about. As they stand, Helmholtz's rules do not tell us how they are related.

1.2.4.4.1.4 The Relationships between a Descriptive Element of a Timbre and its Physical Correlate.

Fourth, based on what we have discussed so far, a descriptive element of a timbre might be .3 sweetness or .9 hollowness. And we need to be able to articulate the relationship between a descriptive element like this and an aspect of the Fourier magnitude spectrum. That is, given that A is .9 hollow, we should be able to say that the even numbered harmonics have energy less than say .05 of the odd

harmonics and the energy above say the ninth partial is less than .1 of the total energy.

1.2.4.4.1.5 Finding Solutions from Simultaneous Constraints on the Magnitude Spectrum.

Finally, we need to be able to invert the relations described above, namely, given twenty such numerical constraints on the magnitude spectrum, we need to be able to reconstruct the magnitudes of the twenty partials.

1.2.4.4.1.6 Summary of this Approach.

In summary, we observe that a synthesis technique based on this approach is quite formidable, even if the dimensionality is eight or nine (i.e., even if we can only hear the effects of eight or nine partials), not to mention the difficulty in obtaining well-defined and stable psychophysical measurements. So, as a result, we want to explore an alternative approach starting from the opposite direction, i.e., starting with the magnitude spectrum.

1.2.4.4.2 Combinatorial Approach by Quantizing the Magnitude Spectrum Quadrant.

An alternative approach to the above would be to partition the Magnitude Spectrum Quadrant into a fixed number of bins. For example, we might want to have the frequency axis divided into nine parts, one for each of the lowest eight harmonics and one for all the harmonics remaining above the eighth (and below the cutoff frequency). And we might want to have four levels of intensity in the magnitude direction (e.g., 0, .25, .5, and .75). This gives $4^9 - 2^6 + 1 = 131,009$ possibilities. (The counting argument here is that if we want n–here $n = 9$–choices of the numbers 0, .25, .5, and .75, ordered, then there would be 4^n possibilities. But the silence state of all zeros should not be counted, nor should any state consisting of all zeros and .5's, or all zeros and .75's (since it will sound the same, except for loudness, as a state consisting of all zeros and .25's). This works out to $4^n - 1 - 2(2^n - 1) = 4^n - 2^{n+1} + 1$.) Of course one can start with a smaller partition, say one with two levels in the magnitude and five levels in the frequency, which would give fifteen possibilities, and approach the problem hierarchically. One would then synthesize these combinations of partials. The advantage of this is that if we go far enough, we exhaust all perceptually distinguishable magnitude spectra. These combinations would include odd harmonic spectra, even harmonic spectra, spectra with different

formant regions, spectra with different magnitude envelopes, etc. We would then see how existing musical timbres of the steady-state variety would fit into this table, whether every possibility is a musical timbre, whether perceptually similar timbres fall into neighborhoods in magnitude-frequency plane, how smooth the transitions are, etc.

The point of this is of course to outline how we might go from magnitude spectra to labels such as sweet and hollow, instead of going from the labels to the spectra. Even though starting with the magnitude spectra would appear to be much simpler, we must remark that the number of tests required would be astronomical, and this is in fact one reason for the approach that we take below, which is local in nature rather than global. Because even if we could somehow carry out the thousands of experiments required with the magnitude spectra, that wouldn't even begin to address the issue of what to do about transient signals, so it seems more appropriate to study the timbres near a given timbre, and how to interpolate between two timbres of different natures. More on this in chapter 2.

1.2.5 Perception of Transients.

Just about any acoustical phenomenon contains transients as Winckel later would emphasize (see 1.4.2, below). So what does Helmholtz have to say about the perception of transients? Notice that Helmholtz considers any irregular, hence aperiodic, vibration as noise. Thus observations he makes in connection with noise often have bearings on acoustic transients important to musical timbres of our time.

1.2.5.1 The Response of the Mechanical Resonator to Transients.

Regarding the mechanical resonator's response to acoustic transients, Helmholtz [Helmholtz, 1877, p. 403] writes:

> "On account of the question raised on p. 150b as to the behaviour of the basilar membrane of the ear for noises, we are interested further in the integral of an equation in which $A \sin nt$ [the forcing function] of equation (4)* is replaced by an arbitrary function of the time ψ_t. Of course, if this

---

* That is,

$$m\frac{d^2x}{dt^2} = -a^2x - b^2\frac{dx}{dt} + A\sin nt,$$

where the term on the left is of the form mass times acceleration, a is something like a spring constant, and b is a term to represent resistance (from the fluid).

function vanishes for very great positive and negative values of the time, it could be transformed, by means of Fourier's integral, into a sum (integral) of terms such as $A\sin(nt + c)$, and then for each one of these terms, the solution just found [for a sinusoidal forcing function] might be applied, and finally the sum of all these solutions might be taken. But this form of solution becomes incomprehensible, because it exhibits a continuous series of tones each of which exists from $t = -\infty$ to $t = \infty$. Hence we must proceed differently."

1.2.5.2 The Principle of Superposition and the Convolution Character of the Filter Response.

The linearity of the mechanical resonator's response implies that we can invoke the principle of superposition to obtain the response to an arbitrary excitation. The idea is that an arbitrary excitation can be viewed as a sequence of narrow pulses similar to the output of a sample-and-hold device to which a continuous function is applied. If the narrow pulses are approximated by impulses of the same energy (determined by the area under the pulse), and if the impulse response to the linear filter is known, then the filter response to an arbitrary excitation is simply the superposition of the impulse responses appropriately scaled and shifted according to the excitation function.

The result is mathematically the convolution integral of the excitation function with the filter's impulse response. This means that the instantaneous response of the filter in general is a linear combination of the past, present, and future excitation weighed according to the impulse response function. If the filter is time-invariant, the weights are always the same for the same distance into the past, and similarly for the future. If the filter is causal, which we will assume to be the case, given our current state of knowledge, then only the past and present contribute to the response. If the impulse response follows an exponential decay, then the filter response is one-step Markov, i.e., one instant in the immediate past determines the whole past.

1.2.5.3 The Impulse Response and the Eigenfunctions of the Damped Oscillator.

The second-order linear differential equation that describes the motion of the damped oscillator in the absence of any driving force, i.e., the homogenous equation,

has two characterstic or eigen solutions; these can be put in the form $e^{\lambda t}$ where λ assumes the form $\lambda_{\pm} = -\alpha \pm i\beta$, functions of the mass (inertia), the stiffness, and the dissipative parameters of the system. Since the impulse response must be the solution to the homogeneous equation for $t > 0$ (assuming the impulse is 0 for $t > 0$), and since any solution to the homogeneous equation must be a linear combination of the two eigen-solutions, the impulse response must be a linear combination of $e^{\lambda_+ t}$ and $e^{\lambda_- t}$ and must be real.

Without going through the tedious derivation involving solving simultaneous equations on two sets of boundary conditions (remember that an impulse really is an idealization of a rectangular pulse which divides the domain of interest into three consecutive segments with two boundaries), one can argue that (up to a multiplicative constant) the impulse response must have the form $e^{-\alpha t} \sin \beta t$ since it must be zero at $t = 0$, the instant the system is subject to the action of the impulse (we assume the system is causal). As a result, the general response is

$$r(t) = \int_0^t \varphi(t - \tau) e^{-\alpha \tau} \sin \beta \tau \, d\tau$$

where $\varphi(t)$ is of course the input. (Here, α and β are "generic" constants coming from solving a second degree differential equation. They are functions of the damping and restoring constants of the differential equation.) Therefore, knowing the physical properties of the mechanical resonator, such as its inertia, compliance, and damping characteristics, which can in turn be determined empirically from resonance experiments in the form of resonance frequency and bandwidth, we can derive the mechanical response to acoustic transients or any excitation provided that the model for the resonating device is adequate and the assumptions valid.

1.2.5.4 The Inadequacy of a Fourier Representation (Interpretation) for Acoustic Transients.

As we have seen from the quote given in 1.2.5.1, Helmholtz, in considering the mechanical response of the ear to an aperiodic excitation, opts for a linear filtering interpretation instead of a Fourier one. He describes the result of a Fourier approach as "incomprehensible"*! To better understand his statement, it seems worthwhile to make the following considerations.

* in translation

First, the ear is fundamentally a mechanical device, while Fourier Analysis is a (mathematical) representation. Therefore, given an excitation function $f(t)$ as physical data, we are really talking about whether the response behavior of a mechanical device is equivalent to the salient features of a mathematical representation of the physical data. Although one can specifically compare the response of the mechanical device with the result of a Fourier Analysis, and the filtering operation and the transform operation are both transformations of physical data, there are fundamental differences. Fourier Analysis is a transformation in the sense of transforming a mathematical object from one projection (or picture) to another, whereas filtering is a transformation of a mathematical object to another mathematical object in the *same* projected space. With a filter *bank*, the transformation is to a product space (whose dimension is the number of filters involved). Clearly what the ear does corresponds to the latter. The question is then whether there is a meaningful interpretation that would allow us to link the two phenomena together within the context of some sort of equivalence.

Now, if we agree that the membrane response is equivalent to that of a bank of linear filters with well-defined frequency selectivity varying smoothly over a range of the frequency continuum, then it seems reasonable to divide the frequency axis into bins and assign to them output values of fixed filters in the filter bank. However, we immediately run into difficulties. The output of a filter is a time function whereas any value a Fourier Analysis assigns to a bin on the frequency axis is a single complex number, composed of a real and an imaginary part, or given by a phase and magnitude at that frequency. Of course, if the output of a filter is a steady sine tone, then apart from a scale change and a phase delay, a pair of numbers suffices to specify the output time function. If we have invariant data on scale change and phase delay, then it seems reasonable to establish a Fourier interpretation of the filter bank output. And this is precisely what Helmholtz has shown in explaining the mechanical basis of perception of steady-state tones. When we are dealing with acoustic transients or some arbitrary response function, we would like to know if there is any way to get around the problem mentioned above.

Helmholtz mentions that we could perform a Fourier transform (note that a Fourier series does not exist for non-periodic functions and while the domain of analysis is 0 to 2π for Fourier series, it is $-\infty$ to ∞ (all times) for a Fourier trans-

form) on the acoustic signal and proceed to derive the response to each sinusoidal component for each filter. Then for each filter, the responses to all sinusoidal components in the input are summed (or integrated) to yield the true filter response to the particular excitation function. The following gives some typical cases:

If the damped oscillator equation is given as

$$\ddot{r} + a\dot{r} + b^2 r = \varphi(t),$$

where a is a constant that measures damping or resistance, b is a constant that measures restoring force (like a spring constant), and $\varphi(t)$ is, say, an arbitrary square-integrable function, then we define $\Psi(\omega)$ by

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \Psi(\omega) e^{i\omega t}\, d\omega.$$

We also define $R(\omega)$ by

$$r(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} R(\omega) e^{i\omega t}\, d\omega.$$

We have

$$-\omega^2 R(\omega) + ai\omega R(\omega) + b^2 R(\omega) = \Psi(\omega)$$

or

$$R(\omega) = H(\omega)\Psi(\omega)$$

where

$$H(\omega) = \frac{1}{(b^2 - \omega^2) + ia\omega}$$

would be called a transfer function by electrical engineers. Of course we can obtain the same result from the convolution integral. In any event, the equation means that if the input is $e^{i\omega t}$, then the filter response is the same $e^{i\omega t}$ multipled by a complex scalar $R(\omega)$ whose amplitude (modulus) is $|H(\omega)|$ and whose phase (argument) is $\Theta_H(\omega)$. In other words, $e^{i\omega t}$ is an eigenfunction of the damped oscillator equation. Note, however, that neither $\cos\omega t$ nor $\sin\omega t$ are eigenfunctions of the damped oscillator equation, although the amplitudes and phases of the responses are scaled and shifted in the same way that those of the complex exponentials are, provided that the filter impulse response is real.

This is just like the Fourier interpretation of the linear filter action for steady sinusoidal input except now the general output of any filter in the filter bank is not going to consist of just a complex exponential $e^{i\omega t}$ but an uncountable infinite set of them over the frequency continuum. Only in the special cases of a steady sine tone or a steady periodic vibration which contains a sinusoidal component of frequency within the frequency resolution of the characteristic frequency of the filter will we find the correct context for a Fourier interpretation. Even then we will have to wait for the transients to die and the eigentone to build up.

Take $h(t) = e^{-\alpha t} \sin \beta t$. We consider three cases:

Case (i). Let

$$\Psi(\omega) = \frac{1}{2}[\delta(\omega - \beta) + \delta(\omega + \beta)],$$

where δ is of course Dirac's delta function. Then

$$r(t) = \frac{1}{\alpha\beta} \sin \beta t.$$

Case (ii). Let

$$\Psi(\omega) = \frac{1}{2} \sum_{n=-N}^{N} a_n \delta(\omega - n\omega_0)$$

with

$$\beta = k\omega_0, \qquad 0 < k \leq N.$$

Then

$$r(t) = K + \frac{|a_k|}{\alpha\beta} \cos(\beta t + a_k + \phi_k) + \sum_{n=1, n\neq k}^{N} \frac{|a_n|}{\sqrt{\beta^2 - n^2\omega_0^2}} \cos(n\omega_0 t + \vartheta_n + \varphi_n),$$

where K is a constant.

Case (iii). Let $\Psi(\omega)$ be arbitrary, i.e., consider a (non-periodic) $\psi(t)$. Then

$$r(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{(\beta^2 - \omega^2) + i\alpha\omega} \Psi(\omega) e^{i\omega t} \, d\omega.$$

Then $r(t)$ is in general a *nonvanishing* time varying function for each filter. That is (1) $\psi(t)$ does not elicit selective response from the filter bank, and (2) the responses are not sinusoidal motion, as Helmholtz has noted (see quotations in 1.2.5.1 and 1.2.5.7).

1.2.5.5 The Notion of a *Steady-State* Response to a *Periodic* Motion

In 1.2.5.4, we showed that it is in general impossible to observe a sinusoidal response for each fibre in the basilar membrane if we subscribe to the mechanical resonator model of Helmholtz. We see that sinusoidal responses can occur only under very special circumstances, namely, ***periodic motions***. Even then, the steady-state motion must *take time* to build up and the transients die down. This time-dependent behavior does not show in the responses we have derived because we made the simplifying assumption that the input is a δ-function or a linear combination of δ-functions in the frequency domain, which means that a sinusoidal input or a linear superposition of them has been present since the infinite past. Helmholtz [Helmholtz, 1877, p.404-405] shows in mathematical detail how a mechanical resonator takes time to vibrate "sympathetically" (more and more vigorously towards a steady-state amplitude) to a pressure function which fluctuates with the characteristic frequency of the resonator. Physically, it has to do with the fact that the response is not instantaneous, i.e., the instantaneous response is a sum of present and past. The "memory" invokes the past to add constructively (a term from the theory of wave interference) if the ripples of the resonator's characteristic response to an impulse are "in phase" with the applied stimulus. From an information-theoretic viewpoint, it is the mechanical fibre's way to reject information not consistent (or commensurate) with what it is supposed to have. This view is in fact consonant with the idea that the convolution integral can be rearranged to look like a statistical correlation function. (And it is this correlation property that is the basis for extraction of harmonics in a Fourier series in Beauchamps' analysis of complex tones and Moorer's hetrodyne filter. This correlation property is also used in the maximum likelihood detection algorithm that we use to find local pitch in this thesis.)

But then, of course, the involvement of the past depends on the decay time constant of the eigenfunction of the resonator. If it is large, then the response involves a long history, therefore the steady-state response would be strong but it would take a long time to achieve it. At the same time, the "smearing" of the input would be extensive. Conversely, if the steady-state response is observed to be weak, then the decay time constant must be small and the response would follow the input very closely, i.e., there would be a short rise-time and little "smearing."

This latter situation is realized in the basilar membrane fibres of high characteristic frequencies. Finally, we can therefore see why acoustic transients are not resolved into sinusoidal motions, especially on the high frequency end.

1.2.5.6 The Basilar Membrane as a "Short-Time Fourier Analyzer."

Let us rewrite the convolution of 1.2.5.3 as

$$r_{\alpha,\beta}(t) = \int_{-\infty}^{t} \varphi_s e^{-\alpha(t-s)} \sin\beta(t-s)\, ds, = \Im\tilde{r}_{\alpha,\beta}(t)$$

where

$$\tilde{r}_{\alpha,\beta}(t) = e^{i\beta t} \int_{-\infty}^{t} \varphi_s e^{-\alpha(t-s)} e^{i\beta s}\, ds$$

and α and β are constants, functions of the physical parameters (damping and restoring) of a fiber. If

$$h_\alpha(\xi) = \begin{cases} e^{-\alpha\xi} & \text{if } \xi > 0 \\ 0 & \text{otherwise,} \end{cases}$$

then

$$\tilde{r}_{\alpha,\beta}(t) = e^{i\beta t} \int_{-\infty}^{\infty} \varphi_s h_\alpha(t-s) e^{-i\beta s}\, ds = e^{i\beta t}\Psi(t,\beta).$$

Now, a mechanical resonator of eigenfunctions $e^{-\alpha t \pm i\beta t}$ has a response which can be formally interpreted as the imaginary part of a time dependent Fourier integral $\Psi(t,\beta)$ with a time-dependent phase shift of $e^{i\beta t}$. Here, $\Psi(t,\beta)$ is the short-time Fourier transform defined by J. Flanagan [Flanagan, 1972, p. 142]. We write

$$r_{\alpha,\beta}(t) = A(t,\beta)\sin(\beta t + \vartheta(t,\beta)),$$

as usual, α and β being constants depending on the damping and restoring forces of a fiber. This form can be seen as the basis for the phase-vocoder analysis of sound by Flanagan and Golden [Flanagan and Golden, 1966]. So indeed instead of the classical Fourier interpretation, the instantaneous membrane response can be seen *formally* as a Fourier transform which varies as a function of time. However, because of its time-dependent nature, the response of each fibre in general does not resemble a sinusoidal motion.

Rather, it is a time-function approximated by little segments from a period of a sine function. Each segment can be of a different scale or taken from a different place in the period. It is of course fundamentally no different from approximating a curve by linear segments or cubic splines. And in general, it is common knowledge among

workers familiar with short-time Fourier analysis that the phase angle $\vartheta(t,\beta)$ is extremely unstable numerically in the transient segment of a sound. Therefore it is usually ignored. But while the phase angle may not be crucial in cases where selected basilar membrane fibers vibrate strongly according to their own sinusoidal motion, it is clearly indispensable when we use them to reconstruct our arbitrary function from the filter response. Indeed, work by Serra [Serra, 1986] on the marimba shows that a marimba tone reconstructed with only $A(\beta,t)$ loses some of the bright, "noisy" character in its attack.

1.2.5.7 The Collective Nature of the Perception of Acoustic Transients.

If there is a mechanical basis for the perception of timbre as Helmholtz suggests, and if the long-time response to periodic vibration brings out the individual character of certain components in the excitation, then is it not reasonable to expect that the instantaneous response to irregular vibration, such as impulses and transients associated with the growth and decay of the acoustic excitation, should bring out the collective behavior of the sound as a whole? On this, Helmholtz says [Helmholtz, 1877, p. 150]:

> "So far the theory which has been advanced refers in the first place only to the lasting sensation produced by regular and continued periodic oscillations. But as regards *the perception of irregular* motions of the air, that is, of noises, it is clear that an elastic apparatus for executing vibrations could not remain at absolute rest in the presence of any force acting upon it for a time, and even a momentary motion or one recurring at irregular intervals would suffice, if only powerful enough, to set it in motion. The peculiar advantage of resonance over proper tone depends precisely on the fact that disproportionately weak individual impulses, provided that they succeed each other in correct rhythm, are capable of producing comparatively considerable motions. On the other hand, momentary but strong impulses, as for example those which result from an electric spark, will set every part of the basilar membrane into an almost equally powerful initial motion, after which each part would die off in its own proper vibration period. By that means there might arise a simultaneous excitement of the whole of the nerves in the cochlea, which although not equally powerful would yet be proportionately gradated, and hence could not have

the character of a determinate pitch. Even a weak impression on so many nerve fibres will produce a clearer impression than any single impression in itself. We know at least that small differences of brightness are more readily perceived on large than on small parts of the circle of vision, and little differences of temperature can be better perceived by plunging the whole arm, than by merely dipping a finger, into the warm water."

In fact we shall see that recent experiments in general point to the outlook that we perceive acoustic transients in collective terms.

1.2.6 Insufficiency of the Analytic Approach to Timbre Perception.

1.2.6.1 Helmholtz's Approach is Essentially Analytic.

In dealing with the steady-state response to periodic acoustic stimuli, Helmholtz shows that timbre perception is essentially analytic in nature, i.e., one which analyzes the stimulus time function into a multitude of sinusoidal motions, similar to that of a Fourier series. Helmholtz does not concern himself with how these motions might eventually fuse into a single auditory image. Although Grey [Grey, 1975] suggests that Helmholtz invokes some kind of unconscious inference doctrine in the analysis stage as well as the synthesis stage, there is no evidence that Helmholtz actually has done so for the latter.

He seems to believe that the harmonics actually exist simultaneously and independently since one can be trained to pick them out individually. The gestalt sensation describable by simple labels such as "sweet", "hollow", or "rich" does not seem to contradict this notion. For example, the connotation of being rich certainly implies the simultaneity of many separate elements. The spatial connotation of being hollow points to the same outlook. It is harder to place "sweet" or "pleasant" in this category, but then they are precisely those problematic descriptives that might be impossible to separate from feelings and emotions as discussed above.

1.2.6.2 Beyond the Steady-State Response to Periodic Stimulus.

In 1.2.5, we have seen that in response to acoustic transients, the bank of resonators responds broadly instead of selectively. In other words, the response is collective instead of on a fiber-by-fiber individual basis. The similar response behavior of the resonators and their gradually changing physical characteristics imply that their impulse response functions are smoothly interpolated between those of the high frequencies and those of the low frequencies. And the short response time

of the high-frequency resonators implies that the general response closely follows the excitation function unless the latter is periodic with frequency commensurate with that of the fibers and sufficient time has elapsed to permit a build-up of the sinusoidal motion. Therefore, whatever the filter responses look like in detail, (1) they are all in general sensitive to the detailed shape of the single excitation function, and (2) the elements that organize the input must similarly organize the output vector. Indeed, studies by Charbonneau (section 1.4.5), Schaeffer (section 1.6) directly or indirectly support this view.

1.2.6.3 Synthesis as a General Element of Timbral Perception.

Starting from Helmholtz's analytic viewpoint, one might ask how, in general, does the ear synthesize the multitude of filter outputs (from the filter bank analysis) into a single "fused" auditory image? How does it organize them? What are the possible organizing elements? Does Helmholtz's theory suggest any hints?

1.2.6.3.1 Spectral Energy Distribution in the Acoustic Stimulus.

It is well accepted that formant distributions and spectral envelopes, when they occur, are important organizing elements for the perception of timbre. In other words, it is not the independent existence of the harmonics, rather the relationship among them, especially one which carries simultaneously continuous as well as contrasting features that are perceptually important. Although Helmholtz's analytic model provides the basis for the spectral viewpoint, it cannot be over emphasized that collective behavior in the form of relationship or simply *organization* provides a crucial element to the fusion of an auditory image.

1.2.6.3.2 Seebeck—a Proponent of the Synthesis View.

Seebeck (see [Helmholtz, 1877]), one of Helmholtz's contemporaries, whose view on tone perception Helmholtz opposes, notices that when higher harmonics fail to resolve, they tend to strengthen the impression of the fundamental more than what is implied by its Fourier contribution. He observes that the higher harmonics seem to "fuse" with the fundamental. This observation is of course related to the famous phenomenon known as the missing fundamental through the great work of Schouten [Schouten, 1940] and others. It is now generally accepted that the cause of the missing fundamental phenomenon is best explained as a temporal phenomenon, thus vindicating Seebeck at the expense of Helmholtz's narrow interpretation of the filter bank as a collection of independent elements. Note that Helmholtz does

abandon this interpretation when he discusses the membrane response to noise, or acoustic transients (see 1.2.5.7). But implicit in Seebeck's temporal interpretation is the idea that the filter-bank in the ear is collectively excited and their reponses to the linear superposition of the higher harmonics as an amplitude-modulated waveform of the input resemble that of the fundamental by virtue of extrapolation of the motion across the filter bank (the membrane). Again one can invoke the "united we shine" analogy Helmholtz uses when he attempts to explain why we should perceive the individually tiny but collectively preponderant response that one has for irregular vibration forms.

1.2.6.3.3 Global Temporal Features as Organizing Elements.

Since Seebeck, many psychoacousticians, notably Schouten, have come to recognise the importance of global temporal features for timbral perception. In contrast to the definition offered by the American Standards Association, Schouten suggests [Erickson, 1975, p. 5] that global amplitude envelope, pitch glide, vibrato, tremolo, etc., are important timbral features. This view is in fact supported by prominent contemporary musicians and researchers, notably Erickson of the U.S.A. (see 1.6) and Schaeffer of France (see 1.7). McAdams [McAdams, 1984], supposing that the stimulus arrives as superposition of harmonic partials and analysed by the ear as such, raises the important question of how in the end do the "concurrent elements in the acoustic stimulus appear as a fused auditory image." The salient features of his findings suggest that temporal features in the acoustic stimulus such as slowly changing frequency modulation (in the sub- audio range) and a sense of coherent motions (or response) among different fibres as evidenced by the logarithmic dependence of the modulation depth as a function of center frequency, provide oganization cues for the synthesis of the "fused" auditory image. It is not hard to see that a coherent amplitude variation as well as a coherent shift in the "place" response pattern are at work. And they are in response to the *form* of the stimulus. Similarly, pitch glide and vibrato, and global amplitude envelope and tremolo are respectively global variation in amplitude and in frequency. Therefore the form or the global temporal organizing elements seem logical to provide neccessary cues for the perception of timbres, especially where the stimulus is nonstationary, which is most ubiquitous. (And it is equally logical that global temporal featues may not seem significant in perception of stationary signals.) The notion of form in the

context of timbral perception is used by Schaeffer's school and is apparently borrowed from music analysis. The question is whether the dramatic change in the time-scale from the macroscopic form of a piece of music to the microscopic form of a sound event still invokes the same kind of feature-extraction strategy from the ear. The answer seems to be positive provided that we interpret the borrowed terms with care. Although short-time Fourier-transform type sound analysis techniques have become fashionable with the advent of digital computer technology (see 1.6), and the importance of these global timbral features seems to be ushered into the background either on the assumption that Fourier Synthesis would automatically reproduce these features or that they are derived features secondary to the analytic elements, Charbonneau's work and recent experience have pointed again to these global features as organizing elements even within the Short-Time Fourier-Transform context. More recently, Serra has shown the inadequacy of this latter approach for highly percussive instrument tones such as those of the hard-struck marimba. And finally, we will show that we can carry this notion, viz., the global amplitude envelope as a central synthesis element, to the extreme and produce synthesis that duplicates the orginal with imperceptible difference on the marimba-type tones (see 3.4 for a discussion on the choice).

1.3 The Psychoacoustic Calculus of Stumpf.

The foundation for a physical interpretation of the psychological phenomenon of tone perception laid down by Helmholtz spurs interest in discovering the physical correlates of perceptual attributes related to the timbres of sounds. The spectra of stationary sounds reveal a great deal about what we actually hear in the steady-state. The notions of spectral envelopes and formant regions bring fruit to understanding the perception of speech vowels and of timbres of musical instruments. This understanding in turn brings about successful construction of versatile electronic organs later, as well as the making of traditional acoustic instruments of improved quality.

But in order to know how we can actually make an impact on any such instrument modification task, we really need to know the relationship between perceptual changes and the corresponding changes in the physical correlates. What Helmholtz does by throwing timbre into bins, i.e., classifying them, cannot accomplish this task. Here we have a psychoacoustic calculus at hand. We must ask how much change in the physical parameters is necessary to produce a perceptible change in the timbre, or a certain timbral dimension. This question is similar to asking how much voltage must one apply to the control port of a transistor, e.g., the base-emitter junction, in order to produce an appreciable current across the output (collector) resistor. And then we might ask how much change at each input voltage must be supplied to produce a fixed amount of current change at the corresponding output current value. The result is a response curve with the input voltage as a control variable.

Noticing the deficiency in Helmholtz's study, Stumpf [Stumpf, 1926] approaches the subject of timbre perception essentially along the lines discussed in 1.2.4.4.2, except restricting his studies to the class of vowels and using a coarse quantization. Items of interest are:

(1) He rejects the usage of labels.

(2) In one extensive study, he considers the set of vowels whose coarse perceptual structures are essentially determined by the lowest two formants.

(3) He then scales the timbral variation by these two physical parameters, i.e., he varies the latter systematically to obtain a (double) series of just

noticeable shifts in timbre (by listening).

(4) He maps the result of his scaling experiments into cells in the plane of the two parameters. Thus the values of these two physical parameters bounded by the edges of each cell describe perceptually identical timbres.

Stumpf's work represents one of the most important advances in our attitude toward psychophysical studies. Although he was not the first to derive the mappings between the relations of perceptual data and the relations of physical data (e.g., Fechner and Weber's psychoacoustic calculus on one dimensional sense data are well known, see 1.4 and [Jeans, 1937]), the multidimensional nature of the mappings we discussed above influences the work of many, including Peterson and Barney [Peterson and Barney, 1952]. And the local nature of the mappings provides details of the psychoacoustic relations among sense data and between them and the physical data. This approach contrasts with the global approach of multidimensional scaling of Kruskal [Kruskal, 1964], Shepard [Shepard, 1966, 1972], and others, as it is used by Plomp (in a non-metric sense) [Plomp, 1970], Wessel [Wessel, 1979], Grey [Grey, 1975], and others. The latter approach does *not* lend insight into such details and hence does not provide a means to regenerate timbres or make new ones. Stumpf's specific approach to the study of perception of vowel timbres by formant mappings (as described above) serves as a model for modern synthesis of timbres. The Fant speech synthesizer [Fant, 1960] is of course one of the earliest examples modelled on his approach.

1.4 The Communication Model of Timbre Perception of Winckel.

In the dialectic evolution of psychoacoustics, while Stumpf moves away from Helmholtz's methodology of psychological measurements, his view on the physical causality of timbre perception remains consonant with that of Helmholtz. His decision as well as that of Helmholtz to focus on spectral dependence of timbre is probably a wise one in view of the technology available in their times. However, it has become clear to many that the spectral description of timbre is less than complete.

1.4.1 Winckel's Precursors.

1.4.1.1 Backhaus.

It is obvious that sounds grow and die. So an important question is how they grow and die, e.g., how fast? To this end, Backhaus [Backhaus, 1932] investigated the physical characteristics of the onsets of many important musical instruments as well as of speech sounds using the Fourier approach.

He spoke of transients in terms of partials in the language of Fourier analysis and he found that the onset as well as the decay process varies from one instrument to another and in particular the relationships among the onsets and decays of the partials vary from instrument to instrument. And he concludes that the dynamical relationships among the growth and decay of the partials, especially during the growth phase, are responsible for the perception of the sound's timbre. And this explains why certain musical instrument timbres "sound more clearly defined" than others, apparently using the form idea in musical analysis. The non-uniform growth behavior among partials can be understood in terms of the response characteristics of the eigenmodes of the instrument as we have discussed within the context of the ear as a mechanical resonator, and is later described as a kind of non-linear growth behavior in the evolution of the waveform (Beauchamp [Beauchamp, 1975], Lo [Lo, 1986]).

1.4.1.2 Meyer and Buchmann.

Meyer and Buchmann [Meyer and Buchmann, 1931] made exhaustive studies of the spectral content of the sounds of many Western musical instruments. They discovered that timbres of sound coming from rather different acoustic cavities appear similar largely because of the similarity in the attack—the way the production

of the sound is articulated initially. For example, the plosive character of the speech sound "dah" is very similar to the timbre of the attack of a trumpet.

Secondly, they observed that even with very detailed knowledge of the spectral content of the stationary parts of the sound, additive synthesis, i.e., synthesis by superposition of sine waves, based on the spectral information alone does *not* give timbres very close to the original.

1.4.2 A Modern Interpretation of Winckel's System Model.

Winckel's work in the early 1950's represents one of the early attempts to present a coherent argument for the importance of onset transients in the perception of timbre missing from Helmholtz's treatment. In summarizing his view, he writes [Winckel, 1967, p. 34] "In musical sounds the characteristic overtone spectrum (formant) and the onset and decay transients are of equal importance. This is unfortunately overlooked in recent works on musical aesthetics, which again and again deal only with the stationary part of a sound through its overtone structure." Winckel's argument consists essentially of the following:

1.4.2.1 The Communication Model.

The sound that enters the ear has a context, namely, where it comes from. From the information-theoretic viewpoint, the sound is first produced by a source, and then transformed as it is transmitted before it is received by the ear. In fact, as sound passes through the ear, it is further transformed in a significant way even on a mechanical level. Each part of the chain of communication, from excitation, through transmission, to reception, involves some kind of mechanical action. Behaving in this fashion, each component in the chain of communication is subject to the laws of physics.

1.4.2.2 Source Characteristics.

Source characteristics deal with the properties of signals that may be independent of the ear. These properties can be observed objectively from the acoustic waveforms themselves or deduced from the nature of the sources that produce them. In the next few sections, we will discuss salient characteristics in the spirit of Winckel's writings. For another in depth and very interesting treatment of sound perception from the viewpoint of source/receiver relationships, see [Huggins, 1952].

1.4.2.2.1 Real Acoustic Signals are not Periodic.

In the real world, there are many *apparently* periodic entities. Electromagnetic

waves, the spatial structure of many solid lattices and their vibrational structures, the Foucault pendulum, the planetary orbits, and the seasons are just a few. But they are only apparently periodic, and the degree of conformity varies from example to example. The real measure is the window of observation. If one waits long enough, we will discover that nothing is exactly periodic.

Real acoustic signals are not periodic either. We will examine this in the following sections. But the key point to keep in mind is the degree of stationarity the ear "sees" with respect to its observation window(s). These windows are typically on the order of milli- or even microseconds long—in great contrast to the window of observation for the periodicity of the earth's orbit, for example. Winckel's argument relies more on a philosophical ground than on the numerical scale between the variation in the acoustic signal and the ear's observation windows (which we will call ***auditory windows*** in chapter II, and which we will there elaborate upon in detail). Nevertheless, it is important to first see his argument, keeping in mind the ear's role.

1.4.2.2.1.1 A Pure Sine Wave is an Idealization.

Winckel argues that first of all the mathematical object of a sine wave is not logically realizable because someone would have to turn it on a long time ago before one could observe it or listen to it. In other words, a pure sine wave is an idealization.

But one might argue that in practice one can realize a "sine wave," i.e., something like a sine wave, in the sense that within a given window of observation, we can observe a waveform which is practically not any different from one observed through the same window of a "sine wave" that has been turned on much earlier. So why the fuss?

1.4.2.2.1.2 Global Window of Observation.

As long as the "sine wave" is turned on far ahead of the *global* window of observation so that the transients have died down, one would indeed *see* a window of the "sine wave" in the sense described in 1.4.2.2.1.1. The lead time is lower-bounded by the time-frequency uncertainty product relation. That is, if the energy surrounding the frequency of the sine wave is small enough so that the variance in the frequency about the sine frequency with respect to the frequency distribution (as a result of the onset transients) is $(1\text{Hz})^2$, and if this condition corresponds to what we call practically no difference between the windowed "sine wave" and

its asymptotic analog, then roughly one second of lead time should be sufficient provided that we don't turn up the sine-wave generator so fast as to cause the device components to oscillate wildly.

1.4.2.2.1.3 An Imperfect Sine Wave has a Practical Consequence.

The concept of a window of observation is indeed the key to Winckel's discussion. The point is that in any real application where music is concerned, an instrument or a loudspeaker is continually excited in the course of the music. The process of turning on and off an acoustic event is continually exhibited before the listener and cannot be hidden outside the window of the listener's listening activities (as opposed to the auditory windows of the ear). In fact, the window of observation contains the entire piece of music. Therefore our inability to avoid transients has a practical consequence in musical application.

1.4.2.2.1.4 Perfectly Periodic Acoustic Signals do not Exist.

We have seen that within a normal window of observation or listening where musical application is concerned, a sine wave does not exist. Using exactly the same argument, we can conclude that a periodic acoustic signal, continuous or discontinuous, does not exist. In other words, we have to turn on and off these signals "real-time," i.e., within the window of musical listening. Still put in another way, we cannot leave out the on/off transients of the acoustic events in a piece of music. In spectral terms, we cannot expect line spectra in real world sound: Each line is broadened by the sound's finite life-span, i.e., by the abruptness in which sound comes and goes. Of course, different acoustical sources project different acoustic transients to the sounds they generate. The latter are characteristic of the sources. We will have more to say about this later.

1.4.2.2.2 The Impossibility of Changing Anything Instantaneously.

1.4.2.2.2.1 Inertia of the Source.

Winckel argues that neither can we change a signal from one shape to another instantaneously because we must change the *state* of the source to do so. But the source, being a physical device, has inertia; therefore it takes time to change a signal from one shape to another. In particular, it takes time to turn on and off a signal.

1.4.2.2.2.2 The Inertia in the Impulse Response.

As we have seen in 1.2.5.3, a damped oscillator modelled as a second order

linear differential equation with constant coefficients

$$\ddot{f} + a\dot{f} + bf = 0$$

has a pair of complex eigenvalues $\lambda_{\pm} = -\alpha \pm i\beta$ where both α and β are functions of a and b, a being the damping constant and b being the restoring force constant. The corresponding eigenfunctions are the complex exponentials $e^{\lambda \pm t}$ and the impulse response is $e^{-\alpha t} \sin \beta t$. This says that even if the excitation is an impulse, the vibration form of a damped oscillator will grow initially as $\sin \beta t$ and decay later as $e^{-\alpha t}$ (as it rings away at the angular frequency of β). Now for a general form of excitation which can be considered as a sequence of impulses appropriately scaled individually, the vibration form of a damped oscillator is determined by the convolution integral between the impulse response and the excitation function. Regardless of the shape of the impulse response, (1) the integral is always smoother than the excitation function since it is being weight-averaged over time, and (2) the integral as a function of time must be initially small (as long as the impulse response is a causal function) because initially the past is a string of zeros or silence. Now it would be easy to see how the same argument can apply to a three-dimensional acoustic source in the real world. There the vibrating system can still be modelled as a system of coupled linear damped oscillators. And an arbitrary vibration can be seen as a linear combination of some uncoupled vibration modes, or eigen-modes. These eigen-modes in some properly chosen coordinates obey exactly the same equations as the simple damped oscillator does. Each mode is complex exponential. In other words, if we have a system of N degrees of freedom, we will have N complex eigenvalue pairs. Although a general solution to the system of equations in absence of force is a linear combination of the $2N$ eigenfunctions, the real solution always appears as a linear superposition of the forms $e^{-\alpha_k t} \sin \beta_k t$. This form crudely approximates the partials of Backhaus and others. Thus we can see that the inertia in the acoustic source is reflected in the attack transients of the waveform the source produces.

1.4.2.2.3 The Interpolated Nature of Source Response to Changes.

Thus we see that the onset of the acoustic response to a source excitation is somewhat between that of an abrupt change and that of a steady-state. And different sources have different forms of growth and decay. In Fourier terms, the

growth and decay of the partials are described by the complex eigenvalues of the vibrating body.

Although Winckel does not mention it, the inability of the source to change abruptly means that the acoustic waveform is constrained somewhat in the way it changes from one vibration form to another. In other words, the waveform will take time to evolve as the source gradually responds to changes. This source characteristic will be seen as providing a basis for the timbre model we are going to develop.

1.4.2.2.4 The Limitation of Helmholtz's Model of Timbre from a Source Viewpoint.

The implication of Winckel's discussion is that as far as Helmholtz's approach is concerned,

(a) The latter leaves out an important part of the source characteristics, namely, the onset characteristics, and, to a lesser extent, the decay characteristics. The decay characteristics are usually closer to those of the steady-state for a variety of traditional Western instrument tones.

(b) The line spectrum representation of the musical timbre neglects the anharmonic energy that reflects the non-periodic character of the acoustic signal and hence loses the natural flavor of the timbre.

In general, if we have a waveform $f(t)$ over an interval $[0,T]$ and if we repeat it once (over the interval $[T,2T]$), then if the Fourier transform of $f(t)$ is $F(\omega)$, then the new functions are $g(t) = f(t)+f(t+T)$ and $G(\omega) = (1+e^{i\omega t})F(\omega)$. Notice that whatever the nature of $F(\omega)$ is, $G(\omega)$ has an additional two zeros in every interval of length $\Delta\omega = 2\pi/T$. If $f(t)$ is repeated N times over $[T,(N+1)T]$, then

$$\begin{aligned} G(\omega) &= (1 + e^{i\omega T} + e^{i2\omega T} + \ldots + e^{iN\omega T})F(\omega) \\ &= \frac{1 - e^{i(N+1)\omega T}}{1 - e^{i\omega T}} F(\omega) \\ &= e^{\frac{iN\omega T}{2}} \frac{\sin\frac{(N+1)\omega T}{2}}{\sin\frac{\omega T}{2}} F(\omega). \end{aligned}$$

Now $G(\omega)$ has $N+1$ (additional) zeros in each interval of length $\Delta\omega = 2\pi/T$ and as $N \to \infty$, the Fourier transform of the one-sided periodic extension of $f(t)$ picks up countably infinitely* many zeros between the "harmonics" of the spectrum. It

* That is, the number of zeros is infinite, but they can be put into one to one

can be shown that completing the periodic extension of $f(t)$ from the left actually removes all the mass between the "harmonics" and leads to a line spectrum modulated by the envelope of $|F(\omega)|$, as expected.

So we see that periodically extending a finitely supported signal gives us a more coherent signal in the sense that as the repetition grows, the frequency distribution becomes more and more concentrated around the frequencies where the harmonics of a purely periodic signal should be. To see this in discrete mathematics, let us consider a sample sequence sampled at the rate of f_s samples per second. Let the sample sequence consist of N_T samples over a real time of T seconds. Then the discrete Fourier transform of the sample sequence has exactly N_T points corresponding to the frequencies $\frac{kf_s}{N_T}$ where $0 \leq k \leq N_T - 1$. Now if we repeat the N_T samples exactly, then the DFT has $2N_T$ points corresponding the the frequencies $\frac{kf_s}{2N_T}$ but the points with odd index have no energy. However if the second N_T points differ from the first by a sequence we call $\{\varepsilon(k)\}$, and if $\sup\{|\varepsilon(k)|\}$ is small compared with the original signal, then the points of the DFT with odd index, although non-zero, will be small. Similar considerations can be made if there are m repetitions instead of 1, and we get a frequency resolution of

$$\Delta\omega = \frac{f_s}{mN_T} = \frac{1}{mT},$$

where f_s is the sampling frequency, N_T is the number of samples, and T is the length of the time interval. We can then say, the longer we extend the signal by increasing the integer m, so that the total duration of the signal is now $\tau = mT$, the better we can resolve the frequency. That is, we can find a well-resolved neighborhood around each nominal harmonic such that the variance of the frequency with respect to the harmonic varies inversely with the total length of the repeated signal.

What we have just shown is the discrete version of the spectral behavior of a signal as we extend it either periodically or quasi-periodically. In the asymptotic situation where the extension is completed over all of time, we have a line spectrum over the real frequency continuum, as expected. But in practice, we don't have that. Instead, we have a typically continuous spectrum with concentrations of

---

correspondence with the positive integers 1, 2, 3, ... Not every infinite set is countable, of course; e.g., the real numbers do not form a countable set. See for example, [Rudin, 1955].

energy except in cases where the onset transients are strong, i.e., the onset is quick and powerful. Then the spectrum is wide-band and does not provide clearly interpretable information about the signal, other than some generalized facts in terms of band-width or high frequency concentration, etc.

1.4.2.3 Receiver Characteristics.

Receiver characteristics deal with the response properties of the ear to acoustic stimuli. These properties are complex and have been only partially deduced from experiments. Many of these properties are scattered in the literature and not well organized into a system of viewpoints. Perhaps it is still too early to formulate such a system, yet Helmholtz did, and his mechanical resonator viewpoint is still essentially valid, insofar as timbre perception is concerned, even after a century of research, including that of Békésy. One reason why an occasional attempt may be productive is the natural questions that emerge as one tries to synthesize different observations for a particular study. They may spur debate and further experiments to resolve the issues.

In the following sections, the views are generally those of the author. They are presented here because they are in the spirit of Winckel's communication model idea of sound perception within the context of timbre perception.

1.4.2.3.1 Ethological Consideration—the Evolutionary Context.

By ethological we refer to the listening condition outside of the laboratory. From the evolutionary viewpoint, the ear and its host have co-existed with various sound sources for ages. If there are characteristics in a sound that reflect the nature of its source, it would be logical to assume that the ear, as a receiver, has a "fiduciary" responsibility to the host for its well being to recognize these source characteristics and to decide in a preliminary fashion whether to warn or not. For example, the immediate response of the ear to hearing a sound might be to ask: Is it a sign of danger? In order to come to an intelligent resolution of this matter, searching the memory for source characteristics and comparing them with those of the stimulus seems a necessary and automatic part of the auditory function.

1.4.2.3.2 Mechanical Consideration—the Mechanical Response Property of the Ear.

We have discussed the principle of mechanical response characteristics of the ear in fair detail in 1.2 in connection with Helmholtz. We can see also that much

of what has been discussed in 1.4.2 in connection with source characteristics should equally apply to the ear as a mechanical element. It has inertia and obeys causality. In the linear version, which the device obeys at least approximately in the limit of moderate excitation, the principle of superposition applies. The response is the convolution of the impulse response with the excitation function and the impulse response is a linear combination of the eigen-modes of the mechanical system. Furthermore, the responses involve transfer of energy and momentum and depend on the efficiency of the transfer. It takes time for the mechanical motion to cohere in the sense of either growing into a steady-state, or periodic, motion or being destroyed by destructive interference, and there is intrinsic bandwidth to a mechanical resonator determined by its damping characteristics. The issues that have not been appropriately discussed will be developed further below.

1.4.2.3.3 Variable Threshold of Hearing and Timbre Perception.

It is a well-established fact that there is a threshold of hearing. Certainly part of it is due to the fact that sufficient energy must be supplied within a coherence time characteristic of the constraints of the device and their interaction to generate any recognizable motion other than the Brownian motion it experiences constantly on the atomic level.

Another part of it is probably due to need, for the ear must have learned that not every stimulus is necessarily important. For example, acoustic signals originated from a great distance may render the signal so weak that it must also be rather unreliable because of all kinds of interfering acoustic elements (such as the howl of the wind and the murmur of the stream). So, to be efficient, the threshold level must have been further adjusted to the need of the animal's current survival requirements. Certainly, this threshold is further periodically adapted to the animal's daily cycle of activities. Furthermore, if the ear has a fixed resource requirement like most other devices, then adjusting the threshold can optimize its dynamic range to achieve the most efficient detection of the class of signals most relevant to the survival need of the animal. This phenomenon of variable threshold will be seen to be important in the ear's ability to discriminate for and against various features in the presence of other features in timbre perception (see 2.3.3).

1.4.2.3.4 Context and Timbre Perception.

From a communication viewpoint, if it is plausible to attribute efficiency as a

cause for the existence of a threshold of hearing (which is a function of the ear's host and is adaptive), it should also be plausible to assume that the auditory detector would go to sleep at any arbitrary level of excitation unless subsequent changes are significant enough to warrant its attention. Certainly the meaning of "significant" varies from species to species and function to function. But the qualitative knowledge of these threshold changes is important to the psychoacoustic student in general and the timbre student in particular. This qualitative knowledge is needed to decide what in the complex form of the acoustic wave must be prescribed to ensure the survival of a certain perceptual feature in the acoustic event. The decision must be based upon the evaluation of the current change in amplitude level, for example, against what has been in the waveform in the form of an ℓ_∞ comparison* within a window (probably characteristic of the fibre's damping constant), i.e., sample by sample comparison or in the form of window-averaged comparison. In any case, the recent past as represented in whatever form of a window characteristic of the ear provides the notion of a context against which perceptual decisions come to play. A simplified but useful observation has been around in the form of Fechner's Law which states "the intensity of our sensation does not increase as the energy of the exciting cause, but only as rapidly as the logarithm of this energy" [Jeans, 1937, p.223-4—pages 221 through 224 discuss this law rather thoroughly].

In chapter II, especially in 2.3.3, we will show how context might be used to determine what acoustic features may form a timbral feature or part of a timbral feature and how they might affect synthesis of timbre.

1.4.2.3.5 Asymmetry of Time and Timbre Perception.

A characteristic distinction between the Fourier transform and the physical behavior of a mechanical device such as the ear's basilar membrane is that, for the former, since it is a mathematical creation, causality is irrelevant, whereas every hitherto known fact [Pippard, 1985] suggests that a mechanical device obeys the law of causality. In other words, the latter is partially a function of what happens before the present but is uninfluenced by what is to come. Naturally, a wave A given by a sine wave modulated by an asymmetric trapezoidally-shaped amplitude envelope

* An ℓ_∞ comparison is one in the supremum norm, i.e., one involving the maximum of the differences, as opposed to an ℓ_2 comparison, which would involve squaring the differences and summing.

$G(t)$ over an interval $(0, T)$ does not sound the same as a wave which is the same sine modulated by $G(T-t)$ although the Fourier transform of both are identical up to a sign change in the overall phase angle. Of course any wave (besides just a sine) will be subject to the same effect. In fact, the timbre of a sound played backwards is so dramatically different for most sounds that Schaeffer wrote [Schaeffer, 1966] that he was shocked by the perceptual asymmetry. Naturally, this asymmetry has to do with the fact that the basilar membrane is a mechanical device. Therefore its response obeys causality:

$$\int_0^t \psi_A(t-s)h(s)\,ds \neq \int_0^t \psi_A(T-t+s)h(s)\,ds = \int_0^t \psi_B(t-s)h(s)\,ds,$$

where $\psi_A(t) = G(t)\sin\omega t$, $\psi_B(t) = G(T-t)\sin\omega t$, and $G(t)$ is an asymmetric trapezoidally-shaped amplitude envelope.

1.4.2.3.6 Efficiency of Energy Transfer and Timbre Perception.

As with any mechanical element, the magnitude of a mechanical resonator's response depends on whether energy is efficiently transferred from the excitation to the resonator. In fact, the characteristic frequency of the resonator plays the moderator's role in whether to accept or reject a certain acoustic waveform. From temporal considerations, as we have discussed in 1.2.5.5, Helmholtz [Helmholtz, 1877, p. 404-405] shows that if the excitation is "in phase" with the characteristic frequency of the mechanical resonator, after some "coherency time" which determines how well the frequency is brought out compared with others, then there is a strong resonance, i.e., energy is efficiently transferred to the device.

A consequence of the discussion is the following: Suppose we press down two keys on the piano "simultaneously." Of course, the onset of the response of the two sources cannot be quite synchronous. Furthermore, there is no reason to believe that they should have identical response characteristics. Therefore, when the superposed sound arrives at the ear, one set of fibres would respond strongly to one source and another set (possibly overlapping) to the second source because of the energy transfer principle we have just described. Now of course the source characteristic of one note will be reflected in a coherent fashion across one set and that of the second note will be reflected in a coherent fashion across the second set except for the overlapping part. Even though the superposed waveform varies from one occurence to the next (if we perform the experiment more than once), the coherent features

of each source remain at each fibre where efficient energy transfer is possible. This example illustrates how it is possible for the ear to have a single "fused" timbre percept associated with a source and at the same time to resolve timbres associated with different sources when the latter are excited together. It also refutes the notion that the ability to hear the notes in a chord necessarily means that the ear is phase-insensitive between partials simply because the superposed waveform varies from instance to instance because of a lack of synchrony in the onsets of different notes [Crawford, 1980, p.67].

1.4.2.3.7 The Observer's Uncertainty Principle and Timbre Perception.

There are two levels of limitations we must contend with. On the most universal level, there is physical law that everything must obey. On a more local level, the ear itself might have its own resolution limits. How the acoustic stimulus appears to the ear is a time function of amplitudes. The amplitude resolving limit has been discussed in terms of Fechner's logarithmic law, and is reflected in our common use of the *bel* scale. The frequency resolving limit on the one hand depends on the time span of observation on the global scale and on the other hand is a function of frequency itself because of the mechanical property of the ear.

The frequency resolving limit is a direct consequence of the fact that frequency is not a concept independent of time. In fact, a clearly defined notion of frequency is associated with periodicity of the time function of amplitude variation. For example, how does one know that a periodic function is varying at 1000 Hertz instead of 999 Hertz? A most intuitive way is to "beat" it with a sine wave whose frequency is swept across the 1 kiloHertz mode from left to right, so as to discover the coincident pattern between the two functions. But in order to do so, we have to observe the interference for at least half a second before we can tell whether the waves agree in frequency to within one Hertz, and similarly half of ten seconds to tell whether they agree to within .1 Hertz. In general, we observe that we cannot assess the frequency of a periodic signal more accurately without at the same time expending more time to do so. More precisely, the inaccuracy Δf obeys

$$\Delta f \cdot \tau \geq \frac{1}{2},$$

where τ is the time length of the observation.

If we consider the limitation on time, it is at first a little difficult to understand what it means to take more frequency to ascertain the time when something takes

place or to assert that we lose more and more frequency information as we attempt more and more precisely to determine the time of an acoustic event. We are not used to dealing with the concept of frequency except where repetition is involved. But we can imagine that in order to ascertain the time a sine tone burst takes place, one sends out a series of clicks of a certain separation τ at the same time. (The series of clicks forms a sort of "yardstick" in time.) If the tone burst is long, i.e., distributed, then we must lengthen τ correspondingly in order that the tone burst continue to fall within two successive clicks. Now the frequency spread of the tone burst is

$$\Delta f \sim \frac{1}{T},$$

where T is the length of the tone burst. (We have discussed this in 1.4.2.2.) Therefore

$$\tau \Delta f \sim \frac{\tau}{T} > 1,$$

where Δf is the frequency spread, τ is the separation time between clicks, and T is the length of the tone burst. Of course, if we want to ascertain the time of the tone burst's occurrence better, assuming that the tone burst is more localized in time, then we must reduce the window time τ until $\tau \leq T$; at that point it becomes meaningless to say that the event takes place in the window of τ. In general we see that the better we are able to localize the time, we must also lose the frequency resolution and therefore the spectral cues for pitch detection of the tone burst. However even if we are willing to give up frequency resolution indefinitely, the ear does not go on indefinitely in order to ascertain the time of an acoustic event.

Winckel quotes a time constant of ~ 50 milliseconds. But it could be as small as ten milliseconds based on our fusion experiments with time-separation pitch.* In any event, fast succession of acoustic events would most probably be perceived as timbre rather than melody.

1.4.2.3.8 Collective versus Individual Response and Timbre Perception.

As we have discussed in 1.2.6.3.3, 1.2.5.7 (which also includes a passage of Helmholtz's similar opinion on this matter), and in 1.4.2.3.6, collective response

* We are here referring to certain experiments we have conducted, in which the subject listens to pulse pairs separated by various time intervals. If the time interval is sufficiently short, for example, ten milliseconds, then the pulses "fuse" into a single sound with a faint pitch which is a decreasing function of the time separation.

reflects a lack of efficient transfer of energy from the signal to a particular resonator, but at the same time a smaller but nevertheless coherent motion that reflects the source characteristics–especially those special to the transients. By contrast, the individual response reflects an efficient transfer of energy from the stimulus to a small selected number of fibres. If the onset transients can somehow be isolated, the phase relationship among the sinusoidal responses may become of secondary importance.

For *long* sustained timbres with harmonic quality, i.e., waveforms with quasi-periodic structure, the sinusoidal responses of selected fibers dominate over the collective response. As the ear focuses its attention on these selected fibers, the temporal behavior is regular and predictable. Thus, it is easy to imagine that the ear actually focuses its attention on the distribution pattern of the strengths of these oscillations along the cochlea's "place" dimension. This reasoning explains why some experienced listeners assert that they don't hear timbral difference as the phase relationships of the source harmonics change. Since the relative strength between the individual response and the collective response is affected quantitatively by the efficiency of energy transfer, it is understandable that even under the stationarity assumption certain phase configurations in the source harmonics could give rise to some degree of timbral changes. This observation dates back to Helmholtz and has been reaffirmed—see [Plomp, 1964] and [Schroeder and Mehrgardt, 1982]. In fact, in the experiments of Schroeder and Mehrgardt, the timbre of a sharply contrasted periodic waveform is compared with that of a noise-like periodic waveform with the same components but a set of phases designed to "flatten" the waveform.

On the other hand, for sounds with prominent non-stationary characteristics, such as strong attack transients, the collective response is important. Therefore, it explains why phase relationships of the source harmonics are considered important timbre synthesis parameters among modern workers in this field; see [Grey, 1975] and [Charbonneau, 1981] for example. In general, we expect that the relative importance of the collective and the individual responses provides a continuous distribution for the "phase" dependence of timbre.

1.4.2.3.9 Attacks and Decays as Natural Timbral Features.

Since attacks and decays are a manifestation of the manner in which the source is excited and of how it responds to the excitation as a mechanical device, the ear,

through years of listening, must have learned to recognize these features as some kind of signatures of the instrument.

For example, the timbres distinguish between the violin, viola, and cello of the bowed-string family and the guitar and harpsichord as a plucked-string family and the piano as a struck-string family. Furthermore, the unique physical characteristics of an instrument, as manifested in its size, shape, and material, induce a different mechanical response to a given type of excitation, thereby supplying another level of distinguishing feature among members of a given family. These distinguishing response features must surely be noticed by the ear. The manner of articulation introduces still another level of distinguishing response features for a given instrument. We can go on and on but the main point here is that the acoustic vibration resulting from exciting a physical object represents the characteristic response of the object to the excitation and this characteristic response defines the shape of growth and its eventual decay.

Thus from the communication model viewpoint, it is easy to see why the ear as the receiver should have learned to recognize the source or its signature through years and years of repeated listening. Since this signatory source-receiver relationship is manifested in the growth/decay characteristics of the sound, it would only be natural to expect that the latter constitute important timbral dimensions. In fact, many musicians use these terms to describe these dimensions collectively.

Although our experience shows that loudness is also a function of growth and decay, the functional dependence is of a different nature. Loudness has more to do with being a measure of all the energy being stored in the membrane at a particular instant of observation, as is manifested by how vigorously the fibres are vibrating, whereas the timbral dimensions have more to do with being a measure of how the acoustic waveform is evolving, or the form of the evolution. While loudness clearly reflects the fibre response sensitivity and therefore is a function of how the acoustic energy is distributed across the membrane, the timbral dimensions which are functions of the growth and decay of the waveform reflect more the temporal element that unifies the response behavior of different fibres. While both are functions of the collective response of the fibres and both are functions of time, loudness is a summation—a scalar, whereas the timbral dimensions follow a profile of patterns. From the communication viewpoint, loudness is, through evolution, a distance cue

(e.g., how close is danger?) whereas the timbral dimensions characterize the source (e.g., what type of danger?). While changes in growth and decay effected by varying the strength of source input induce changes in both loudness and timbral dimensions, the source response characteristics are usually strongly retained in the timbral dimensions. Therefore the ear as a receiver also adapts to this kind of timbral change to insure a reliable recognition of the source. If a violin is plucked more vigorously, the sound is still recognized as that of a "plucked" violin. When a lion roars more vigorously, the little animals still shudder at the roar instead of mistaking it as a loud yawn.

1.4.2.4 Transmission Characteristics.

The waves that are excited in the mechanical elements of the ear are an image of the air vibration that enters the ear. This image is perceived as the image of the vibration of the source. This image of the source is often distorted by the boundaries (if any) of the chamber the sound is produced in, and obstacles the wave must bend around or be reflected from. This type of interference can significantly change the timbre quality of the source.

First, the effect of the normal modes of the room introduces spectral modification in a stationary acoustic signal. And then, the spectrum of the signal arriving at the ear is sensitive to the latter's location relative to the walls. For example, the sound of a fountain in a courtyard surrounded by concrete walls is heard differently at different locations. The difference is more than loudness. If one is close to the wall, the timbre has an additional softer component, which gets sharper (as in pitch) as one stands closer to the wall. Similarly, consider an acoustic transient of sufficient length, say, a half-second marimba tone, in a square room of side ten feet. Noting that the speed of sound is approximately in 1000 feet per second at room temperature, the transient signal will take twenty milliseconds to return to its origin of propagation, after reflection from the front and rear walls, if, say, the marimba is in the center. The resultant waveform, being the original algebraically added to a scaled version of itself, shifted in time about twenty milliseconds, is in general quite different from the original. If there are smaller objects in the room, the vibration pattern is further altered, because fluctuations of higher frequency (i.e., corresponding to wavelengths much less than the dimensions of the objects) will be directly reflected, whereas those of lower frequency will be deflected or scattered. (Enough

of these objects might actually randomize the effect of the individual alterations.)

The "culvert whistler" is an example of the effect of room acoustics on the timbre of an acoustic transient. In this case, clapping one's hands sharply in front of a culvert produces a "zroom," which starts at a high pitch and drops to a low pitch within a fraction of a second (see [Walker, 1977]).

Singing in the shower furnishes a widely experienced example of the effect of room acoustics on the perception of timbre. It has been pointed out [Knudsen, 1963] that the exciting voice one hears isn't one's own voice. Rather, it is one's voice filtered through the shower's frequency response.

Of course, when one talks about electronic simulation of acoustic halls (see, e.g., [Borish, 1984]), one implies that the timbre is changed from one acoustic environment to another. What is interesting is the assumption that we still hear the same music (or speech) even though the "sound quality" or timbre is changed in some perceptually significant way.

The reason why this may be the case (i.e., one is hearing the same speech or music) is usually given as the ear's ability to adapt, to organize, and to recognize distinctive features, much exceeding what an oscilloscope as a passive observer is capable of doing. Under many listening conditions, the timbral quality of the acoustic event may change significantly from the source point to the receiver point, but the information content remains intact.* In traditional Western music, timbre acts as a carrier (see [Erickson, 1975]) of information. The information, or the set of interesting patterns that continue to draw a listener's attention is in the pitch structure. But what happens if timbre is used not merely as a carrier of information but also as information itself, much like speech, especially in poetic speech where musical quality is high (see 1.6.1), i.e., timbres themselves providing musically interesting patterns, functioning as structural elements in a composition? It may then require a highly special ambience for this type of musical listening experience, accomodating only a few listeners in some specially designed locations. Given the effect of transmission on timbre, room acoustics must be an obvious consideration if timbre composition will ever be a successful musical form.

---

* The well known phenomenon of the ear enjoying the "same" music or speech despite the horrendous situation dictated by the effect of room acoustics as we pointed out above is known as the precedence effect (see [Benade, 1976, p. 201]).

1.5 Grey's Timbre Research by Computer.

When Grey set out to study timbre by digital computer, several factors had come to the forefront. First, a preponderant set of evidence had been accumulated that suggested that certain temporal elements, especially the attack of a sound, played a critical role in the timbre percept, at least for many familiar instrument tones. Second, Risset had tilled fertile ground for digital computer synthesis of sounds of all sorts and a question for Grey was whether there was some general approach that would allow easy manipulation of timbre. Third, Moorer had just developed a sophisticated and yet intuitively simple technique, based on Fourier decomposition, to first analyze a digital waveform and then either resynthesize it from the analysis data or modify the latter and then perform the resynthesis. In principle it can handle any slowly time varying quasi-periodic sound. And finally, a statistical correlation technique that is capable of finding a representation of the relationship among a set of empirical data, optimal in some sense, had been recently applied to psychoacoustic studies by Plomp and Wessel. This technique, ideal for scaling data known to have more than one dimension, is capable of handling a large amount of data and is well suited for the computer and is known as the multidimensional scaling (MDS) method.

1.5.1 Grey's Precursors.

1.5.1.1 Risset and his Computer Synthesis Catalogue.

Benefitting from the foundation laid out by Mathews at Bell Labs, Risset was the first to explore the capability of digital synthesis of sounds over a wide range of timbres. While the synthesis goal was by no means aimed at perceptually perfect duplication of known timbres, the results were so encouraging that Risset wrote down the detailed recipes for each synthesis and assembled them into a catalogue. The main conclusion was that digital synthesis is indeed a promising approach to sound making because of its high degree of manipulability, controllability, and reproducibility. Controllability refers to the quality or precision of synthesis at a given cost or constraint. Reproducibility refers to the ability to repeat a certain performance given exactly the same instructions and the same data. Manipulability refers to the ease with which one can modify a given set of data to produce a new sound of some definable expectation. Furthermore, Risset demonstrated his synthesis in musical passages, thus establishing the computer as a powerful means

of musical composition.

1.5.1.2 Risset's Trumpet Studies.

Risset used the digital computer to analyze the time-variant properties of trumpet tones. By systematically simplifying the complex, analyzed parameters of the tones in various ways, Risset concluded that three particular features were aurally important: " [1] The attack time, with faster build-up of the low-order harmonics than the high-order ones; [2] for certain tones, a quasi-random frequency fluctuation; and, most importantly, [3] a peak in the frequency spectrum between 1000 and 1500 Hz and an increase in the proportion of high-order harmonics with intensity."

It is easy to see that these features are consistent with the source characteristics of a trumpet tone. (1) is a direct consequence of the issue of efficient transfer of energy, where the long wavelength energy is more readily radiated out while the short wavelength energy spends a long time bouncing back and forth inside the tube (see 1.4.3.6). (2) is a statement of the nonstationary character of the attack in frequency terms (see 1.2.5.7). (3) is a manifestation of the nonlinear character of a sustained vibrational feedback. This study reflects the fact that a sound such as a single trumpet tone confirms the notion that transitions from one vibration form to another are perceived as timbrally significant. Risset further breaks down the physical correlates (the vibration form) in spectral terms as temporal relationships among harmonics, and as spectral relationships among these harmonics with a continuous distribution of energy.

1.5.1.3 The Trumpet Studies of Risset, Chowning, Beauchamp and Morrill and their General Implications.

One noteworthy result of Risset's study is that the brass-like quality is preserved when the amplitude functions for the higher partials are scaled from that of the fundamental by some fixed relation. This result confirms the conjecture that (a) there is some kind of organizing element in the source characteristic that is picked up by the ear, (b) this element is some kind of amplitude envelope together with a transition of the spectral contents, and (c) exploitation of such coherent features in the source (as they are picked up by the receiver) leads to significant data reduction, which would be a highly desirable feature of a timbre study.

This observation is consistent with Chowning's finding where sounds of good quality can be made using an overall amplitude envelope, a recipe for the spectral

evolution, and the fundamental frequency independent of the basis functions used. The nonlinearity characteristic of the source is certainly picked out by the ear and is the basis for Beauchamp's nonlinear synthesis. Finally, articulation characteristics were also perceived as a timbral element, though not as notable as the "primary" feature above, and were successfully exploited by Morrill.

1.5.1.4 Limitations of Risset's Approach.

On the other hand, we recognize that the period synchronous Fourier analysis approach Risset uses involves a fundamental assumption, namely, a quasi-periodic signal. As it turns out, recent studies have indicated that the growth character of many percussive sounds is so highly nonstationary that they need more than one analysis data point, e.g., the amplitude-phase pair per frequency bin per "period," to provide the needed perceptual fidelity in the resynthesized sound. Similar observations have been made about the timbre of many speech sounds [R. Shannon] where the vibration pattern can change significantly and non-uniformly from one "period" to the next.

1.5.2 Modern Timbre Studies of Grey.

Grey studies the timbres of sixteen members of the orchestral instrument family—a choice in the tradition of Helmholtz. Like Helmholtz, Grey identifies his timbres with their physical instruments on a one to one basis, i.e., focusing on the signatorial or invariant aspects of the instrument as they are registered in their respective waveforms, ignoring all other timbral features that might vary within the same instrument. But unlike Helmholtz, Grey brings in a certain perceptual naturalness that characterizes these instruments by including the attack segments of the tones. By doing so, his studies open up a new dimension that is lacking in Helmholtz's original study of the orchestral instrument family. But unfortunately, unlike Helmholtz, Grey does not examine further the process by which timbre is perceived, choosing instead to follow Helmholtz's Fourier analysis view, even though the scope of the stimuli he considers is significantly different from that of Helmholtz, thus limiting the applicability of his study.

Grey's work is influential for the following reasons:

(1) He is the first to recognize the importance of having an adequate analysis/synthesis environment, i.e., a timbre operating environment, for successful timbral research so that it is possible to conveniently generate

related timbres by modification of analysis data followed by synthesis of the modified data.

(2) He is the first to address the issue of recoverable analysis, i.e., an analysis that permits one to recover the sound from the analysis data.

(3) Recognizing the implications of distinctive feature-extraction in simplification of the data base, Grey is the first to address the issue of data reduction in connection with effective timbre manipulation and generation.

(4) He recognizes some of the dynamic characteristics of timbre previously overlooked by timbre researchers and demonstrates the dominance of attack in perception of many natural musical timbres (although of short duration).

(5) He is one of the first to apply metric scaling to delineate the multidimensional relationship among timbres.

(6) He is the first to address the issue of timbral interpolability as a means to probe the way we organize timbres in our memory, e.g., in discrete bins separated by brick walls or in a continuum, and as a means to explore timbre space. Furthermore, he uses interpolation to test the adequacy of his analysis method.

(7) The shortcomings of his work force us to reexamine many issues in timbral perception which he helped start to address.

In his thesis, Grey states the following goals for his timbre research.

(1) The analysis and synthesis of natural timbres from distinctive features.

(2) The simplification of the complexity of physical information in timbre.

(3) A general exploration of timbre perception using multidimensional scaling.

(4) An examination of the continuous versus categorical nature of timbre perception.

1.5.2.1 The Analysis and Synthesis of Natural Timbres from Distinctive Features.

In expressing his major concerns for timbre research, Grey stresses the importance of having a research program which performs "analysis and synthesis of natural timbre for distinctive features" (p. 16) (by digital computer).

1.5.2.1.1 The Issue of Control.

He argues that computer analysis provides information about the physical properties of the timbres under study "in levels of detail not previously obtainable" He further argues that such levels of detail are "absolutely necessary" to determine perceptually important physical features because "until one can specify the physical features in sufficient detail, no psychophysical correlation is possible". In other words, the digital computer provides the means necessary for arbitrary control over waveforms for the definition of timbre needed for perceptual experiments. It is clear that such a degree of control is needed because in order to assess the importance of a certain physical (acoustic) feature of a timbre, one would most likely need to obliterate other features or modify the feature of concern in fine enough steps that meaningful perceptual scaling is possible. Therefore the role of a digital computer in timbre analysis is perfectly obvious. Control provides the same reasons for synthesis of timbre by digital computers.

1.5.2.1.2 Analysis and Synthesis Based upon Each Other.

Grey further argues that analysis and synthesis must go hand in hand because "a necessary test of the analysis technique is that it could provide information for a re-synthesis of an analyzed tone such that the synthesized tone would be *indistinguishable* from the original tone" (p. 16, emphasis in original), and because modification of physical features in the analysis could be immediately tested perceptually with controlled expectation. As a result, one can "[pin] down the critical physical dimensions in detail." Grey's idea of a research program involving analysis and synthesis of natural timbres for distinctive features by digital computer can be summarized by the diagram below.

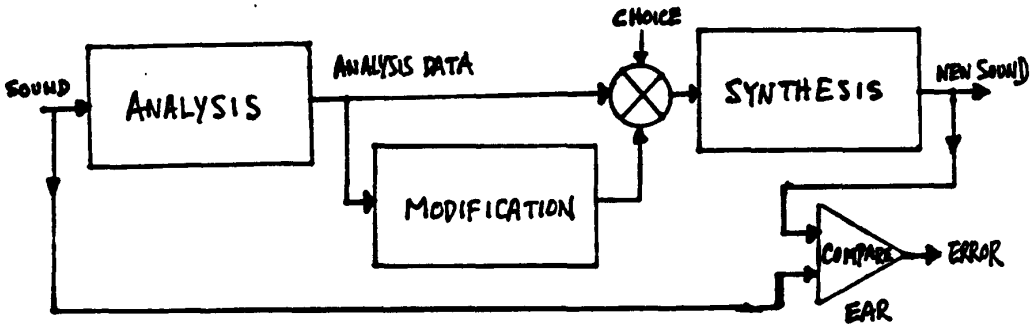


figure 1.5.2.1.2

1.5.2.1.3 The Issue of Distinctive Features.

Grey is therefore ahead of everybody in articulating this issue of paramount importance in no uncertain terms. However, for pedagogical and expository purposes, we should also point out that Grey also makes a fundamental assumption about what he calls "distinctive features" for natural timbres. He assumes that the amplitude envelope function and the frequency trajectories for the time-varying harmonics of the timbre are "distinctive features" without a debate. His assumption may or may not be valid but it is important to point it out because not everyone sees it that way (see for example Schouten's and Erickson's numbered list of timbral features in 1.6 and Schaeffer's rules on timbral perception in 1.7). But Grey's unwritten reasoning is not hard to guess.

First, the notion of timbre as correlated with its time-varying harmonics is a direct extension of Helmholtz's notion of timbre for periodic sounds.

Second, the notion of the ear as a "Fourier analyzer" makes it reasonable for Grey to believe that the time-varying harmonics are physical features by default. (Actually, the ear's role as a "Fourier analyzer" is an idealization applicable only to asymptotic conditions for certain stimuli; see 1.2.)

Third, Luce's work [Luce, 1963] uses very similar choices of stimuli and very similar analysis/synthesis techniques, namely, windowed Fourier analysis and its inverse transformation. Furthermore, Beauchamp [Beauchamp, 1975], Keeler [Keeler, 1972], Risset [Risset, 1966], and others have all approached the problem of sound analysis and synthesis from Fourier based methods (with varying degrees of success). In short, it was a very popular method when Grey came along.

Fourth, the acoustic properties of many musical instruments seem to be equally suitable for Fourier analysis—on the surface. (Actually, musical instruments, like the ear, are physical devices whose behaviors are suitable for Fourier analysis only under asymptotic conditions of stationarity which, as Winckel rigorously points out (see 1.4), are physically unrealizable (even within the ear's window of observation). Furthermore, the normal modes of vibration of many physical instruments are not part of a harmonic series. This in fact is the case with many percussive instruments.)

Fifth, Moorer's hetrodyne filter analysis/synthesis appeared on the horizon as Grey came along. It would only be natural that Grey take advantage of it.

Sixth, it is reasonable to assume that the relative magnitudes of the harmonic

amplitudes provide discriminating information for "distinctive features." But first of all, not everyone agrees that such information alone is indeed what the ear uses as we have pointed out above in regard to Schouten, Erickson, Schaeffer, and others. In fact, Charbonneau's data reduction experiment (see 1.5.7) later contributed a view concerning the perception of timbre that is more consistent with the organizing capability of the ear. Second, these magnitudes can wander from "harmonic" to "harmonic" as a function of time, as many percussive timbres do. Thus it seems that a more sophisticated pattern organization must accompany the ear's perception of timbre.

Finally, we should also note, for pedagogical and expository reasons again, that analysis of ***distinctive features*** is neither necessary nor sufficient for synthesis. It is not necessary because perfect synthesis does not require ***nontrivial*** analysis. For example, we can "analyze" the sample sequence of the digitized waveform apart into a sequence consisting of the samples with even index, and a sequence consisting of the samples with odd index. Then we can interleave these sequences for a perfect synthesis. It is not sufficient because if we do not include enough physical features under the rubric ***distinctive***, perceptually identical synthesis is impossible (e.g., the set of distinctive features normally associated with a speech vowel in phonetics is not sufficient to recapture the timbre of a particular enunciation because speaker characteristics have not been included—the description of the "universal" /a/ does not suffice to regenerate the sound of Jane Doe's /a/.) unless we include every physical feature as a distinctive feature. The point is that the issue that Grey tries to address is a little more complicated than the way he poses it. We will address it in detail in 2.4 in terms of the notion of analysis and synthesis of perceptual importance trees.

1.5.2.2 Simplification of the Complexity of Physical Information in Timbre.

Grey correctly points out the importance of data reduction as a means to eliminate "those factors which are not perceptually potent," and as a way "to make manipulation of timbres easier using a much simplified representation of tonal material."

1.5.2.2.1 Helmholtz's Data Reduction.

Helmholtz is the earliest to take advantage of the physical consequence of a stationary process. He recognizes that for the class of periodic signals significant data reduction in an analytical description of their timbres is possible. That is, if we know that the signal is exactly periodic with period P, then we need only $R \cdot P$ numbers instead of $2 \cdot R \cdot T$ to completely specify a signal of length T and Nyquist rate R. Note that T is usually much greater than P for musical timbre. Note also that the DFT of the period is also $R \cdot P$ wide, which accounts for both the magnitudes and phases of the nonredundant Fourier components. From the viewpoint of the receiver, once a sinusoidal response is established, why should the ear have to keep track of what is going on until something changes? In other words, there is nothing to *update* as time passes for as long as the response of a particular fiber is in a sinusoidal motion. This observation together with the one that the ear has a non-infinitesimal perceptual gridsize means that one can take advantage of the ear's data reduction behavior and perform data reduction in the control space as well. What we have just said is that even without taking the further step of discarding the phases of the Fourier components, Helmholtz's model of timbral perception implies that the ear performs data reduction whenever it is given highly redundant acoustic information. Later, we will show how we might be able to go further to take advantage of the ear's data reduction capability for natural timbres for which the periodic condition fails. (Note that by data reduction we always refer only to the possibility where the perceptual identity of the timbre is not distorted. On the other hand, we try to distinguish between data reduction that occurs in the ear and that which occurs in the control space. Our purpose is to discover the ear's data reduction behavior so that we can emulate it in the control space). But first we will examine what it might all mean from a perceptual viewpoint.

1.5.2.2.2 Notions of Data Reduction.

There are physical data reduction and perceptual data reduction. We will only concern ourselves with the latter. Perceptual data reduction hinges on the assumption that not all of the physical data in the acoustic stimulus of a timbre is equally important perceptually and that some of the data can even be ignored without affecting the perceptual integrity of the timbre.

1.5.2.2.2.1 Notion of a Perceptual Grid.

The existence of a perceptual grid of finite gridsize in the ear, in both the amplitude and time dimensions, is evident from Fechner's or Weber's Law. Our experience tells us likewise. For example, certain alteration in the sample amplitudes of a timbre's waveform may not lead to a perceptibly different timbre, although other alterations will. The difference, however, does not always lie in the magnitude of alteration on a sample by sample basis. There is a strong indication that the difference lies to a large degree in context. That is, the extent to which a local wave pattern can be changed without alerting the ear depends on the local pattern of changes. For example, alteration in a single sample amplitude depends on the original amplitude variation in the neighborhood of change. We will discuss this point in more detail in 2.2. Here the principle behind this type of data reduction lies in the finiteness (as opposed to being infinitesimal) and elasticity of our auditory grid of perception and, has been expressed in various forms, in amplitude and in time, by the laws mentioned above. It is therefore possible to perform data reduction on quasi-periodic signals the way Helmholtz does by thinking of them as truly periodic. For if the ear's updating mechanism decides that the small changes in the vibration pattern from period to period do not constitute innovation, or new information, then they are simply ignored and the ear is said to have performed (perceptual) data reduction. Grey's notion of data reduction is essentially the same.

1.5.2.2.2.2 Analysis-Dependent Data Reduction in Control Space.

There are a few subtle differences between Grey's data reduction implication in his line-segment approximation and the idea discussed above. First, the physical data on which data reduction is performed is not the waveform itself. They are the amplitude functions of the harmonics of the analysis, or the parameters in the control space of synthesis. If it can be linked to data reduction in the sample amplitudes in the waveform, then it is equivalent to the data reduction we discussed above based on the ear's update mechanism and the finitenes of the perceptual grid of audition. Otherwise, it is something else.

In his work, Grey observes that alteration in the microfluctuations in these harmonic amplitude functions do not have perceptual significance, therefore the data involved in specifying these fluctuations can be replaced by line segments. Thus, horizontal line segments for all the harmonics concerned over a certain dura-

tionn imply that we have a locally periodic signal in Helmholtz's sense. In general, data reduction in the analysis data is considered a consequence of perceptual data reduction in the ear in the sense discussed above.

However, from a pedagogical viewpoint, we must point out that many of the microfluctuations could have come from channel cross-talk or other analysis limitations. For example, when the original analysis is viewed against its spectrogram, we often notice that the wild fluctuation in certain harmonics of the former is not reflected in similar fluctuation (of the same time scale) in intensity over corresponding frequency bands of the latter in the same time region. (See, for example, graphical illustrations in [Strawn, 1982]). Therefore,data reduction in the control space is specific only to this particular analysis approach and not necessarily others. By the same token, other analysis approaches may not need the same kind of reduction because they might not have produced the volume of data this particular approach generates in the first place.

1.5.2.2.2.3 Perceptual Hierarchy of Features.

Another kind of data reduction hinges on some kind of hierarchical organization of timbral features the ear might actually exercise. As a result, obliteration or modification of different features would have different effect on the overall perception of the timbre in consideration. Naturally, context also plays a role but in a more complicated way because the features are usually composition of more elemental features. And it is the relationship of a particular feature relative to the rest on the same level that determines its perceptual importance. And it is the hierarchy that determines the impact of data reduction on any particular feature or features. Grey's work does not address this type of data reduction and we will discuss this in more detail in terms of the metrical perceptual importance tree description of timbre in 2.4 and 2.5 as part of our dynamic formulation of timbre.

1.5.2.3 General Exploration of Timbre Perception using Multidimensional Scaling.

In discussing the importance of a multidimensional scaling (MDS) approach to exploration of timbre space, Grey underscores the importance of modelling the space by a small number of "salient dimensions or features of classes of sounds" (p. 18).* He reasons that the notion of a psychological distance function commonly

* There were 20 subjects who participated in one experiment, and 15 who par-

understood among musically trained ears is most appropriately modelled, because of the spatial notion implicit in the psychological measure, by a Euclidean distance measure in some low dimensional geometric space. Based on these considerations, Grey makes the conscious choice of a low dimensionality and applies the metric version of Kruskal's MDS program [Kruskal, 1964].

The advantages of the MDS approach are:

(1) It has the ability to handle a large amount of data easily and derive a reasonable (but not necessarily unique) representation of the data relationships in some prescribed distance measure by means of a digital computer.

(2) It has the ability "to generate a data structure on the sole basis of the perceptual judgments which then may be interpreted with respect to the physical parameters of the sounds" (p. 18).

(3) It does not involve the "very precarious if not dubious tactic" of verbal rating scales.

However, for expository reasons, we must note that (1) the method makes assumptions about the Euclidean nature of the timbre space. (2) It does not provide information on the dimensionality of that space. (3) It does not specify what the dimensions are.

Specifically, Grey's result reveals a very unsettling question in light of his concern for an analysis/synthesis program for distinctive feature extraction. Namely, the MDS dimensions are not the physical features in the analysis/synthesis program but are only vaguely related to certain trends among them. Grey considers these MDS dimensions (three of them altogether) to represent salient features of classes of sounds. Several questions naturally arise:

---

ticipated in a second, making a total of 35 sets of data. The data were similarity judgments on pairs of timbres. There were sixteen tone pairs which were rated on a scale of 1 to 30, 30 being very similar, 1 being not similar at all. The timbres were from the oboe, the French horn, the bassoon, the trumpet, the flute, the saxophone, strings, the English horn, the trombone, and the clarinet. The tones were processed so that they had the same duration and pitch (and thus the steady-state portions were clipped).

(1) If these MDS derived features are salient or distinctive, how can we use them for the purpose of synthesis and modification?

(2) If these MDS derived features are not meant for synthesis and modification, what is the meaning of an analysis/synthesis program for distinctive features of timbre?

(3) If these MDS derived features are a result of the auditory processor's data reduction effort, what is the transformation that specifies their functional dependence on the physical features used in the analysis, synthesis, and modification?

(4) Is the dimension count of three really sufficient in specifying timbres of the class of sounds under study?

(5) Are sixteen data points really sufficient to reliably describe the nature of the space? Can we, from them, answer such questions as whether the dimensions are independent or orthogonal (since orthogonality depends on the notion of angle and since angle is defined solely by the covariance function of the data)?

(6) Does the surprising confusion in the placement of data have something to do with MDS or with the nature of the stimuli (which have the steady-state portion truncated and which, in the case of some timbres, are placed in an unnatural pitch range for the sake of control—that is, for the sake of equalization of pitch)?

1.5.2.4 Examination of the Continuous versus Categorical Nature of Timbre Perception.

Grey's loop is generally thought to demonstrate that timbre is in general interpolable because the tones sound interpolated. But we must recognize that the tones are very short and perceptually very close in timbre. In a certain sense, these tones are just as artificial as the steady-state tones of Plomp because the long sustained tones characteristic of most of these instrument tones such as the French horn, the oboe, and the bowed string instruments are truncated.

In fact, we don't think the French horn and some of the other tones sound very much like those we normally identify with the classical orchestral instruments. Furthermore, there is no way the breathy character of a flute can be easily realized with Grey's analysis-synthesis algorithm, and in fact is absent. Finally, timbres of

certain types are more strongly characterized by their steady-state periodic behavior (because of a lack of identifiable attack pattern, see Luce [Luce, 1963]) while others have very strongly identifiable attack characters. This diversity leads us to think that interpolation can be better achieved from a dynamic viewpoint.

Therefore, while Grey takes the first step to try to examine this issue, he remains limited by his methods, namely the static notion of timbre (even though he recognizes the functional dependence of timbre upon time) and the rather cumbersome way of doing analysis using the hetrodyne technique.

1.5.3 The Data Reduction Studies of Charbonneau.

The apparent contradiction between the large amount of data necessary to define a timbre from a short-time Fourier analysis and the small number of dimensions attributed to the identification and differentiation of timbres (by similarity judgment) by multidimensional scaling studies suggests that not all of the details of the harmonic amplitude and frequency functions are perceptually significant. At the same time, the high degree of perceptual approximation that frequency modulation (FM) synthesis is able to provide for many musical timbres with such an economy of description, suggests that Fourier-based additive synthesis has not addressed the issue of perceptual minimal requirements adequately. In this light, Charbonneau observed that "Data reduction is thus not only of obvious interest for the synthesis, transformation, or transmission of sound. It also permits a deeper understanding of the truly relevant features of hearing (specifically the invariable elements of sound perception)."

Starting from the sixteen reference tones provided by Grey, Charbonneau investigates the data reduction possibilities in the harmonic amplitude functions, the "harmonic" frequency functions, and the onset times of the harmonic amplitude functions of a tone. His findings include the following:

(a) The spectral envelope is an important perceptual attribute in timbre recognition.

(b) The harmonic amplitude functions evolve in a coordinated fashion so that each one can be described by the same amplitude function scaled by the harmonic's peak value (as a function of time). That is, a given amplitude function is a scaled and shifted version of any other:

$$f_j(t) = A_{j,k} f_k(t + C_{j,k})$$

where f_j and f_k are the amplitude functions for harmonics number j and k, respectively. Equivalently, the harmonic amplitude function can be replaced by a "spectral envelope" evolving as a function of a single time variable. That is, there are time indices $\alpha_1, \ldots, \alpha_n$ such that if $f_{j,\beta}$ denotes the amplitude of harmonic number j at time β, then

$$(f_{1,(\alpha_1+\delta)}, \cdots, f_{n,(\alpha_n+\delta)}) = \lambda_\delta(f_{1,\alpha_1}, \ldots, f_{n,\alpha_n})$$

where δ is any time increment and λ_δ depends only on δ. α_k can obviously be chosen to be the time when harmonic amplitude function k reaches its maximum (as Charbonneau chooses). As a result, we are not talking about a spectral envelope in the strict sense, because the α_k are not necessarily the same.

(c) The harmonic frequency functions do not have individual perceptual significance. Rather, they all evolve as some constant multiples of the fundamental frequency function.

(d) The onset times can be approximated by a polynomial function of the harmonic number (index) of degree much smaller than the number of harmonics. In other words, the harmonics seem to move in some coherent fashion.

Charbonneau's results reveal much about the mechanism of data reduction for the purpose of synthesis, transformation, and transmission of the tones involved within the context of short-time Fourier analysis. But more importantly, the observations point to a pattern of perceptual organizing features that help reduce the large dimensionality of timbre into a much more manageable and well-organized set of dimensions, i.e., the group of dimensions subsumed under the notion of an evolving spectral envelope of size $m \cdot \overline{n_A}$ (rather than $\sum_{k=1}^{m} n_{A,k}$, where $n_{A,k}$ is the number of breakpoints for the $k^{\underline{th}}$ harmonic amplitude envelope, $\overline{n_A}$ is the number of breakpoints for the averaged amplitude envelope, and m is the number of harmonics involved in the analytical description of the sound under study); the group of dimensions describing the single fundamental frequency trajectory of size n_P, where n_P is the number of breakpoints for the fundamental frequency trajectory; and the group of dimensions concerning the initial phase relationships among the harmonics n_O where n_O is the degree of the polynomial that approximates the onset times.

1.6 Erickson's *Sound Structure of Music* and the Notion of Relationships and Organization in Timbre.

Erickson's *Sound Structure* does not propose a research program in Popper's sense of a falsifiable theory [Lakatos, 1978], and his experimental findings are usually within the context of informal listening, but it brings us to the forefront of the issues of timbral organization, the organizing role timbral features (especially those of *temporal* origin) play, the way these features figure in our perception in a hierarchical order, and in general the notion of dynamic relations among timbres, drawing examples from a wealth of "world" music.

1.6.1 Musical Context.

Contemporary research in musical timbre is heavily accented towards spectral composition of timbre and much has been done within the context of isolated tones. However, Schoenberg [Schoenberg, 1969] writes that "A triad standing alone is entirely indefinite in its harmonic meaning." From the viewpoint of timbral compositions, the harmony analogy may be quite relevant. It seems therefore that without addressing the issues of timbral relationships and organization, we are not prepared to provide a composer the timbral operating environment to create compositions in which timbre *functions* as a structural element. Also by working with either isolated timbres or groups of timbres mainly characterized by their spectral compositions, researchers may have unnecessarily obscured certain aspects of timbral perception that make the perception of timbral relationships from sound to sound possible. A study of perceptual attack time by Gordon [Gordon, 1984] suggests that when temporal character comes into play, there is a kind of form relationship among the sound events on the micro-scale that the ear can detect and try to organize. (See section 2.1.6.3 for greater detail.) To discover these relationships, a proper selection of musical context seems to be in order. Also it turns out that listeners tend to have more difficulty establishing relationships among timbres which are mainly characterized by their spectral composition. There are many well-known examples. To some, the timbre piece in Schoenberg's *Five Pieces for Orchestra* belongs to this unfortunate category. On the other hand, it is well known that perception of the local timbre relationships is an integral part of speech perception, at least, in the English language.

In this respect, ancient Chinese poetry is in fact a genre of timbre composition in the true sense, i.e., one in which we can hear the timbre relationships, the timbre functions, and the timbre structure, clearly, unencumbered by metric structure in pitch or rhythm. (The kind of structure we are referring to here is typical of a tune where pitch and rhythmic patterns (melodies) fuse with and transform (even distort) the tonal quality of the lyrics.) An interesting aspect of the timbre relationship in these poems is the prominent role timbral features of temporal origin plays. It seems that these temporal features really help bring out the timbral relationships and these temporal features epitomize the gestural features and nuances reflected in so much of contemporary music, especially among the expressionists.

While this may be a modern adventure in the history of Western music, it is what one may describe as the normal mode of musical listening in many Third World cultures, particularly China's—at least in the more sophisticated part. One may speculate that the Chinese, the Indochinese, and the Indians, having been relatively undeveloped scientifically and technologically when their art came to maturity many centuries ago, did not understand resonance very well, and therefore they could not make very richly resonating instruments. In particular, opera singers in the Chinese tradition are often criticized, by those familiar with Western operatic style, for not singing but shouting in loud passages. But for all the spectral richness they miss, their aesthetic sense probably has taught them to explore the *expressiveness* of the temporal dimension. The Chinese are known to relate sounds to some notion of form in a mysterious way, but when one listens very carefully to the way ancient Chinese poetry is recited, the notion of temporal form in timbre, as manifested in the loudness, pitch, and spectral dynamics, within a single speech sound, is immediately apparent. In fact, the calligraphy, which usually goes with the poem, often reflects these temporal forms in the dynamics of shape, thickness, and strength. This is perhaps no accident, especially in light of the central role timbre composition plays in a classical Chinese poem. Erickson, in his *Sound Structure*, describes ancient Chinese lute playing in similar terms. The musical context of the two examples cited here is strongly form oriented in the temporal sense as opposed to being spectrally or pitch oriented. (This is not to say that the spectral or pitch characteristics do not figure.)

1.6.2 Timbre as a Temporal *form* Phenomenon.

At one extreme, pulse trains of various shapes run at various rates constitute a kind of formless stationary temporal phenomenon very much like sine waves. The timbral characterization of these type of sounds is *grain-like*. These include what is known as rustle noise. At the other extreme, we have all kinds of variations of nonstationary temporal phenomena known as attack, decay, amplitude modulation, frequency modulation, frequency glide, spectral glide and spectral alternation that make up the character of most sounds in the world. In between, we have the more exotic kind of sounds ranging from chord-like timbres from fusion of a small combination of pulses or pulse-like waveforms, to a spectro-temporal phenomenon known as drone music as well as the ebbs and flows of audible beats. The point of making this list is to emphasize the fact that timbres described by a spectral distribution uniform over time is a very restricted class indeed.

1.6.3 Hierarchical Perception of Timbre.

1.6.3.1 Grey's "Dilemma."

As we have seen in 1.4, Grey's analysis and synthesis model consists of a set of amplitude envelope functions for the harmonic partials of the tones under investigation. If there are M partials and if each partial requires N_k samples to specify it, then the dimensionality will therefore be between MN_{max} and MN_{min}, where N_{max} and N_{min} are the maximum and minimum of N_k over the M partials. Therefore, even if there are only a few perceptually significant harmonics in existence, the number of dimensions is fairly sizable even if there are on the average three or four segments in each amplitude function. Yet Grey's MDS of the timbres determined by these harmonic partial amplitude functions yields the result that there seem to be only three distinct timbral dimensions. On the one hand, one can make the statement that these amplitude functions are merely control parameters for the timbres. But on the other hand, if they do not explicitly describe certain timbral behavior, what is their purpose from a perceptual standpoint other than the way that they constitute some known method of recovering a waveform? One could argue that timbre actually has a huge theoretical dimensionality but somehow the number of *independent* dimensions is very small, maybe two, three, or four. And somehow the harmonic amplitude functions reflect that huge theoretical dimensionality. Wessel, in fact, in one of his MDS's of stationary timbres, reports that the

harmonic amplitude functions are not independent. However, it is not clear how the process of dimensionality reduction takes place. In fact, the process involves not reduction alone but a transformation to new dimensions, i.e., they are no longer amplitude functions any more but some functions of them if one can indeed detail such a transformation.

1.6.3.2 Hierarchical Organization of Dimensions.

It might be more productive on the other hand to assume that in fact the perceptual dimensionality does correspond to the physical dimensionality but the dimensions are not in general the partial amplitude functions but rather a hierarchy of features at the root of which are some global or anchoring features such as the growth and decay characteristics and the spectral content of the stationary region, and these are reflected in the MDS results of Grey, Wessel, and others, where further discrimination is either not possible because of the coarseness of the perceptual data matrix or not intended by the experimenters. This is in fact more compatible with the knowledge that both Plomp and Slawson have devised a four dimensional representation for their stationary stimuli. For in their result there is no temporal feature discussion whereas in Grey, Wessel, and others, there is always at least one stimulus made of nonstationary acoustic events.

If the perceptual process is indeed hierarchical, then:

(1) Further discrimination along the growth and decay dimension may lead to separate growth and decay dimensions. And further discrimination along the growth dimension may differentiate still more dimensions such as the average rate of growth against the instantaneous behavior of growth that involves the question of whether the growth within a period (if it exists) is linear or nonlinear. And the decay dimension may bifurcate into a dimension which has only one decay time constant such as is the case with the marimba tones (hence the dimension represents the decay time continuum) and a dimension which has two or more time constants such as is the case with the piano tones with the characteristic "singing" quality.

(2) The perceptual features may be correlated with easily observable physical characteristics of the waveform where they are also organizable (according to their perceptual importance) into a hierarchy or tree. The existence of a perceptual importance tree of physical data tells the analysis what to

look for in a hierarchical manner and then tells the synthesis how best to trade perceptual quality for computation economy for a given algorithm. In other words, the importance tree will form the basis for designing the analysis and synthesis algorithms of sounds in the map of their timbres.

(3) The physical features or timbre dimensions are also hierarchically organized. As a result, timbre dimensions can be naturally organized into a hierarchy. For instance, the marimba and the piano may have the same node *viz.*, the decay dimension, in the hierarchy of dimensions, whereas all percussive string instruments with more than one decay time constant may share the same node dimension.

Note that it is the dimensions which are hierarchically organized. Since the dimension distance can vary from dimension to dimension between any two timbres, it is not at all clear whether all timbres can be organized into hierarchies as Lerdahl attempts to do. The conflict becomes clear when competitors for dominance in distance occur at the same level in a hierarchy. For example, the marimba may be close to the piano in the decay dimension but not in the spectral dimension at some comparable time location. In this case, it is possible that some timbre without a decay may sound closer to one or the other in an overall judgment. In this case, we have just brought up the issue that we have not decided (and may never be able to)—the relative importance of two or more dimensions at some given level of a hierarchy, say the $k^{\underline{th}}$ level of an M-ary tree. However, Erickson correctly points out the intuitive place the notion of hierarchy has in the perception of timbre.

1.6.3.3 Pitch as an Organizing Element.

Schoenberg regards pitch as a dimension of timbre. Arguing from a dimensionality point of view, a waveform of N samples quantized to M levels (corresponding to perceptually discriminable quanta) has in theory N dimensions (for a total of M^N perceptually distinguishable waveforms). This description remains valid if we model a waveform as a realization of an independent increment stochastic process. (Note that from an information-theoretic viewpoint, each acoustic waveform the ear receives is indeed an outcome of a random process with a certain probability). Such an independent increment process can be organized into a tree whose branches all originate from the root and each extended branch constitutes a realization, i.e., a waveform, with a certain probability. A well-concentrated probability distribution

effectively limits the number of practically encountered waveforms and hence the set of all possible timbres (possibly smaller in size).

However, if the waveform is perfectly periodic, then the waveform space is reduced significantly. For example, if the Nyquist sampling rate is N_s samples per second, and if each period contains P samples, then $N_s/P = f_P$ is the frequency of the periodic vibration and the dimension reduction is f_P-fold. In this case, there are at most P dimensions (and M^P perceptually distinguishable members) for the class of periodic waveforms of period P whose Nyquist rate is N_s. The point here however is that periodicity is recognized by the receiver as an organizing element. Perhaps the dimension-reduction, as it appears through time, is perceived as providing no further new information and the timbre is consequently perceived as boring or "electronic," whereas natural waveforms are at best quasi-periodic. In that case, new information keeps coming in and the sound is perceived as "live," for as long as the new information or "innovation" is not so dramatic that the receiver cannot adapt its measurement to organize the data against existing ones.

If the period length changes from P to $P+1$ in two cycles (with uniform amplitude change), then the dimension calculation becomes $P+1$ (M^{P+1} waveforms) compared with $2P$ (M^{2P} waveforms) when there is no repetition and P (M^P waveforms) when the repetition is exact. Note that the organizing that takes place over a quasi-periodic signal in general involves adaptation along the time axis as well as the amplitude axis. The ability of the ear to adapt (to some extent) to resolve a piece of new information is generally regarded (by a Darwinist) as a reflection of the ear's host's desire to be in harmony with, or to know its surroundings, and this desire has been realized as indicated by the host's and ear's symbiotic survival. This reasoning implies the following:

(a) The growth and decay characteristics in the (overall) amplitude envelope of the waveform as well as the period trajectories can be thought of as organizing elements for the ear and these perceptual characteristics are reflected as timbral features.

(b) The adaptation characteristics imply that the analysis and synthesis algorithm should certainly reflect this receiver behavior if it is in fact receiver-based.

It is in fact not surprising that the period trajectory should control how we

perceive a timbre, hence a timbral feature or dimension in the hierarchical sense elaborated earlier. The extent of deviation from perfect periodicity in terms of excursion and rate is certainly reflected in the timbral character simply from our experience. The same can be said about the description of vibrato. In fact, we will discuss later that our laboratory experience shows that essentially the same two period trajectories which are different only in a continuous phase shift of one or two samples per period during an octave drop result in very different timbres. In this case, the initial and final frequency remain the same between the two waveforms and the frequency uncertainty is so large on a per-sample time-window that there is no basis for us to believe that one can normally perceive the instantaneous "frequency." But the cumulative phase shift leads to a pattern significantly different from the starting pattern. Therefore the interpolated pattern in between becomes much more different from one period to the next than is the case if the end patterns are significantly in phase and hence much more similar.

1.6.4 Relationship, Organization, and Structure of Timbre Space.

We can understand why Erickson went out of his way to study the relatively unknown music discussed in 1.6.1, i.e., Chinese lute playing and Tibetan religious chants. For while traditional Western instruments can provide a lot of insight into the inner workings of timbre perception because of their familiarity, i.e., the strongly established source-receiver relationship, because of their perceived naturalness in contrast with laboratory tones, because of their strongly resonating properties, and because of the body of scientific knowledge we have accumulated about them, there is at least one disadvantage in limiting our studies to the timbre of these instruments only. Since Western music is so rooted in pitch relationships as the central structural element of a composition, we are not very conscious of timbral relationships. We are more trained to identify the sources and contrast them, especially by name of the sounds, than to discern differences in some metric sense of timbre, as we do for, say, pitch. It may have to do with the multidimensional nature of timbre. But people have begun to ask the metric question lately.

Once we have a hierarchy of timbral features, then we can see how one feature relates to another in importance on the one hand, and how the same feature of one instant relates to that of another instant on the other hand. As a result, we can then relate a timbre feature of one sound to the same feature of its neighboring

sounds, one step removed or two or more steps removed. Or we can then relate how a timbre feature changes to another at comparable stages in the evolution of two or more neighboring sounds. Therefore, understanding the internal organization of timbre permits organization among timbres of discrete sound events.

Within a sound, a timbral feature can be a context for what follows. Similarly, as a group, relationships among timbres can form a context for what comes later. These relationships can enlighten us as to what seems to be an element of extension, an element of contrast, and an element of balance. Continuity, parallel translation, contrast, and balance become executable functional elements in a timbre composition.

1.6.5 The Parallelogram Analogy of Timbre of Ehresman and Wessel.

Similarly, a structure in timbre space on a local scale can be described by a constructive approach using the notion of interpolated continuity. Once the local structure is delineated, we may contemplate going onward. But if the method of exploration is based on the notion of interpolated continuity, then the structure is always associative. But since the organization of dimensions is hierarchical, we can see how a hierarchical structure, such as one envisioned by Lerdahl, may emerge from an associative structure.

However, the interpolation between two timbres involves the principle of constructive approach above applied in several dimensions. An intuitive way would be analogous to the parallelogram approach introduced by Ehresman and Wessel following the paradigm of Remelhardt *et al.* However if the dimensions are actually hierarchically organized as we seem to have arrived at earlier, then this approach would be applied recursively. That is, one subtree of dimensions must be exhausted through a push-down approach before "popping" out into another subtree, for all the relevant branch-dimensions to be visited and interpolated. Certainly, the temporal character must be consistently taken into account. Therefore the dimensions which are of spectral nature by themselves must be visited on a dynamic basis in a temporal sense. Thus, we see that timbre interpolation must be regarded as a process of successive approximation and we will discuss this in a separate chapter in more detail.

1.7 The Sound Objects of Schaeffer.

In his monumental treatise, *Traité des Objets Musicaux*, Pierre Schaeffer attempts to translate the insight from his vast experience with electroacoustics and music concrete, as a composer and as a researcher, into a pedagogic paradigm in the classification and treatment of sounds and of musical events as a way to cope with the new universe of sounds brought about by new technology as well as by a shrinking geographic world. In fact, he is calling for no less than a revision of the fundamental view of Western music. He sees the need for a global convergence of order as our sound universe continues to expand, from synthetic means and from non-Western cultures.

1.7.1 Cycle of Listening and Distinctive Feature Reduction. He recognizes the key to this order lies in the link between the source and the receiver, i.e., our perception to all those sound creatures, some of which by themselves can be quite monstrously perceived. So he looks to music's famous cousin, *viz.*, spoken language, for answers. So if the linguist can exhaust the "alphabet" of spoken language, by writing down a small number of phonemes from which all speech sounds are built, then perhaps such objective atoms as the phonemes may have their analogues in a language of sound. This leads him to formulate the functional cycle of listening.

He proposes that there are four phases, not necessarily chronological or physical, in the process of listening, i.e., from the moment an external sound event hits the ear to the moment when we form an opinion of what we hear, such as "what is this?", "what is that?", when we try to identify the sound with its source. The first stage is necessarily concrete, objective, and tied to the source (this of course becomes not necessarily meaningful in the realm of computer generated sounds). Then the second stage includes the crude analytical listening in the ear, probably corresponding to the unconscious analytical processing of Helmholtz's mechanical resonators of the inner ear. This stage is still a physical phenomenon, but an internal one and is so complicated that he calls it concrete but subjective

The third stage involves a higher level involvement which selects and controls elements of the analysis result of the second stage consistent with certain objectives at high level. Here the context, attention and all the efficient action from higher level makes him consider the stage both subjective and abstract in the sense of

abstracting something from something else and of being no longer tangible in the physical sense. But in order that this abstract object could be communicated, the idea must be objectified into some common terminology and language. This is stage four which he sees as an emergence of the "content" of a sound that is referential, i.e., that which one can reference to or relate to in a communication sense and contrast its detachment from the source with the external sound object which is attached to the source.

This notion is of course of paramount importance *if successful* in view of what the notion and classification of phonemes have done to all languages. Moreover, it could be very important in our attempt to understand the fundamental question of timbral interpolability. More precisely, if every timbre sequence sounds smooth, then there is no problem. But if some timbre sequences do not, then a natural question to raise is: Is it because we haven't captured the essence of our perceptual function or is it because fundamentally we are cognitively tied to the source, thus leading to a kind of perceptual entrapment from which a progression away cannot occur? Within this context, Schaeffer's reduction program would go a long way in answering the question of whether source-independent timbre features can be consistently identified under very general conditions, provided such reduction can be achieved. These source-independent features are described in Schaeffer's perceptual rules of timbre. Aesthetically, of course, such reduction is useful for certain compositional objectives. The purpose of this reduction exercise is to train the ear to perceive a sound in terms of its form and its contents, the common determinants of all sounds. The form is referred to by the manner in which it is realized on a sound as treatment or *facture*. It refers to the global temporal features such as growth and decay of a sound. The contents can be thought of as the local pattern of variation or its DFT in Fourier terms. Thus the perceptual reduction exercise essentially tries to get the ear to focus on two initially factorizable features of timbre on the root level of the perceptual hierarchy. From those, differentiation of features is pursued further. We can think of the sound object typology given in *Traité* as a coarse division of the first hyper-quadrant defined by the group of dimensions of *form* and the group of dimensions of *content*.

Still another question, even more fundamental, is "Is reduction possible?" How do we answer this question? In linguistics, we know we can. But at the same time,

our experience with continuous speech tells us that it is not at all the individual phonemes that furnish complete, even crucial, information on the perception of continuous speech. It is in fact the transitions, or relationships between phonemes, especially between consonants and vowels that provide key discriminability. And here the aesthetic value of music differs fundamentally from the intelligibility value of language. So it is an important question that Schaeffer's thesis has raised that would only be answered in time and practice.

1.7.2 Schaeffer's View on Distinctive Timbral Features.

Schaeffer observes: *"La perception musicale est qualitative. Les mêmes causes (physiques) n'ont pas les mêmes effets (musicaux)."* He cites the case of the variable rate pulse train and its bandpassed derivatives as an example to show how the sounds change qualitatively from a discrete sequence of short sounds through a kind of grain-like timbre to the timbres of continous sounds, as the rate is changed monotonically. This example has in fact generated much interest among musicians with the advent of new technology—consider for example Stockhausen's *Kontakte* (see Slawson [Slawson, 1985] for a description). This phenonemon essentially confirms our common experience that the ear is more complex than the behavior of the physical appearance of sound might suggest. Underlying this mysterious surface is the manifest of interplay of various feature-extracting mechanisms that our perceptual faculty is capable of. Surely there is a threshold and a region of dominance of each mechanism. In the particular example our basilar membrane can be set to vibrate no lower than a certain minimum frequency on the one hand, and there is a temporal resolution threshold on the other. When the threshold is crossed, there is a switch in the detection mechanism. Nevertheless, the transition is smooth which demonstrates again the versatility of the basilar membrane as a simultaneous time-frequency detector (but not necessarily mysterious in view of the mechanical nature of the detector). He therefore observes:

"Corollarie: aucune correspondance n'est assurée entre une progression graduée en paramétres et une échelle de valeurs musicales." Even from a perceptual viewpoint without the assertion of musical values, this cannot be more true according to our experience in timbral interpolation. But of course, again, it does not mean that Schaeffer's corollary means that timbral interpolation is impossible. All it means from our viewpoint is it is crucial to discover perceptually relevant data. In fact,

our experience shows the dynamic relationship of local timbres within a sound may just provide the key to solving this in general difficult problem involving transition that Schaeffer points out.

1.7.2.1 Perception of Attack.

Schaeffer observes that *"la perception musicale d'attaque était en corrélation d'une part avec la* dynamique générale du son, *c'est-à-dire avec l'*évolution énergétique, *et avec le* contenu harmonique *d'autre part"* (page 224). While not practiced by researchers using short-time Fourier analysis, this view is nevertheless widely accepted. Later, we will see that the dynamic description of timbre we derive from perceptual considerations is very similar to Schaeffer's observation. Schaeffer went on to note that there appears to be a certain temporal threshold that quantizes our perception of attack: "In the case where the attack time is between three to ten milliseconds, the slope of the attack is perceived to be the same" (our paraphrase of a passage on page 228). While the exact number is different, our experience in interpolating between a bell-tone with a vertical slope for the attack and a bird song of octave trajectories seems to support Schaeffer's assertion. This threshold appears to impose real (physical) barriers to timbral interpolation of this type, i.e., not relating to our lack of understanding of our perceptual function.

He also observes that within a certain time range (~20-50 milliseconds), the attack is a function of the steepness of the global amplitude rise but is insensitive to fluctuation around the slope. While it is common experience that attack is a function of the rise time (say to 99/100 of the peak value) or of the slew rate or of the slope of the amplitude envelope, the issue of the latter part of the statement was never corroborated by other sources. Surely our knowledge of the attack transients continues to improve from Helmholtz's day. While high precession computer technology based analysis/synthesis has been available for a couple of decades now (after Schaeffer's work), it has not been a popular view that the overall amplitude envelope (*"amplitude global"* in Schaeffer's terminology) can be used for high quality analysis/synthesis. But recently, as we shall see in later chapters, the analysis/synthesis approach, derived from the dynamic theory of timbre this thesis will be devoted to, shows that the detailed variation of the amplitude envelope *does* seem to be important perceptually, contrary to Schaeffer's observation, one which is accomplished by techniques less precisely controlled than available on a digital

computer.

1.7.2.2 The Dynamic Role of the Attack in the Overall Perception of Timbre.

Concerning the conditions which determine the importance of attack in the overall perception of a sound, Schaeffer reports:

(1) For nonsustained tones, such as the piano or other percussive tones, the attack plays a decisive role.

(2) For sustained tones of average duration, the importance of the attack diminishes and the attention is on the evolution of the sound.

(3) For sustained sounds with vibrato, the role of the attack becomes rather negligible, as demonstrated by the strong identification of violin and oboe tones whose attacks have been cut.

The conditional importance of the attack in the overall perception of the sound based on what is also present in the sound certainly suggests the need of a dynamic description, especially in the event of dealing with the idea of timbre interpolation.

1.7.2.3 Distinctive Features of the Attack Timbre.

Schaeffer observes that the ear distinguishes two features in the sound quality of the attack: one, called the "attack timbre," is associated with the harmonic content, and the other, called the "steepness of attack," is associated with the dynamics. In general, Schaeffer speaks of the musical perception of attack as being correlated with the general dynamics of a sound, or the evolution of energy and with the harmonic contents.

1.7.2.4 Timbral Physical Correlates.

Finally Schaeffer makes a general observation about the physical correlates of timbre of a sound in general.

"Le timbre percu est une synthèse des variations de contenu harmonique et de l'évolution dynamique; en particulier, il est donné dès l'attaque lorsque le reste du son découle directement de cette attaque" (page 231). The significance of this statement is the observation of amplitude envelopes that he calls dynamic variations, pitch trajectory (or period trajectory) which he describes in terms of variation in harmonic content as in vibrato or pitch glide, and the harmonic contents themselves and separately as elements of timbre correlates.

In summary, Schaeffer's work is important both because of its forward looking

character, seeing the need to revise our attitude (aesthetic values), as well as our method toward sound processing, in particular timbre processing, as our sound universe continues to expand. Schaeffer is also ahead of his time in seeing a need for a dynamic description of timbre. It is noteworthy that Schaeffer also observes that although the amplitude envelope, harmonic contents, and pitch trajectory may vary independently, in fact steep attacks are found to go with rich high harmonics which sound bright and rich, and soft attacks go with low harmonics which sound weak and poor. This points to a certain "grammar" that the ear has been accustomed to hear in combination with these physical correlates. This in fact has been observed in the process of the author's attempt to bring an /a/ with a mild envelope to sound more like a marimba attack. The lack of high frequencies in the /a/ onset sounds hollow in a steep attack envelope.

Finally, it is also worthy of pointing out that Schaeffer does seem to have a dynamic notion of timbre when he points out the advantages of listening for the harmonic content of a sound played backwards. This is especially true when our perception seems to shut off after the attack despite a lingering decay, characteristic of a percussive sound. He also thinks that such a listening process may help to objectify the natural or original sound we want to comprehend because of its unnaturalness with respect to the source. However he was shocked by the experience. The problem is of course not just that we are not used to the sound played backward. It has to do with the mechanical filtering nature of the ear. Anyhow, the opportunity to listen to various parts of the sound still exists from a dnyamic theory of timbre viewpoint, which will be described in a later chapter.

Chapter II: Theory

2.0 Introduction.

As stated at the beginning of chapter I, one of our goals consists of finding a systematic description of timbre in terms of its distinctive features or physical correlates, for a wide class of musical timbres, such that the same description will enable us to describe the relationship among timbres and eventually lead us to describe the structure of timbre space. Since we hope that a language which describes the internals of timbres will also describe their external relationships, the approach we adopt to explore timbre space will be one that begins locally with pairs of timbre. We will explore the issue of timbre relations later.

At this point, we want to stress the importance of trying to discover the desired language from a perceptual foundation. That is, we want to derive the language from experimental evidence and reasoning based on the principles of psychoacoustics. Recall that Helmholtz tackled the problem of justifying Ohm's acoustical description of timbre by showing how the ear might respond to periodic waveforms. Starting from a system of damped resonators as a model for the ear, he proceeded to rationalize the study of the physical features in the sound that the ear, in his view, perceives. From these, the description of timbre is entirely based on the Fourier magnitude spectrum and the rules classifying timbre are based on physical feature extraction from the magnitude spectra of the class of timbre he subjectively defined. His approach to the treatment of timbre is summarized in figure 2.0 (a).

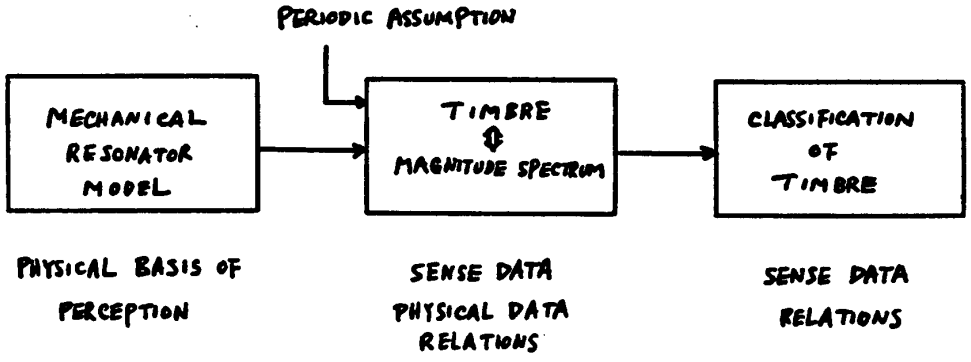


figure 2.0 (a)

We will follow Helmholtz's approach to some degree. We believe that Helm-

holtz's mechanical basis for timbre perception provides a reasonable point of departure. But we will make a stronger assumption about the ear, i.e., we take the view that the ear is an *active* observer, and see how Helmholtz's mechanical basis might provide a pattern recognition basis for perception. It is well known that the ear can adapt, organize, and extract features. Minsky, among others, has recently proposed that consciousness can arise from pure physical phenomena alone in some kind of societal basis. It is our goal to try to discover a reasonable path that leads from Helmholtz's passive mechanical model of perception to an active one that permits a systematic pattern organization, recognition, and extraction environment for the ear to do what we know it does for timbre. We will see that a space-time single input multiple output response (SIMOR) pattern of the damped harmonic oscillator type provides redundancy and coherence, as well as selectivity that forms the basis of a space-time distinctive feature extraction "program."

We will use arguments from information theory, pattern recognition theory, and physics, as well as psychoacoustic experiments to establish our claim. From the study of the perceptual foundations of timbre, we will show how distinctive timbral features and their physical correlates will emerge. And finally, we will show how a language can be formulated that would describe both the internal and the external dynamics of timbres.

The approach we use looks like Helmholtz's (figure 2.0 (a)) except the connections are now two-way (figure 2.0 (b)). This is partly because the scope of timbre is greatly expanded and partly because the only sources of information for a meaningful argument come from experiments on known perceptual relations.

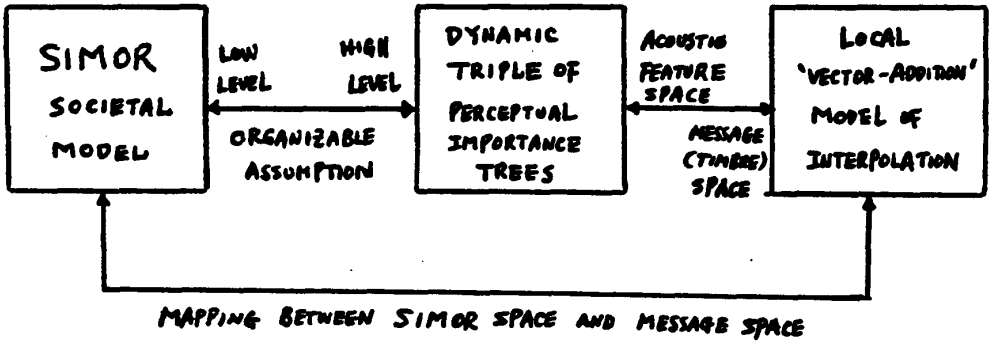


figure 2.0 (b)

We will also attempt to present a new description of timbre that will reduce to the more common description in limiting (special) cases.

2.1 Conceptual Fundamentals of Timbre.

This thesis is about timbre; it is an attempt to provide a more modern treatment of timbre. But one might ask, "what is timbre?" or "What do we mean by 'timbre' when we use the term?" It's a natural, reasonable question. For part of the human experience is to ask "What is this?" and "What is that?" Furthermore, a meaningful discussion of timbre requires some kind of understanding of the terminology essential to this discussion. This understanding will be gradually established as we explore the ways timbre has been understood, or the things the term has meant in practice.

In the first chapter, we assumed we had some idea of what the term denotes and we assumed that we had some knowledge about timbre from experience. But it is important to point out, even at the risk of being repetitive, that timbre is at once a very complex concept in the realm of perception and is at the same time very poorly understood [Schubert, private communication]. Therefore, we will take the approach mentioned above—i.e., gradually bringing forth ideas about timbre by considering what it has been held to mean, and thus arrive at an understanding of timbre that will lead to productive results. In due course, a new definition of timbre will emerge from this thesis that will be compatible with existing ones in limiting cases and at the same time broad enough to be consistent with the emerging notions of timbre we have reported about in Chapter I.

2.1.1 The "Force" Analogy—a Methodological Model.

At the time Newton discovered the law of gravitation, the notion of force was not well understood. Although it was commonly known (from experience) that applications of force resulted in motions, the notion of force was vague and confusing. People had no conception of instantaneous velocity let alone the rate of change of a velocity. It was not known that force was directly related to acceleration. The notion of momentum was not defined and its relation to an applied force was not understood. Everything in the science of mechanics that we know to-day was almost nonexistent.

The primary aim of mechanical studies at the time was to discover (to be able to predict) the dynamic relationships among the celestial bodies, i.e., the relative positions of these bodies as a function of time—to discover their orbits (trajectories).

It was therefore considered by some as unnecessary to discover the "cause" of these motions. We should point out that the positivists or presentationists who rejected causality were also supporters of Copernicus's ideas.

The reason for the loss of support for Ptolemy's idea of describing the observed orbits by systems of circles (even though the original of approximating an orbit by a circle seemed both reasonable and elegant) was largely based on the principle of economy. To the positivist, the kinematic description of Kepler sufficed but it was actually Newton's metaphysical or representational approach to the problem that provided a simple explanation for why Mars' orbit should be an ellipse instead of a circle or a system of circles as Ptolemy would have it. The cost of Newton's metaphysical approach is the additional machinery to support the causal relationships of his mechanical description, but it seems well justified. Although force was an intuitive idea in mechanics, it is not directly observable as position and time or displacement and duration are. When the law of gravitation states that planetary motions are due to the pulls among the celestial bodies and the pull or force between two bodies treated as point masses is proportional to the product of the masses and inversely proportional to the distance squared, force is not *a priori* defined. It is simply assumed that force (in particular, such a force as described above) exists and the idea is to find ways to determine it or measure it. The way Newton succeeded in doing that was of course through his laws of motion. In particular, Newton suggested that force is a function of the mass and acceleration of the body in motion to which the force had been applied. So if we could observe acceleration, then we could observe force. And in order to make measurements precise, Newton developed calculus so that we could measure acceleration from velocities and conversely, determine velocity from acceleration—with precision limited only by the instruments employed. From such a scaffolding, we see that the law of gravitation emerges as a meaningful statement about nature and at the same time, the notion of force, momentum, the rate of change of momentum, and so on, become clarified and well-defined.

One point is clear: it was not necessary for Newton to first labor with a definition of force before he set out to do what he did. Perhaps we can benefit from using the master's method as a model. And this is the reason we discussed the analogy above.

2.1.2 Notions of Timbre.

In order to start our discussion, we must begin somewhere. That is, we must assure ourselves that we understand each other when we talk about timbre even though we have assumed that we have some knowledge about timbre from experience. The way we will do this is to survey various existing notions and examine their strengths and weaknesses.

Some of these notions, e.g. in 2.1.2.1 and 2.1.2.2 are commonly shared without controversy. Others, like in 2.1.2.3, are a little controversial and may not have been articulated in the literature in such a manner, but we nevertheless believe to be reasonable and not exactly new. The notion in 2.1.2.4 is best known to have originated in Winckel's treatment of timbre, although the notion is generally accepted in the psychoacoustical field. The notions in 2.1.2.5 through 2.1.2.8 represent various forms of the spectral notion of timbre originated by Helmholtz (2.1.2.6) and refined by Moorer and Grey (2.1.2.7 and 2.1.2.8), or misapplied (2.1.2.5). The pattern recognition/feature extraction notion (2.1.2.9) is well known in speech perception and vision and is suggestive in Charbonneau's work based on Moorer and Grey's framework. In fact, Schaeffer stresses the role of distinctive feature extraction in timbre perception and uses linguistics as a model paradigm. The multidimensional attribute has been articulated by Plomp, Wessel, Grey, Slawson, and others, although not necessarily in consonance with our view.

2.1.2.1 Timbre as Identifier of Musical Instrument.

To someone who has never heard of the term timbre before, the easiest way to explaim to him or her what the term describes is to say something like: "Are you familiar with the sound of a violin? Are you familiar with the sound of a piano? O.K. Can you tell that the sounds of these two instruments are quite different? O.K. The qualitative difference between these two kinds of sounds as we hear it is referred to as a difference in timbre and the sound quality of each instrument that gives rise to this difference is referred to as timbre." Essentially, in using this description, we are saying:

(1) For each instrument, there is a timbre associated with that instrument so that the timbre of a piano for example is the sound quality associated with, or characteristic of, the piano.

(2) Timbre is an *identity* of a musical instrument; that is, it is different

from one kind of instrument to another.

(3) Timbre is a musical concept.

The advantage of this notion is that from experience it is easy to understand; therefore, it is a simple way to convey the notion of timbre to those who are otherwise unfamiliar with it. But there are many disadvantages:

(1) Although timbre arises most naturally in connection with musical instruments, they are not the only place, nor even the most common place, where it occurs. In fact, speech timbre is actually closer to home, although it is harder to explain to a novice because of possible confusion he or she might have in distinguishing the role timbre plays in speaker identification and in linguistic (phonetic) functions, such as the discrimination between the vowel sounds *ah* (/a/) and *uh* (/∧/), for example. On the one hand, different speakers have different voice timbres but on the other hand, all speakers of a given language share the same speech timbres in saying all the vowels and consonants in order to communicate.

(2) The concept of timbre is not necessarily tied to music. Historically, timbre served useful functions for information exchange long before music played a part in human activities.

(3) Even within a single musical instrument, timbre changes significantly over the entire pitch range. It also changes over the range of loudness, manner of articulation, etc.

Therefore, identification of musical instruments is not a very precise way to describe timbre.

There are many more fundamental shortcomings that are shared by other notions of timbre and will be discussed momentarily.

2.1.2.2 Timbre as Identifier of the Source.

This notion is similar to that of identification of a musical instrument. The scope is now broadened to include any physical object that is capable of generating sounds by virtue of its response to some form of excitation of physical origin. Its advantages are:

(1) Timbre is no longer tied to the notion of musicality so that noise, speech, or any sounds, ethological or synthetic, can be understood to have timbre and we can appreciate what it is by extrapolating from what we

have understood in terms of timbres of musical instruments.

(2) Timbral variations within a musical instrument can be understood because each device or manner that generates a particular tone is by itself a separate source from those of the rest of the tones. It is a matter of understanding the source characteristics of individual devices in order to understand the timbral variations.

(3) Historically, the ear must have served the function of source identification for the purpose of survival. To know your friends and enemies and to know your food sources (prey) from environmental noise.

The disadvantages are:

(1) We still don't understand how speech timbre works since it is independent of the speaker (although species-dependent).

(2) We have not seen how timbre is related to the receiver, i.e., the ear. Is it possible that we can understand timbre by considering merely the source alone?

(3) As an extension of (2), it is not clear how computer generated sounds different from those of acoustic sources are actually perceived. As a consequence, a notion of timbre detached from the notion of the receiver seems ill-equipped to exploit the potential that digital synthesis of sounds seems to offer.

(4) A notion of timbre based on source identification is inherently limited in permitting the use of timbre as a structural or organizing element in musical composition because the best one can do is to organize the music around acoustic instrument families. Again, the notion of acoustic instrument families defeats the advantages digital synthesis of sound offers. Furthermore, careful listening shows that timbre is hardly so simple as being only instrument family oriented, and it contradicts the great success people have achieved in developing spoken language as a means to communicate with one another.

2.1.2.3 Timbre as Image of Sound.

This idea tries to separate timbre from the source and hence source identification.

The advantage of this notion is to make clear that sound, not the source of

excitation of the sound, is received and therefore perceived by the ear.

(1) First of all, we recognize that the "path" between the source and the ear in a typical listening environment is variable and complicated. Imagine the long sustained "tutti" $E\flat$ major chord with which Beethoven introduces his listeners to his *Fifth Piano Concerto* without the reverberant characteristics of the transmission environment. We can almost be sure that the master does intend to convey a sound of very dramatic and majestic quality because it is repeated two more times over a span of at least a couple of minutes sustained in between by a sort of variation on the same effect by the piano with sustained pedalled arpeggio and other torrential runs. The impact is clearly made on more than a few listeners, for the opening chord mentioned above has become the signature (the most memorable and recallable element) of the long piece. It has attained an importance for the piece analogous to the attack timbre of a percussive instrument tone for a single sound event. But if the impact is so great for the listeners, it is equally important to notice the frustration and disappointment they experience when the room acoustics are deficient in projecting that rich acoustical image the master had in mind for the listener.

Although it is useful to draw a distinction between chord and timbre in certain situations, and although the example being discussed certainly involves, from a sound generation point of view, a chord and literally dozens of sources, the effect is hardly dissociable from that of timbre from the ear's point of view. In our example, the harmonic function is not to be fulfilled for a length of time on the order of a minute in real time. And the ear certainly does not try to differentiate the violas from the cellos and the G's from the $E\flat$'s above and $E\flat$'s below. Finally, the string chorus effect combined with the coloration and reverberation of the acoustic environment gives rise to that majestic roaring timbre we find in the performance. In other words, the timbre we hear has lost much of the identifying characteristics of the source under the combination of these factors. We notice that the perceptual difference (i.e., from the ear's point of view) between the timbre of a single source and the timbre of multiple sources playing a single note or chord can be blurry when room acoustics work in a way

(which is usually the case) to smear the images of these sources. An attempt to insist on a distinction between timbre and chord is artificial and serves no useful purpose for timbre modification (including interpolation of timbre).

(2) The notion of timbre as the perceptual image of a sound is equally important in understanding the timbre of a rich gong sound such as those from the Javanese gamelan. Obviously, the timbre is significantly altered when we damp the vibration and it depends on when we do the damping also. Therefore, one might ask which timbre corresponds to the gong? Naturally, the sound as we hear it is the result of interactions between the air vibrations (originated from the gong vibrations) and the gong vibration itself. That is, some kind of acoustic feedback effect plays a critical role in what we actually hear. The normal modes come and go separately at their own times. It is unnatural and impossible to isolate their interaction and it is equally unnatural to describe the gong timbre by looking at the spectrum over some arbitrary window of time. In short, the underlying element of timbre is most simply and naturally the acoustic waveform.

(3) Then of course sound is the underlying element for speech perception. It is so because the speech we understand is independent of who speaks it. In other words, the timbral characteristics that make speech communicable go beyond the perception of source characteristics. Although we know that the ear is a very versatile message extractor in the sense that it will discard from the acoustic waveform a considerable amount of data which it regards as noise or irrelevant information, it remains true that it is some crude form of invariant acoustic features in the sound that the ear actually must wrestle with. This is in fact the case because if the noise interference is too severe, i.e., when these acoustic features become significantly degraded by noise, the ear will not hear (in the sense of extraction) the message.

(4) Finally, the successes of digital music synthesis are often described in terms of how our ear can be "fooled." The logic behind this description is of course the notion that the timbre we perceive depends *only* on the sound that enters the ear and nothing else.

But what we still need is a description of how the ear processes the sound and

finally a description of the acoustic features that are translated into the percept of timbre.

2.1.2.4 Timbre as Message—Information Theoretic View.

The idea of timbre as message emphasizes the importance of the ear historically as a message decoder both in speech and in source identification in the chain of communication:

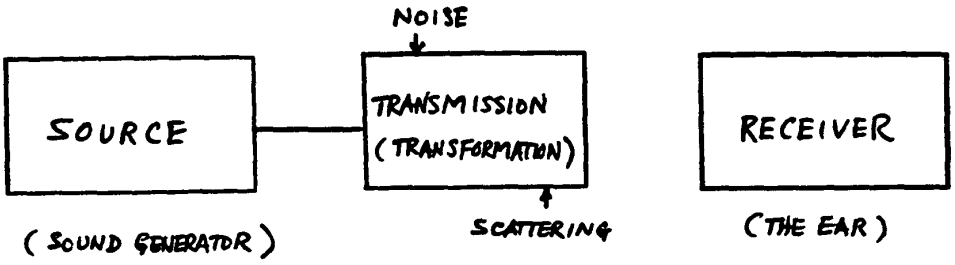


figure 2.1.2.4

The role of transmission is important because of noise interference. It implies that the ear must have learned through stages of evolution to extract important or distinctive features in the messages (from the source) in order to survive as a functioning element. As a result, we expect the ear to process the sound waves received with intelligence, in the sense that it tries to adapt, to organize, to update, and to discard depending on its current knowledge of the state of the communication affair.

But how does the ear do all these? At this point we still lack an understanding of how sound is perceived and what aspects of a sound are decoded into the percept of timbre. In other words, it is still unclear how timbre is perceived separately from pitch, loudness, and duration, for a single sound and what physical aspects of a waveform contribute to timbre. Later, we will see how the complex membrane response in the cochlea might provide a basis for these intelligent activities.

2.1.2.5 The Ear as Fourier Analyzer—Timbre as Spectrum.

The knowledge that the ear exhibits strong frequency selectivity across the basilar membrane, or behaves as a bank of resonators, leads many to believe that the ear behaves as a Fourier spectrum analyzer. As a result it is believed that a signal is "analyzed" into its Fourier spectrum inside the ear and timbre is the ear-brain's interpretation of that spectrum. The advantage of this idea over previous ones is that the mechanics of audition finally plays a role in our attempt to understand timbre. Or more precisely, the Fourier idea is intended to provide an

understanding of what it is in the acoustic signal that the ear "perceives" as timbre based on our knowledge of the mechanics of auditory perception. Furthermore, casual observation suggests that timbre is a function of signal spectrum in the sense that an acoustic signal with a different spectrum generally elicits the sensation of a different timbre. However the shortcomings of this notion of timbre are fairly obvious once one listens more carefully. Sound is a dynamic process and the ear has an auditory window that can pick up temporal changes readily. But this spectral notion has absolutely no room for the time arrow. As a result, it cannot explain why a waveform played backwards should have a different timbre (or even a different sequence of timbres). Note that the Fourier transform of a waveform differs from that of the same waveform played backwards only by an overall phase* that does not affect the magnitude spectrum. Similarly, the spectrum idea does not explain why we hear a succession of well-defined timbres for a succession of notes (played by the same instrument or by different instruments) if time plays no role at all in our perception. [Note that the Fourier transform is performed over all times from the infinite past to the infinite future.] While infinities are mathematical idealizations, we cannot conveniently define them to be the beginning and the end of a note. The reason is the ear is a physical device and it has its own real times independent of the notions of time in the senses of J.B.J. Fourier or N. Wiener—or in the senses of Beethoven or Debussy. On the mechanical level, the ear's time is a function of the physical material of the basilar membrane. In fact, there is a collection of such times. We will see later how the brain can make sense of such times. But it suffices for now to stress the importance of time as perceived by the ear.

2.1.2.6 Musical Timbre as Fourier Magnitude Spectrum.

The notion of musical timbre as Fourier magnitude spectrum originated with Ohm and Helmholtz. This notion (in terms of a Fourier series where the signal is periodic) differs from the spectral one of 2.1.2.5 (in terms of a Fourier integral where the class of signals is not necessarily periodic) in a fundamental way. This notion limits the class of timbre to one corresponding to periodic acoustic signals. This

* If this sign-change is important, how can one use it to measure the different perceptual impacts which different sounds have when played backwards (e.g., a percussive sound, such as the marimba has a very dramatic change, but a sustained sound, such as the /a/ does not)?

restriction means that the "short-time" nature of the auditory window is not violated because the ear needs only to analyze one period of the signal and then "rest" forever. The result of the analysis is of the discrete-type characteric of periodic signals in contrast with that of the continuous-type characteric of Fourier transforms of arbitrary signals. The most significant aspect of this model is its data reduction implication, namely that while each filter output or resonator response is a time function, only two real numbers are needed to specify the entire function (compare with $2T \cdot (f + \Delta f)$ numbers dictated by the Shannon-Hartley sampling theorem, where T is the duration, f is the resonance frequency, and Δf is the "bandwidth" of the filter related to the reciprocal of the damping constant—note that the bandpass version of the sampling theorem does not apply here because it refers to an ideal bandpass filter with a symmetric filter shape but the resonator's filter shape is asymmetric and "passes" low frequencies). Therefore, by assuming that only periodic signals please the ear and therefore are perceived as musical, Helmholtz and Ohm articulate the issue of timbre perception as one of data reduction and feature extraction. Of course, they went further by halving still the number of dimensions in the control parameter space by dropping the role of the phase function. The disadvantages are (1) the unrealistic limit the model places upon the universe of musical timbre, (2) its silence about how we perceive "non-musical" timbre, (3) the phase-deaf hypothesis is found to be invalid even for the class of periodic signals [Schroeder and Mehrgardt, 1982], and (4) higher spectral lines are not well resolved because of the logarithmic nature of the membrane response.

2.1.2.7 Timbre as Short-Time Fourier Spectrum.

The notion of timbre as the way the ear interprets the time-varying spectrum of the acoustic signal through the "short-time" auditory window takes into account the short-time "tracking" nature of the signal and the non-stationary nature of the ear's mechanical response. However it is not without shortcomings because:

(1) A Fourier interpretation is awkward whenever the window of stationarity T_s in the signal is short compared with the Fourier resolution window τ ($\tau = 1/\Delta f$ where Δf is the frequency resolution of the Fourier analysis), and

(2) the membrane response is physically much closer to that of a system of responses from a collection of damped harmonic oscillators each with

its own damping constant, and hence its own "auditory window" than a "short-time" Fourier analyzer where one must face the choice between adopting a linear orthogonal frequency scale (with constant auditory window) of a Fourier transformation or setting a logarithmic non-orthogonal frequency scale (with variable length window). Note that orthogonality in a DFT does not have an obvious parallel in the case of *windowed* continuous Fourier transforms. The logarithmic frequency transform is known as a Mellin transform [Bracewell, 1978]. Even with the Mellin transform where the frequency dependent auditory window is taken care of, the dynamical nature of the resonator response is such that (i) the response always consists of a transient response and a steady-state response, (ii) there is always some degree of response in every resonator no matter what the stimulus frequency is (one can see this by examining the particular or inhomogeneous solution of the differential equation describing the dynamic system), and (iii) we really don't know the auditory window time constants well enough to give precise meaning to such a Mellin transform.

(3) The short-time Fourier transform interpretation does not address how the ear abstracts relevant information or distinctive features from such a vast array of wave patterns (output time function of the filter bank) that the communication model strongly suggests the ear actually does.

2.1.2.8 Timbre as Time-Varying Amplitude Functions of the Harmonics.

This approach is a variant of the one described in 2.1.2.7 but attempts to avoid the disadvantage mentioned in 2.1.2.7 (3). In other words, an active ingredient is introduced into the auditory processing in the form of distinctive feature extraction and data reduction. But again,

(1) It shares the disadvantages (1) and (2) of 2.1.2.7. In fact, the notion of harmonics does not apply to all timbres of musical or speech interest. This is especially true when the signal is highly non-stationary in which energy concentrated in harmonics is an artifact of discrete mathematics. Literal interpretation would contradict the time-frequency uncertainty product law.

(2) There is a significant body of evidence that many such independent harmonics do not have individual perceptual significance. Rather, their

collective behavior does. See Chapter I.

Therefore, while this notion represents a significant advance in our understanding of timbre, it is not a complete description, especially not in terms of feature abstraction and organization. Note the distinction between data reduction in terms of what the ear discards as irrelevant information (as Grey addresses) and data organization in terms of transformations of data patterns into more coherent forms (as Charbonneau later addresses).

2.1.2.9 Timbre as Patterns and Collection of Distinctive Features.

An example of timbre as patterns can be found in Charbonneau's study as explained in 1.5.7. It represents a significant advance in our understanding of timbre in terms of distinctive feature abstraction. The coherence that binds the multitude of amplitude envelopes, the frequency trajectories and the onset times is actually "heard" by the ear. The fact that major identification characteristics remain, as physical data is degraded, under the constraint of these coherence factors, suggests not only that the ear actually organizes the acoustic data, and abstracts distinctive features, but also does it hierarchically so that destruction of data has a major or minor effect on perception of timbre depending on whether the major or minor features are destroyed.

A typical pattern recognition approach involves further the abstraction, classification, and reduction of data by the auditory processor. These pattern recognition characteristics have been well recognized in linguistic studies. So it is plausible to think of timbre as a list of distinctive features which the auditory processor extracts. But what are these distinctive features? In speech, the most well known is the combination of formant locations, or the distribution of formant energies. Lesser known but gaining increasing recognition are the transition trajectories from phoneme to phoneme when continuous speech becomes the focus of study. Musical timbre in most cases behaves like continuous speech. Therefore, timbre studies with a scope that includes perceptual study of continuous speech seem to make a lot of sense. The questions are (1) what are the features, and (2) how do we arrive at these distinctive features from the resonator response functions or directly from the waveform?

However, while the pattern recognition idea of timbre as implied by Charbonneau's study represents a significant advance, it is not quite a complete description

for the following reasons.

(1) We do not yet have the underlying mechanism to distinguish these patterns from those of the visual kind. Their temporal character is not evident; the kinematic nature of acoustic or fluid motion is not evident.

(2) It does not articulate possible dynamic relationships inherent in the way different sounds are turned on and off and fluctuate and are organized differently, which the "short-time" auditory window picks up.

(3) Charbonneau's study is inherently limited by the limitations of short-time Fourier analysis as listed in 2.1.2.7.

2.1.2.10 Timbre as a Multidimensional Attribute.

From the viewpoint of arranging timbres into possible perceptually ordered sequences, our experience shows that timbre is a multidimensional attribute in the sense that we cannot arrange all timbre into one single perceptually ordered sequence. From the viewpoint of the underlying waveform, we know there are mathematically N possible dimensions from an N-sample waveform. And since modification of the amplitudes of the sample sequence usually leads to different timbres, we expect potentially a high dimensionality in the timbre space.

In other words, we expect the multidimensional nature of the control space of timbre to induce a multidimensional timbre space because of timbral sensitivity to change in the waveform, although the dimensionality may differ. One can arrive at similar qualitative conclusions from Helmholtz's Fourier magnitude spectrum although the control dimensionality is now much smaller as a result of periodicity. If f_s is the sampling frequency and f_0 is the frequency in Hertz, then

$$\frac{f_s}{2f_0}$$

represents the number of harmonics that can fit into the frequency range, and therefore specifies the dimensionality in the control space and puts a bound on the dimensionality of the timbre space.

Plomp, using a statistical data representation technique known as multidimensional scaling (MDS) [Kruskal, 1964], determines that a dimensionality of four is sufficient to reasonably represent his data—taken from the class of stationary acoustic stimulus. It is not clear that this is actually the dimensionality. Unlike MDS studies of color, where the dimensionality of two correlates with the empirical observation that all colors can be determined by the *relative* strengths of the

three primary colors, Plomp's result cannot be independently confirmed. (Note that Plomp used non-metric MDS.)

We have similar difficulty with the harmonic amplitude function representation of nonstationary stimuli. In his MDS study, Grey reports that a three dimensional (Euclidean) space is sufficient to represent his sixteen data points. In this case, it is very unclear how the huge dimensionality (much larger than that of Plomp) in the control space is transformed by the auditory processor to such a small number in timbre space.

In fact, we pause here to briefly discuss dimensionality. First, we ask, what is the purpose or what is the scope of the estimation based on a given set of stimulus data?

(1) One may want to determine the dimensionality or the irreducible set of dimensions of the perceptual space all timbres live in. This is in fact extremely ambitious and quite beyond the means of our knowledge at present. From mathematical considerations (see 2.5.4), dimensionality is a local concept, especially for a space whose topological properties we know so little about. Therefore, with a finite sample size, it is quite meaningless to estimate the "true" dimensionality of timbre space.

(2) One may want to define the local dimensionality of the known samples. The problem is that we cannot be sure that these timbres are in fact objects or points in a neighborhood for which a single meaningful dimensionality can be assigned. If the data set indeed belongs to a neighborhood of a single dimensionality, then one cannot make any statement about the dimensionality of a data set which includes new samples. So we must recognize the limited application of efforts in dimensionality determination from current experimental methods.

(3) We can think of a typical dimensionality and a distribution of dimensionalities in the space and find a set of randomly selected samples for the estimation of the parameter of dimensionality.* Such a job is formidable

* Timbre space and its dimensions are necessarily random variables on the basis of individual (person to person) differences alone. But even within an individual listener, the complex variation in the control parameter space among timbres, even within a fairly local region, implies that even any measurement of local timbre space

because it depends on the ability to select a truly "random" sample of timbres, and a large one for statistical reliability and meaningfulness. The set Grey chose was of course highly biased. But even if we follow (2) and try to determine the dimensionality of the set based on an assumption of uniform behavior in a local timbre space, we must assume that they are more or less independent identically distributed (iid) samples near the point whose local dimensionality we want. Then we must address the sample size. A crude estimation of reliability can be obtained using the assumption that the iid samples have a spread of behavior proportional to the square root of the size of the sample, so that for small sample sizes, the reliability is small.

Therefore Grey's MDS study must be viewed first as an attempt to scale the set of timbres chosen, not an attempt to scale the broad timbre space this set and other timbres live in. Second, the sample size of sixteen timbre points is statistically rather small and is therefore not sufficiently reliable for making any meaningful assertions about the dimensionality of timbre space. After all, one certainly does not think that one is able to regenerate these sixteen timbres with information derived from the three dimensions described by (1) "the physical property of *spectral energy distribution*," (2) "the physical existence of *low-amplitude, high-frequency precedent*

---

or its dimensionality must behave like a random variable.

In other words, if timbre T_1 is said to live in a local space of dimensionality d_{T_1} (note that dimension refers to "internal" dimensional, representing the perceptual features of the timbres in the local space), then d_{T_1} can be thought of as a random variable. If timbres $\{T_k\}_{k=1}^n$ are members of a local timbre space, then each d_{T_k} can be thought of as a random variable with the same mean μ and variance σ^2. Their average

$$d_{\overline{T}} = \frac{1}{n}\sum_{k=1}^{n} d_{T_k}$$

will have the same mean μ but its variance will be σ^2/n (see, for example, [Hoel *et al.*, 1971]), provided that the deviations of the d_{T_k} from their mean μ are uncorrelated. Thus, sampling theory implies that the reliability of the estimates of the local dimensionality from a set of timbres in the local space should be inversely proportional to the number of timbres used for the estimation.

energy," and (3) "*synchronicity* in the attack-decay behavior of upper harmonics" (emphasis in original, [Grey, 1975, p. 99]). Third, if the dimensions mentioned are not meant for the regeneration of the timbres under study, then they must be viewed as merely the *more important* dimensions along which meaningful relationships among the timbre points are obvious. Actually, however, the decision to go for a low dimensionality representation precluded the possibility of finding other dimensions along which additional meaningful relationships among the timbres might exist. Fourth, the study was probably meant to find the minimum dimensionality for an acceptable visual representation of the data, acceptable based partly on the particular technique used and partly on the prevalent view of how the data should look (e.g., instruments of the same family would be expected to cluster together). In fact, Grey dealt at length with the implausibility of a two-dimensional representation. Fifth, however, we may have a better understanding of the three dimensions referred to above if we assume that the perceptual space is hierarchically organized and therefore that these dimensions are the more prominent ones from which lesser ones sprout. In 2.4 and 2.5, we will consider a model for the hierarchical organization of acoustic features which are important for the perception of timbre.

One reason we can argue for a hierarchical dimensionality in a pattern recognition problem is the tendency for the observer to bring subfeatures together by some organizing element to form features and in turn bring features together to form superfeatures by yet some new organizing element in the current level of organization. This approach of course cannot specify the nature of the dimensions either. Therefore it is apparent that while the multidimensional character of timbre is well-accepted, our knowledge of its nature remains insufficient. Still another point to consider is that because of the time-varying nature of the Fourier magnitude spectrum, the control space dimensionality is greatly increased by the data points needed to specify each function. We therefore expect a corresponding increase in the dimensionality of the timbre space. Notice that the class of non-stationary signals calls for a much higher dimensional space than the class of stationary signals.

This is the same with pitch and loudness when they become dynamic attributes, i.e., functions of time. But, unlike pitch and loudness, the multidimensional nature of timbre remains even if we have a stationary signal. This is so because timbre is a perceptual measure of change, especially local change in the acoustic signal.

Therefore, even with a stationary signal, we still have a time function within each repeatable period. Therefore it is not surprising that timbre is a multidimensional attribute. From the spectral standpoint, the changes or vibration patterns manifest themselves in the energy distribution as a function of frequency (or place along the basilar membrane). The many degrees of freedom in the frequency (or place) dimension, i.e., the variation of amplitude as a function of frequency (or place), contributes to the multidimensional character of the control parameter space for timbre generation which in turn apparently leads to a multidimensional perceptual space despite all the organization and feature abstraction that goes on through the hierarchy of perception.

But what is a dimension in the perceptual space beyond the ordering notion? Is the dimensionality even an integer (as opposed to a fraction such as $\frac{118}{7}$)? If the dimensionality is integral, then how does one dimension relate to another? What is the physical correlate for each of these dimensions? How does the dimensionality change from region to region (or even point to point)? Can we talk about independence or orthogonality? At present, our knowledge does not show us how to do this, other than through some statistical approaches that involve the correlation measure and the assumption that perceptual space is Euclidean n-space ($\mathbf{R}^n$). We shall see later that the assumption that perceptual space is $\mathbf{R}^n$ may not be compatible with the hierarchical nature of the way the auditory processor organizes data. If the acoustic features are actually hierarchically organized and extracted, then statistical error may lead to the loss of quite a few features or dimensions of lesser importance. And the notion of angle necessary to establish orthogonality of dimensions is at best ill-defined when the statistical error is high. In 2.5.4 we will attempt to address some of these issues in somewhat greater depth.

2.1.3 A Relativistic versus an Absolute Description of Timbre.

2.1.3.1 The ASA Definition.

The American Standards Association (ASA) definition is an example of the relativistic notion of timbre. It states that a timbre is that (auditory) percept that makes it possible for one to judge two sounds of the same pitch, loudness, and duration to be dissimilar. The advantage it has over a definition based on source identification is clear. First, as we have seen, the immediate object underlying the percept is sound or acoustic vibration, not the source (although as we have also seen, the manner in which we perceive and make sense of a sound depends to a significant extent on source identification in many instances), and usually the source image is perturbed by noise and other transmission characteristics, such as echo, reverberation, spatial effects, velocity effects, etc. Second, the fact that we can communicate (send and receive information) among ourselves through combinations of a set of phonemes, regardless of loudness, pitch, stress, intonation, and the manner of articulation, demonstrates that information encoded in that part of timbre which includes these characteristics that distinguish a vowel from a consonant, a stop from a fricative, etc., is quite independent of the source. On the other hand, since we can distinguish the vowel /a/ from other vowels over a range of pitch, loudness, and manner of articulation, it makes sense to learn the vowel quality of the /a/ by comparing it with others, holding constant those variables, such as pitch, loudness, and manner of articulation, that are not part of the coding alphabet (the phoneme set). Similarly, it makes sense to play a violin tone against a piano tone at the same pitch, loudness, and duration to bring out "the" timbre of a violin, i.e., to heighten the salient features of that timbre as opposed to the piano timbre. It is on this basis that the ASA definition is important as an introduction to the important and complex notion of timbre.

2.1.3.2 Weaknesses of a Relativistic Definition.

However, limitations in the definition are well recognized.

(1) *Universal Applicability?* What happens to the timbres of two sounds with great differences in pitch and loudness (e.g., a low fortissimo piano tone and a high pianissimo violin tone)? The definition does not suggest a way to compare them because by "equalizing" the pitch and loudness,

we certainly will have an entirely different pair of timbres for comparison.

(2) *Confusion Between Control and Description.* It seems that from empirical evidence, "equalization" is necessary only when the timbres to be compared are too close to be possibly interfered with by other perceptual features. Thus for (experimental) *control* reasons, we want to set those others constant. But if timbre is clearly observed, such a control strategy should not be necessary, and may not be desirable, or even possible as we have pointed out in (1).

(3) *Lack of Independent Description or Absolute Meaning.* Does it suggest that one must find (synthesize) another sound of a certain pitch and loudness before we can identify the timbre of the sound, say that of a low piano tone? Surely, a typical ear would notice a timbre associated with that low tone. And what about the timbre of a sound without a well-defined pitch? How does one go about synthesizing a reference sound with such an impossible pitch requirement? Surely, the ear has been trained through stages of evolution to recognize timbres to answer the questions "What is that?", "Is it predator?", "Is it prey?", or "Is it just noise?" Surely the ear never had (to search over the memory) to find a second timbre to know whether a timbre just heard is *danger* or not. In other words, timbre should have, from our experience, an independent existence. It should have an absolute meaning.

2.1.3.3 The Pedagogical Merit of a Relativistic Definition.

Having discussed the limitations of the ASA definition, we do recognize the pedagogical value in the relativistic approach. From an epistemological viewpoint, we cannot effectively learn something without a context. Context provides us with the means for comparison, for adaptation, expansion, and growth. For example, for someone who has never learned the structure and organization of a digital computer, explaining it initially as a calculating machine and comparing it to an abacus would be appropriate as something of an introduction. To the extent that we understand sounds through the elasticity of our hearing process, the relativistic approach is pedagogically significant.

2.1.3.4 The Need for an Absolute Definition.

But the abacus analogy for a computer is not adequate because it does not explain many distinctive features of a digital computer. For example, it does not give any hint to the notion of a stored program, one of the most distinctive features of a modern digital computer. And in the end, it takes a hierarchy of distinguishing features to be enumerated in order to completely describe it. And just as we need an absolute notion of a computer, we ultimately need an absolute notion of timbre, i.e., one that involves enumerating a list of all of its common features, perhaps hierarchically ordered, with the most distinguishing ones first, although we might want to start from a relativistic approach for pedagogical or expository purposes. By the end of this thesis, we will have such an absolute notion of timbre, at least in a rough sketch.

2.1.3.5 The Value of a Relativistic Approach to the Study of Timbre Space.

The relativistic notion of timbre actually has a broader meaning than pedagogy. Where timbre is characterized by an absolute description, there is value in a relativistic approach to the study of timbre space and even the internal structure of timbre.

First, it is known that the relativistic approach provides a better understanding of the relational nature of speech timbres and their functions. For example, perception of leading consonants in continuous speech is a function of the transitions of these consonants into the vowels that follow. The relational character of speech timbre as exhibited in these transitions can be brought out clearly by a relativistic approach. Furthermore, we must understand the relationships among timbres (in the sense of being able to predict our response to a given selection of timbres and their arrangement, for example) *if* we will ever succeed in delineating the musical timbre space and creating musical timbre composition. This is so because of the dynamic character inherent in any timbre of a sound which has a beginning and an end. We will discuss this notion in 2.1.6. Notice that we may not have to equalize pitch or other attributes for control reasons if we can specifically name the distinctive features contained in an absolute description for comparison.

It is therefore our goal to strive for development of an absolute notion of timbre which is at the same time capable of articulating timbral relationships.

2.1.4 Timbre as a Distinctive Perceptual Feature and its Relation to other Perceptual Features.

2.1.4.1 Timbre as a Distinctive Percept.

The notion of timbre as a dominant perceptual feature seems well accepted. Our experience shows that as long as a sound has enough energy to be heard, we hear a timbre, i.e., we associate a timbre to it. In many cases, a sound's timbre is almost like the sound's identity, in the sense that other perceptual features, such as duration, loudness, and pitch, seem almost unnecessary for a complete specification of the sound from a perceptual viewpoint. In other words, timbre occupies a unique place in perceptual space as the major information carrier. But what are the roles of other perceptual features?

2.1.4.2 Timbre as One of the *Distinct* Auditory Percepts.

Whenever a sound impinges upon our consciousness, so do the percepts of duration and loudness, and in many cases, pitch, in addition to timbre. They are usually considered distinctive features because of their ready identifiability. They are considered distinct auditory percepts, in the sense that they are outputs of different (simultaneous) detection mechanisms. But are they distinct in the sense that each is a necessary dimension which together uniquely specify a sound from a perceptual viewpoint?

Before we attempt to answer this question, we will discuss the differences between timbre and the other percepts.

2.1.4.3 Timbre is not a Simple Percept.

In contrast to timbre, pitch, loudness, and duration are simple percepts, or perceptual primitives. When we speak of pitch, it is understood to convey some sort of highness notion in the auditory sense. Even a naive listener can articulate pitch differences between two sounds (if the sounds have pitch). Similarly, how loud or how long a sound is is intuitively understood. As a result, an experimental subject has a well-defined task in his or her own hand. These perceptual primitives are analogous to those in vision, such as size, height, and width. The latter are also intuitive concepts or perceptual primitives which need no elaborate explanation or definition. Such is not the case with timbre. Neither is the notion of the look (appearance) of a geometric object in vision. Both are very difficult to describe or define. Clearly, both are "multidimensional" attributes in their respective perceptual domains. But

our difficulty seems to go beyond their multidimensional nature. For if it were *not* the case, we could just articulate each attribute by enumerating their respective lists of features or dimensions. At least part of the problem is our ignorance of these "dimensions." But our ignorance seems to go deeper. Namely, we don't even know how the sensory processors for these percepts organize the complex patterns excited by their respective stimuli. It appears that within each sensory domain, excitation by a stimulus is reflected by a multitude of patterns, some of which turn out to be intuitively simply without the need of careful articulation. These are the notions of hot and cold, tall and short, high and low, loud and soft, and the like. But apparently, there remain features that might appear to be organized and simple to detect by the auditory processor or other sensory organs but do not have a simple notion in the thought processes that control our verbal communication skills.

2.1.4.4 Lack of an Adequate Description of Timbral Complexity.

This phenomenon is nothing new, geometry and other mathematical disciplines are those other languages that complement our verbal languages. Their existence stems entirely from the limitation of our verbal ability in expressing our relation with the physical universe we live in. For example, calculus was invented to articulate the laws of motion. In view of the mechanical nature of the ear and the demonstrated spatio-temporal response patterns of the membrane, it is very plausible that the ear's innate language is geometric, or rather kinematic, considering the temporal factor. Anyhow, it should be quite different from that of our verbal language. And therefore it is not surprising that timbre and look are such complex percepts to wrestle with.

2.1.4.5 Timbre is not Necessarily Independent of Other Percepts.

Returning to the question of whether timbre is distinct from pitch, loudness, and duration in the sense of independence in specifying uniquely a sound, we observe that the latter three percepts seem to be quite independent of each other at least in terms of statistical correlation. But is it always necessary to specify duration in order to uniquely specify the sound? Similarly, is it always necessary to specify pitch and loudness in order to uniquely specify the sound?

It seems that if we hear a stationary acoustic stimulus, then it is almost certainly necessary to specify the perceived duration to complete the specification of the sound perceived. For otherwise, there is no other means to distinguish this

sound from another sound consisting of the same vibration pattern except for a difference in physical length. But if the waveform is non-stationary, it can be distinguished from another by specifying the timbre difference. In other words, each sound can be completely specified without the duration description which has been absorbed as part of the timbre description. In this case, perceived duration is a redundant feature in so far as specification of the sound goes. Similarly, if two sounds are distinguished by their pitch difference only, then each sound must be specified by its pitch in addition to other attributes. But if the timbres are also different while loudness and duration remain the same, and if no two pitches correspond to the same timbre, then the timbre information is rich enough to render the pitch information redundant. The reasoning is that a sound with a pitched timbre would have some pitch for that particular timber by definition and that pitch can be completely defined by a complete description of the timbre. The only time timbre is not sufficient to describe a sound is when the same timbre has several possible different pitches. On the other hand, for the nondegenerate cases, specifying the pitch will not suffice to specify the sound because pitch is not rich enough to contain information of the timbre the pitch is related to.

Consider the timbre of a piano tone two octaves below middle C with that of one two octaves above middle C. Is it possible to obtain their individual timbres without simultaneously affecting their respective pitches? In other words, is it possible to obtain their individual timbres with some other pitch than the pitch that would produce the same timbres? We in fact answer this by asserting that there is no pitch such that there are two tones A and B of the same pitch with A having the timbre of the low piano tone and B the timbre of the high piano tone. In other words, it is not possible to obtain their individual timbres with some other (common) pitch.

We can similarly apply the same argument to loudness with respect to timbre. The only occasion where loudness is an indispensable variable in describing a sound is when timbre remains the same under different loudnesses. When timbres are different, the variables necessary to describe timbre may contain loudness also. For example, if you hit the piano middle C fortissimo and play a pianissimo middle C, there is no loudness level such that there are two tones of that loudness level with one having the timbre of the soft middle C and one having the timbre of the

loud middle C. Another way to look at it is to note that the ear will be able to distinguish between a pianissimo middle C and a fortissimo middle C on the piano independent of amplification (through, e.g., some electronic means).

2.1.4.6 Summary of Timbre's Relation with Other Auditory Percepts.

We summarize by stating that:

(1) Timbre plays a dominant role in the perception of sound. Other perceptual attributes *surface* as functional elements only when certain acoustic features appear to the listener as homogeneous in so far as timbre is concerned.

(2) The duration percept becomes important when there is a global homogeneity in time that can be described as translational symmetry. The pitch percept becomes important when homogeneity manifests itself as invariance under local time-scale change. And loudness homogeneity manifests itself as invariance under local amplitude-scale change. All of these operations normally produce different timbre. But when homogeneity (symmetry) sets in, then other features of the versatile ear come in as funtioning elements.

(3) While pitch, loudness and duration are simple percepts, or perceptual primitives that are readily at our disposal and are manifests of the ear's versatility (or multiple processing capability) they are not necessarily independent in the sense that we need all of them together with the percept of timbre to specify a sound. Often they are redundant when timbre appears to be sufficiently rich in information-content. Therefore, while it is logical to assume that there are intrinsically different processing mechanisms whose outputs are these perceptual attributes, it is premature to think of them as *independent* dimensions in the sense of specifying a sound. In other words, the same parameter dimensions in the control space may simultaneously affect different percepts. That is, the physical data that controls pitch generation may also partially control timbre generation. Similarly, the physical data that controls loudness and duration may also partially control timbre generation.

2.1.4.7 Control Space of Timbre also Controls Other Auditory Percepts.

It is quite reasonable that such interaction takes place. For example, Just as the rate at which the waveform repeats or nearly repeats itself determines a pitch, so does the manner in which such repetition takes place determines a timbre. Thus pitch and an aspect of timbre are different measures of the same physical phenomenon. Similarly, the amplitude envelope describes the steepness of attack, the extent that certain vibrational forms are sustained, and the rate with which such forms disappear. These are important to timbral perception but also describe the rate at which energy evolves in the sound, and hence the rate it is transferred to the ear. Thus these also have an affect on our perception of loudness. This kind of interaction also has a parallel in the perceptual relations between the look (appearance) and the height, the look and the size, and the look and the width. They are not necessarily independent variables but the complex percept often subsumes the simpler ones.

2.1.5 The Scope—Notions of Musical Timbre.

By timbre, we mean musical timbre, unless specifically stated otherwise. But where do we draw the boundaries for the class of acoustic stimuli that give rise to musical timbre?

2.1.5.1 Helmholtz's Definition.

Helmholtz defines musical timbre to correspond to the class of periodic acoustic signals, perhaps so as to justify Ohm's acoustic descriptions of musical timbre. But of course Helmholtz's definition is incompatible with reality. As Winckel has pointed out, no acoustic waveform is ever periodic or even nearly periodic in its entirety, and the phenomenon of acoustic transients is actually a natural ingredient of any musical timbre. Most well-acknowledged musical timbres lie somewhere in the "continuum" between the extremes due to periodic vibrations and noise. And what is considered as musical depends to a large extent on context. In the literature, there have been essentially three approaches to dealing with the scope of musical timbre. The first approach follows Helmholtz's idea and deals exclusively with stationary acoustic signals. The researchers include Stumpf, Plomp, and Slawson. Their stimuli are of course quite artificial in addition to the limitations in the properties the approach can afford to study. We do not know whether their properties are the most important concerning the perception of timbre.

2.1.5.2 Definition of Others.

The second approach, notably followed by Backhaus, Luce, Wessel, and Grey, confines musical timbre to those arising from excitation of traditional musical instruments of Western culture. This approach has the advantage of dealing with naturally occurring sounds and at the same time not having to say what musical timbre is. The disadvantage is of course that it does not say what happens when we go outside the class of instruments they have studied, and especially what happens when we start dealing with computer synthesis of timbres which cannot be characterized by Helmholtz's description.

A third approach, followed by Erickson and Schaeffer, does not attempt to say what musical timbre is, and at the same time, places no limit on where it can come from. This approach considers musical timbres as *given*, determined by the ear and depending on context, or from the viewpoint of synthesis, musical timbre is a function of *treatment* and is still determined on an empirical basis. But how do we articulate these treatments in order to support an analysis/synthesis environment?

2.1.5.3 Some Observations.

We will address the question of what might constitute musical timbre from analysis/synthesis considerations later. But first we want to specify what we mean by musical timbre by a more or less enumerative approach. Then we will attempt to make some general observations about the character of these sounds. The universe of timbre we are interested in includes ideally everything that is potentially musical material. But being musical clearly depends on context. And ultimately it is the ear that will be the judge. We have seen that the universe of musical timbre is not restricted to the class corresponding to periodic vibrations. And yet we don't know what the new boundaries (new with respect to Helmholtz) are. We do know at least a few facts.

(1) Most ears do find musical timbre in traditional musical instruments of the West.

(2) The ear also finds speech timbre musical. In fact, speech sounds have been extensively and successfully used in musical composition from Schoenberg, Berg to Maxwell Davies in the last one hundred years. And it is well known that musicality is often not a mere function of good resonance quality in the sound production but also a function of the "expressiveness"

of the timbre, as is conveniently found in speech voices. We will see that the notion of timbral "expressiveness" might be related to the dynamic character of timbre in general. There is indeed plenty of evidence that the composition of ancient Chinese poetry strongly exploited the dynamic features of timbre for expressiveness, an important source of musicality.

(3) We know that periodic structure is not to be found in the most characteristic segments of most percussive instrument tones nor in most voice sounds, specifically, the consonants.

(4) Nevertheless, neither the highly non-stationary percussive sounds nor the consonants nor continuous speech in general are structureless.

2.1.5.4 Organizability as Criterion for Musicality.

The types of sounds we commonly encounter, in speech or in music, as mentioned in the section above, are not perceived as structureless (at least we would not be able to comprehend continuous speech if the ear operated in that fashion). And there is strong evidence that the ear adapts and organizes. Thus we conjecture that *a timbre is musical if the ear can organize the acoustic image as it appears in the membrane response.* We will make this conjecture precise and plausible in sections under 2.2 and 2.3.2. At this point, we must emphasize that we need to re-examine the mechanics of perception beyond Helmholtz's passive analytic model of perception to see how the ear might actually organize the acoustic image it receives to determine whether it is musical, or how musical it is. We will treat this issue in detail in 2.2. Then we will try to see where the boundaries of musical timbre lie with respect to the space of all timbre (see 2.2.3.5 and 3.13).

2.1.5.5 The Role of Context in the Criterion for Musicality.

But before we articulate this subject in more detail, we want to stress that a musicality criterion is a relative one. In other words, it is not so important to concern ourselves with the decision of whether an isolated sound event should sound musical as with the decision of whether sounds appearing together do sound musical. If we accept this view, then the important question concerning the musicality of timbre is: Given one or more musical timbres (by some subjective criterion), how do we make sure that new timbres generated by modification of existing ones should remain musical (by comparison)? We will see that this notion will actually provide us with insight into synthesis of new timbres.

2.1.5.6 Summary.

With these facts and other previously introduced notions in mind, we will include as musical timbre all speech timbre in the many different ways it is enunciated, timbres from all kinds of resonant acoustic cavities including those of traditional musical instruments of various cultures from all over the world, as well as any sound generated from a digital computer which a typical musically sophisticated listener would consider as musical under some context, and above all, the derivatives of all these timbres as they are modified by their acoustic transmission environments (a concept explained in 2.1.2.3, where sounds, not sources, are the direct physical basis of timbre). The foregoing is based on the assumption that *these sounds satisfy the organizability criterion of musicality.*

2.1.6 The Dynamic Character of Timbre.

2.1.6.1 The Meaning.

By the dynamic character of timbre, we mean more than just the observation that the frequency distribution of the acoustic energy of a sound varies from moment to moment in the course of the acoustic event for most sounds we encounter. That is, we mean more than the observation that our perception of the timbre is a function of this spectral variation as a function of time in some overall sense—like the area swept out by the orbit of some moving object with respect to some reference point. In this example, the area is a function of the time-dependent orbit but in itself, the area is a scalar. It is a static concept.

Similarly, the Fourier transform applied to an acoustic event from $-\infty$ to ∞, or from beginning to end as we would like to think about it, is a static concept. A different waveform, stationary or non-stationary, will give a different Fourier transform, so much so that the transformation is invertible, meaning that exactly one waveform can be recovered from a given Fourier transform and that is the one we originally had. Therefore, we could have used the notion of Fourier transform to describe our perception of timbre if the only significance we could attach to the notion of dynamism regarding timbre was that timbre was some *overall* function of the nonstationary character of the waveform, whether we choose to look at the nonstationarity of the signal from the viewpoint of a Fourier transform or a series of windowed Fourier transforms.

By the dynamic character of timbre, we mean that perception of the non-

stationarity of the acoustic signal actually consists of a series of perceptions, or an act in continuous motion, i.e., a process (journey, as opposed to destination), in short. It is *not* a static concept like an integral or a Fourier transform of a time function, stationary or otherwise.

Therefore, by the dynamic character of timbre, we mean in addition the ability of the ear to assess the importance of each act (or at least each group of acts in some neighborhood—in a continuum sense) in the continuous motion and their relative importance to each other.

Finally, by the dynamic character of timbre, we also mean the ear's ability to relate timbres of a group of sound events on the basis of how important timbral regimes of one sound relate to those of another.

As a result, by a dynamic description of timbre we distinguish between (1) timbre as a (perceptual) *process* and (b) the *dynamic relationships* among the processes. In other words, we have a description of timbre in terms of its internal dynamics, and a relational description of internal dynamics among different timbres.

2.1.6.2 The Reasons.

There are three fundamental reasons for the dynamic character of timbre.

(1) Fundamentally, any sound has a dynamic character simply because it must begin and it must end. This character is intrinsic in sounds of acoustic orgin. Although electronic sounds have become powerful alternatives for sound production and although a sound can be synthesized to display stationary characteristics for arbritrarily long periods of time, the dynamic character of timbre has not been successfully disposed of for one surprising reason. Namely, stationary sounds are perceived to be "electronic" or to "lack a life-like quality." Dynamic amenities must then be reintroduced into the digital synthesis involved.

But more important than the fact that sounds must begin and must end, and perhaps because the non-stationary characteristics have been made part of the routine listening task, communication ranging over all higher animals, including the birds, does make extensive use of dynamic composition of sounds, to code their messages. It is well known in biology that most of these animals have a similar way of intoning spectral trajectories to express messages such as "threat!" (to ward off unwelcome

species), "warning!" (to alert friends and relatives to imminent danger), "appeasement" (to show desire for communion), etc. Similarly, in speech communication, it is well known that transitions of formants of spectral regions are perceptually important. And a continuous speech pattern has yet to be exhibited which does not contain dynamic acoustic passages.

(2) The ear is equipped with what psychoacousticians call auditory windows.* An auditory window is a duration over which amplitude variations of an acoustic signal affect the response of the basilar membrane. The extent of the window is a function of membrane location. If we think of the basilar membrane as a collection of fibers modelled by a system of mechanical resonators as Helmholtz did, then each resonator exhibits a characteristic damping time which determines the extent of its auditory window. The physical parameters that determine the damping constant also determine the bandwidth of resonance. If τ is the time constant of damping, and β is the bandwidth of resonance, then $\tau \cdot \beta \sim 1$. As explained

* Although the notion of auditory windows is a familiar one for the psychoacoustician, it has not been fully exploited in the field of timbre research other than the way it is used in short-time Fourier transform analysis and synthesis of timbre. Even in this case, the window length is constant over the entire frequency range, contradicting the behavior of fiber responses. The critical band analysis and synthesis techniques of Petersen [Petersen and Boll, 1983] do take into account the variations of window length as a function of analysis channel frequency. However, removing some of the sinusoidal components which fall within the same critical band usually distorts the timbre and appears theoretically unjustified in view of the response behavior of the basilar membrane.

The notion of critical band appears most fruitful in the study of loudness perception. It also explains certain pitch discrimination phenomena (see [Schubert, 1980]). In the field of timbre perception, its role is unclear. We feel that there is no evidence a *spectral* interpretation of the critical band concept would add to our treatment of timbre, where a fiber's response to any acoustic signal is a matter of degree and where the main idea of timbre perception is a hierarchical composition or organization of features or patterns in the space-time response of the basilar membrane.

in 1.2.5, the *instantaneous* response of a causal, linear, damped mechanical resonator as described by Helmholtz is a function of past input data weighted exponentially, beginning at the present and extending into the past over the damping time. This means that the dynamic character of sound typical of natural timbre is closely followed. In particular, the time constant of the fiber at a particular location of the membrane measures the extent the acoustic signal is coherent or "periodic" with period equal to the characteristic period of the fiber. Depending on the location of the fiber, the length of the auditory window determines three different temporal regions:

(a) The window time T_W which is some multiple n of the characteristic period of the fiber. This time is responsible for the ear to decide whether the acoustic pattern is organizable into "repeatable patterns," whether it needs to do an update, or go to rest.

(b) Local time $T_L(\ll T_W)$, equal to the characteristic period of the fiber, which determines some microtimbre characteristics to be explained later. In the event the signal "repeats" some local pattern, then these microtimbre characteristics have a spectral interpretation close to that of Helmholtz.

(c) Global time $T_G \gg T_W$. Over this time, the microtimbres might wander significantly and form a trajectory. The dynamic character of sound is thus recognized within the context of the auditory window of perception.

(3) But why don't we pick a local timbre? The answer lies in what is considered natural or what makes more sense. First of all, it is not clear how to delimit a segment of the waveform as physical data for the local timbre since the pattern of variation never begins or ends sharply. Secondly, it is not clear whether it is meaningful to do so during highly non-stationary regimes (many percussive sounds have their most prominent timbre exhibited in these regimes). Thirdly, transients are important perceptual elements. Fourthly, the ear's auditory window follows the sound as a *running* window. Finally, the ear does follow the entire sound event and often

the ear does not make *a priori* decisions as to which part it pays more attention to. In other words, the timbral prominence as a function of the evolution of the sound varies from sound to sound. And this is why a dynamic description seems so desirable.

2.1.6.3 Examples.

We will present examples in classes to illustrate the different meanings we gave to the dynamic descriptive in 2.1.6.1. (Non-stationarity in the acoustic waveform is the prerequisite for a dynamic perception of timbre. But keep in mind the ear's ability to follow it.)

The first class of examples illustrates the fact that perception of timbre is a process, i.e., a series of perceptions in continuous motion. Consider a "sine bundle." A sine bundle is a collection of closely spaced sine waves where the spacing is a small fraction of the frequencies. For example, let

$$f(t) = \sum_{k=1}^{n} \sin 2\pi(\beta + k\Delta)t,$$

where n may be $2, 3, \ldots$. Note that we can write

$$f(t) = A(t) \sin 2\pi(\beta + \frac{n\Delta}{2})t,$$

where

$$A(t) = \frac{\sin \pi n \Delta t}{\sin \pi \Delta t}$$

may be thought of as an amplitude modulation factor. Say in fact $n = 10$. Let β be 100 Hertz and Δ be .1 Hertz.

If we look at the signal, the energy is concentrated completely within the interval from 100 Hertz to 101 Hertz. Also, f is periodic with a period of $P = 1/\Delta$, which is the smallest time commensurate with $\beta + k\Delta$ for all the k's involved. The period P of ten seconds is very long, i.e., the "missing fundamental" is subaudible.

If the ear acts only as a Fourier analyzer, we would not have heard or know the ebbs and flows of the sound because the magnitude spectrum tells us nothing about the temporal property of the sound in the perceptual sense. But there is more than just the ebbs and flows. The ear does not resolve the frequencies. Neither does it perceive a uniform timbre typically exhibited by a steady harmonic tone. Although the same principle of superposition applies to a steady harmonic tone

as to these narrowly spaced sinusoids, their salient characteristics (to the ear) are very different. For the steady harmonic tone, the timbre is temporally uniform but spectrally rich. For the sine bundle, the ebbs and flows of energy due to beating, a special case of superposition where the frequency difference is small compared to the frequencies themselves, are clearly heard. The timbre is spectrally dull but temporally non-uniform. The dynamic character is even more dramatic when we superpose a number of harmonic tones whose fundamental frequencies are spaced very closely together, say .1 Hertz again. If there are say fifteen harmonics, we hear dramatic runs of tones with harmonic pitches with different rates of recurrence. The rates change with the fundamental frequency spacing of the harmonic bundle. These phenomena are very clearly explained by acoustic principles alone. The main point is, however, that the ear follows the temporal pattern.

Therefore, the ear is a device that follows the acoustic events in time very closely. A very similar phenomenon occurs with our perception of bell or gong timbres.

In exactly the same manner we perceive speech. A simple word in the form of a continuous burst of acoustic energy is perceived as a series of clearly identifiable timbres we associate with the concept of phonemes. Taking a Fourier transform from the beginning of the acoustic event to the end of it does not reflect what we hear. It is even true with bird songs. The long sustained utterances have noticeable spectral glides along with their pitch glides.

The second class of example illustrates the fact that the ear actually follows the timbral dynamics carefully and makes decisions as to which part it prefers to identify the sound with. One of the best known examples of this phenomenon is the role the transitions between consonants and the vowels they are connected with plays in their perception. The transitions allow people to discriminate between the *d* and *t* type, the *g* and *k* type, and the *b* and *p* type of consonants. Furthermore, the transitions are even more important in differentiating between something like *b* and *p*, for example, when they themselves are very noisy and hence very difficult to organize and recognize individually. Another example is provided by Luce's study of musical instrument timbre. He reports that the attack waveforms of the string family, the flute, the oboe, and the bassoon are quite irregular and suggests that the ear focuses its attention on the steady-state timbre. As it turns out, Grey's

multidimensional scaling of $\frac{1}{3}$ second long tones puts the flute and the bassoon with the string family. Similarly, a French horn without a sufficient steady-state sounds blunt.

Even a percussive tone like that of a marimba has a continuous transformation of timbre and the ear can be trained to follow it and notice it. An easy way to tell is simply to listen to the sound in pieces. If we ignore the clicks as a result of the abrupt segmentation, we can tell that the tail is definitely dull sounding in contrast to the brilliant timbre in the beginning. We can also avoid the clicks by synthesizing a sound that preserves the beginning timbre of the marimba but lets the second half of the sound decay away with the marimba's spectral content at the peak regime. The tone sounds perfectly natural and percussive. In fact, a highly trained musician thought the timbre to be that of the original marimba, and the real marimba timbre, the one with a dull ending, as a synthetic one. We conjecture that the listener who made this confused observation had been working with frequency modulated sounds a lot in that period of his life and the preponderant energy concentration in the high frequency region characteristic of frequency modulated sounds had influenced his perception.

Still another way to hear the timbre of the ending is through comparative listening where a reference timbre may be one which has a similar ending but a different beginning. This is in fact heard through examples where the reference timbres have the beginning of an /a/ sound.

Finally, an /a/ with an octave drop is perceived as clearly distinct from one where the drop is replaced with a steady continuation of the pre-drop regime intoned with slight vibrato to simulate naturalness. Furthermore, when the drop is shortened in pitch by approximately ten percent in pitch and two percent in time (not by resampling), it is judged by some highly trained musicians as differing more than just the difference in pitch. The description is that the new timbre sounds less *spacious* but otherwise very similar in timbral character. We should note here that a reason why we don't normally associate more than a single timbral percept with a sound has to do with the kinematic nature of sound and the continuous sliding window of observation our ear has. As a result, the transformation of timbre in the course of a musical sound is usually gradual and subtle. But we do follow these transformations of timbre and make decisions about their importance, because if

we distort the attack of the marimba, our perception of it is severely affected. Similarly, distorting the attack of an oboe would not seriously affect our perception of it but distorting the steady-state would.

Finally, the dynamic character in describing the relationship among timbres seems to be a natural consequence of the way the ear perceives the timbre of each sound. If the ear identifies most sounds with a timbral feature that appears halfway into the sound of event A but does so with another that appears in the first five percent of the duration of sound event B, then a transition from A to B would have to involve a transition in the weight of these features as well as the time these features appear. In general, the most important feature, and for that matter, the second most important feature and so forth, will be different from one timbre to another. Not every timbre is most distinguished by its attack, despite the importance it has for the temporal context. In fact, later, we will formally describe the relational dynamics in terms of the perceptual importance trees of the timbre under study. For the marimba and most percussive tones, the attacks are usually the most important. On the other hand, for sustained tones of long duration, vibrato plays a dominant role. Still, in sustained tones of intermediate duration, such as our /a/ with an octave drop, it seems that the importance of the pre-vibrato regime assumes about as much perceptual significance as the vibrato regime. These observations have actually been written down as rules by Schaeffer (see 1.7).

There are in fact concrete examples in the literature which we can reasonably assume to illustrate the dynamic-relational character of timbre. Gordon [Gordon, 1984], in his study of the perception of attack transients in musical timbres, discovered that a melody composed of an arbitrary selection of timbres from the group consisting of the flute, the oboe, the clarinet, the bassoon, the trumpet, and other wind instruments sounds jagged, i.e., uncoordinated, or out of synchrony with the intended rhythmic pattern. He further discovered that by synchronizing the most prominent feature of the attacks of the timbres chosen, he was able to synthesize the melody with the desired rhythm. (Note that these timbres are taken from Grey's $\frac{1}{3}$-second long analyses, therefore Schaeffer's rule concerning the importance of vibrato (or, rather, tremolos for the instruments under discussion) in sustained sounds does not have much influence. In other words, all the important features are largely determined by the consideration of temporal context alone.)

Another example can be found in Morrill's phraseology study of trumpet tones [Morrill, 1977]. He observes that a global amplitude envelope spanning the entire phrase is needed to produce a perceptually coherent sequence of tones. In this case, temporal proximity between the trailing part of the previous tone and the attack of the current tone requires non-uniform treatment over the energy evolution of individual acoustic events, and the necessary treatment is a function of adjacent acoustic events.

An interesting illustration of the dynamic character of perception can be provided by the pulse sequences. Consider the auditory windows at different fibers, using five times the characteristic period length:

$L(2000 \text{ Hertz}) \sim 2.5$ milliseconds

$L(1000 \text{ Hertz}) \sim 5$ milliseconds

$L(100 \text{ Hertz}) \sim 50$ milliseconds

When we have a pulse pair at a separation of say ten milliseconds, then the efficient transfer of energy requirement is not satisfied enough to excite the fibres around one hundred Hertz and the pulses are too far apart for the sensitive fibers at one to two kiloHertz to produce some kind of continuous response through the convolution with their exponentially decaying impulse responses. So, we have two unfused pulses primarily detected over the wide range between one and two kiloHertz.

But if the pulse pair is continued to become a pulse train, then the repetition is now sufficient to excite the fibers around one hundred Hertz which now contribute the dominant response of the ear. So we perceive a timbre corresponding to a continuous wave even though the acoustic event is non-continuous.

Now if the pulse pair separation is changed to 5 milliseconds, the hypothesis of efficient energy transfer is approximately satisfied at the fibers around one kiloHertz but not at two kiloHertz. But since efficient transfer of energy means strong response, the perception of continuous excitation at one kiloHertz (where the pulses are convolved within one auditory window) dominates the perception of discontinuous excitation at two kiloHertz (where the pulses cannot be convolved within one single window)—given that the fiber sensitivites are not too different between the two regimes. So the pulse pair sounds somewhat fused.

Now, if the pulse pair is separated by 2.5 milliseconds, the case is even more favorable for fusion. And it is clear there is a transition from non-fusion to fusion.

So here we see the dynamic nature of auditory perception.

2.1.7 Innate Language of the Ear.

There is no argument that we must have a sound event at our ear drum and a functioning ear with a brain connected to it before we can have a sensation of any sort as a response to that sound event. That sensation is not a random phenomenon—it is a function of the sound event in a very specific way. That very specific way is the way of the auditory processor. First, there is the complex mechanical resonse pattern on the basilar membrane. Then there is an even more complex neural transduction response involving electrochemical activities. It is amazing that somehow we can even identify certain characteristics in the sound in some quantitative fashion. In other words, we can describe that sensation *in words* (to some extent) if we are asked. But our ability to describe remains quite limited. And this limitation is a reflection of our lack of conscious understanding of the auditory activities. In other words, the auditory process is yet to be *understood* by the thought process. This is due partly to the fact that sensation is not naturally intended to be *described.* The inherent language that is used in *abstracting* information from the auditory processor is not necessarily identical to our own spoken language, or involved in our conscious thought processes, both being rather recent developments in our evolutionary history.

In fact, the ear—which has existed since fish appeared on the scene—must be of a much older existence than the mind and spoken language. Therefore, if spoken language is an invention to accommodate the mind and is a result of gradual development, like everything else, it is far from clear that it would encompass the inherent language in our auditory process—what we might call the auditory language.

2.1.7.1 An Example of the Visual Language.

The look (appearance) of a visual figure is equivalent to the perception of the relationships among local forms or local geometric varations. In other words, the look of any visual figure is the perception of spatial variations or geometric relationships of these generally unrestrained spatial variations.

When we talk about these local forms, we imply *constraints* imposed upon an otherwise arbitrary variation in a three dimensional grid. These constraints form the basis for our perception of *distinctive features* in a *hierarchical* fashion. So we

recognize the form of the head, the torso, the arms, and the legs. Further examination within each of these forms reveals finer details. For example, we recognize the eyes, the ears, the nose, the mouth, the brows, the hair, and the cheek within the shape of the head. We also recognize the shape of the hands, the upper arms, and the lower arms within the shape of an arm. Furthermore, we recognize at each level the relationship among the shapes perceived at that level. For instance, we perceive the size, shape, and position of the head relative to those of the arms relative to those of the legs. We observe the ratio of the length to the width among these figures. We see the size, shape, and position of the eyes relative to the nose, to the brows, the cheeks, the mouth, and the chin. The look of a figure is therefore a list of its components or subfigures, plus their relationships.

Therefore, it seems reasonable to formally define the look of a visual figure as a composition of geometric features or a relational description of these geometric features, with each feature in turn as a composition of subfeatures in the neighborhood of the feature.

2.1.7.2 The Auditory Language.

Similarly, we have perceptual primitives of pitch, loudness, and duration in audition, just as we have primitives of height, size, and width in vision. Just as the three dimensional physical object involves many degrees of freedom associated with its objective existence, there are also many degrees of freedom or dimensions associated with the existence of an acoustic waveform. If we can detect perceptual differences in response to small variations here and there in the waveform, just as we can detect perceptual differences in response to small distortions in the shape and construction of a physical object, then there is actually a very large amount of information our sensory organs can actually detect and make sense of.

While we expect our higher level processors for these organs to do feature extraction, data reduction, etc. as they are often discussed in pattern recognition studies, it remains true that the dimensionality of the perceptual space is inherently large because of the fine-grained nature of the mapping from physical world to the perceptual world (and that is why we pick up small differences). But at the same time, it is not clear how we describe this multitude of dimensions after a few simple ones that we have mentioned above. We have a sense of the look of a human figure or his or her countenance. But how do we describe the look in a complete way that

is also clear? In 2.1.7.1, we introduced a sort of a formal definition (description).

The fact is that we usually use a set of adjectives but that is only because we have nothing better to say. Our language is not graphic or geometric and neither is it kinematically oriented or form oriented. It is not easy to describe what a pattern or a geometric form is without mathematics. Therefore, as much as we can limit our attention to say the look or manner in the general attribute we are talking about, i.e., other than the height and size in our visual perception, so must we be satisfied with confining our attention to that overall quality of a sound beyond its pitch or loudness when we talk about timbre. What we are saying is that because of the complexity of the overall percept we cannot begin to pin down every single dimension and call it a different perceptual element.

We have grown used to knowing only a few primitive percepts and calling the rest by a certain name. For example, in vision, the *look* of a figure means all those features and their relationships for which we have no immediate descriptives.

Similarly, in the perception of sound, there is a general understanding that when we talk about timbre, we are essentially saying "let us listen beyond the pitch and loudness, and see what else we do hear." It is quite plausible that the ear does recognize features and their compositions exhibited in the complex membrane response pattern, but that we have no immediate or equivalent verbal descriptives for them. It is quite plausible that timbre is in fact describable recursively as a composition of auditory features analogous to the formal description of form in vision.

So one of the goals of a timbre theory is to discover the inherent language of audition which goes beyond "what is this?" and "what is that?" with regard to the source, and which goes beyond how "sweet" or "bright" a sound is—clearly these constructs are not in the ear's innate language. If we can do this, then we may be able to exploit the ideas we have discussed above for practical use and we will go a long way towards making timbre composition possible (not to mention deepening our understanding of how the ear-brain works). In 2.4, we will see how the idea of timbre as composition of features can be concretely formalized.

2.2 Perceptual Foundation of Timbre.

The auditory pathway is complex and remains not very well understood. However, it is generally believed that perception of timbre originates in the cochlea whenever the complex membrane response pattern to the acoustic waveform gives rise to the percept of timbre, as Helmholtz first suggested more than a century ago.

This belief finds parallel in the perception of pitch, where at least some aspects of frequency selectivity are a function of "place" on the length of the membrane.

Similarly, the perception of loudness is known to be a function of the energy summation over frequency sensitive "places" on the length of the membrane, i.e.,

$$\mathcal{L}(I) = \sum_{f_i \in S_f} c_i \mathcal{E}(f_i),$$

where c_i reflects the sensivity of the membrane over the $i^{\underline{th}}$ frequency and $\mathcal{E}(f_i)$ is the signal energy for the $i^{\underline{th}}$ frequency in the set S_f. In this thesis, we will make the assumption that timbre perception originates with the complex mechanical response pattern on the basilar membrane induced by the acoustic signal. The first thing we do is to point out the aspects in the membrane response that are important to timbre perception and explain how.

2.2.1 The Mechanical Basis—Passive Observer.

Although the process through which timbre is perceived necessarily involves several transactions in the brain, the fact that changes in the detailed variation of the waveform induce changes in the timbre perceived suggests that timbre represents a measure of the acoustic pattern of variation in the stimulus. The acoustic pattern of variation is mechanical in origin and the membrane response pattern is a detailed manifest of the acoustic content of the signal. But how does the ear interpret the response? Minsky's notion of agent [Minsky, 1986, p.18] in a society of ear seems a reasonable point of departure. And Helmholtz's analytic model seems to mesh well with the idea. Therefore the first step in our attempt to interpret the complex membrane response pattern is to follow Helmholtz and treat the response pattern as the response of a system of independent damped harmonic oscillators. The dynamics of each damped harmonic oscillator are described by a linear differential equation of second order with constant coefficients. These coefficients determine the resonance and damping characteristics of the oscillator.

These harmonic oscillators can be thought of as the eigensolutions of the membrane. But unlike many vibrating systems of many degrees of freedom, we can identify each of these eigensolutions with a neighborhood of fibers or a slice of the membrane across the length of the membrane. Therefore each harmonic oscillator is identified by a one-dimensional spatial coordinate. The physical characteristics, i.e., the resonance frequency and damping, of each oscillator are therefore also functions of the spatial coordinate. This is essentially Helmholtz's model of the complex response pattern of the membrane. Some of the salient features of such a model are

(1) The response of each harmonic oscillator to an arbitrary acoustic stimulus is a function of time;

(2) Therefore, the system of responses is a space-time response pattern where the response time functions of the oscillators appear across the membrane "simultaneously" (actually with a characteristic delay discovered by Békésy);

(3) The system of responses can be characterized as a single input multiple output response (SIMOR);

(4) The SIMOR displays interpolated characteristics as a function of place (because the physical characteristics of the responses are smoothly varying functions of place and because each dynamic system is driven by one and the same input). For a graphical representation of a forty point digital computer simulation of the impulse responses of the resonators along the basilar membrane see e.g. Flanagan [Flanagan, 1972, p. 122].

(5) Every harmonic oscillator is forced to respond to an arbitrary function (a consequence of the differential equation that describes its dynamics). But the degree to which a particular oscillator responds to a signal can be described succintly in Fourier terms, which say that the frequency response of the oscillator $R(\omega)$ is given by the product of the Fourier transform of the impulse response of the linear, time-invariant oscillator $H(\omega)$ and the Fourier transform $\Phi(\omega)$ of the input $\varphi(t)$. That is,

$$R(\omega) = H(\omega)\Phi(\omega)$$

where

$$H(\omega) = \frac{1}{(b^2 - \omega^2) + ia\omega}.$$

Of course we can obtain the same result from the convolution integral. The magnitude response, $|H(\omega)|$, assuming the form

$$\frac{1}{\sqrt{(\omega^2-\omega_s^2)^2+\alpha_s^2},}$$

where ω_s is the resonance frequency in radians per second and α_s is the bandwidth of the harmonic oscillator, informs us of the degree to which the oscillator responds to a signal based on its frequency content. In figure 2.2.1, we see that (i) the oscillator responds most favorably to a signal whose energy is concentrated around ω_s. (ii) It responds to signals of all frequencies. (iii) In fact, when the input is a sine wave

$$\begin{aligned}|H(\omega)| &= C\frac{1}{\sqrt{(\omega^2-\omega_s^2)^2+\alpha_s^2}}\frac{1}{\sqrt{(\omega^2-\omega_y^2)^2+\alpha_y^2}} \\ &= \frac{A}{\sqrt{(\omega^2-\omega_s^2)^2+\alpha_s^2}}+\frac{B}{\sqrt{(\omega^2-\omega_y^2)^2+\alpha_y^2}},\end{aligned}$$

where A, B, and C are constants, y denotes the input of an exponentially decaying sinusoid at ω_y with bandwidth α_y, such as the vibration of a tuning fork. Note that since α_y, is the reciprocal of the decay time constant and since most sounds have a significantly longer lifetime than the life of a membrane fiber modelled by the oscillator, the response consists of the sum of the oscillator's only transient (at ω_s) and the forced oscillation of the input (at ω_y) with the latter dominating the vibration pattern—a very advantageous characteristic of a message receiver. (iv) When the input has a flat spectrum, i.e. $|Y(\omega)| = 1$, the oscillator's own characteristic vibration prevails and a sine-like waveform emerges. This example illustrates the bandpass characteristic of the oscillator. (v) But the bandpass property is not an ideal one—one that requires an impulse response extending from $-\infty$ to ∞, violating causality. It therefore responds to signals of all frequencies. (vi) The asymmetry of the filter characteristic (as a function of frequency) is well documented in psychoacoustic and physiological experiments. See, for example, Schubert [Schubert, 1980].

(6) Therefore there is always a collective response in the fibers to external stimulus. Regarding the collective response behavior of the oscillators,

Schubert [Schubert, 1980, p. 50] says, "It looks as though the auditory system attempts a solution to this problem [the time-frequency product uncertainty] by having not one analyzer but several. Out at the cochlea, where ideally both options would be kept open, a dual function looks quite likely."

(7) In addition, if the signal is periodic or nearly periodic, there are fibers which respond not only to the signal but vibrate sympathetically or strongly;

(8) Therefore, in general, there is a simultaneous foreground (consisting only of strong individual responses) and background response(by the collective response of all the fibers). Regarding this point Schubert [Schubert, 1980, p. 50] says, "If we look closely at the modern revision of the frequency response of a segment of the cochlear partition ... it appears that the fiber associated with this segment may contribute to fine frequency analysis when some component of the signal lies near the characteristic frequency—the peak of the curve—but contribute to a different aspect of analysis when the signal lies in the parts of the spectrum further from this best frequency. The entire cochlea may be contributing, at the same time, outputs from a narrow filter for optimum frequency separation and outputs from a fairly broad low-pass filter that preserves more of the timing information at the expense of fine frequency resolution. This possibility is supported by some experiments of Kiang and Moxon (1974) indicating that low frequency information is available in the tails of the tuning curves of fibers with high characteristic frequency. ... it is also evident that some fibers with high characteristic frequency respond over a wide frequency range when the signals are of moderate level."

(9) Each individual response is characterized approximately under normal circumstances by the convolution integral of the input and the fiber's impulse response which is in turn characterized by the resonance and damping characteristics of the fiber. The resonance frequency is a logarithmic function of place and the damping time constant is governed by the constant Q law.

(10) Each damping time constant presents to the ear an *auditory window*

so that a strong individual response would result only if the input contains a *pattern* capable of efficient energy transfer, i.e., one which contains quasi-repeatable subpatterns whose "repetitive" frequencies agree with the characteristic frequency of the fiber or are a small multiple of the latter so that efficient energy transfer is still possible within the duration of the window. This means that if the "periodicity" or the frequency of some repeatable subpattern of the input changes faster than the auditory window, there will not be a strong individual response at the fiber under study even if the frequency sweeps past the fiber characteristic frequency. The auditory window essentially determines the extent of the past of the input that must play a role in determining the present response behavior.

(11) The space-time pattern of the membrane response plays a critical role in all forms of auditory perception except those related to spatial or directional hearing. In reference to the ear's time-place dual capability in decoding information from the auditory image, Schubert [Schubert, 1980, p. 62] says, "Considering the complexity of auditory processing and the variety of pitch perceptions, there seems little purpose in choosing one to the exclusion of the other. A sufficiently versatile and adaptive sensory system should make use of whatever clues yield satisfactory information about events of interest in the environment, and will probably be characterized by redundancy of information rather than parsimony. We have ample reason to believe both kinds of information are available to the auditory processor."

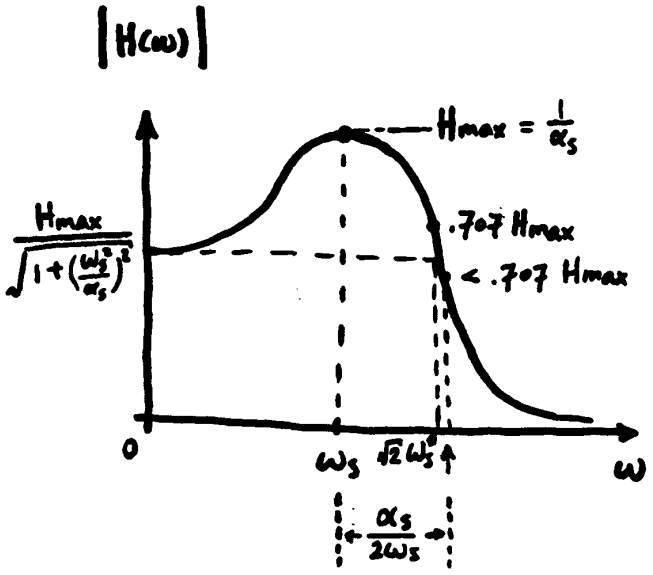
(12) The response is essentially a passive phenomenon in the same sense as the bars of a xylophone being excited. Therefore a mechanical response of this kind cannot be a complete description of the process of perceiving a timbre. Invoking some kind of unconscious doctrine as Grey suggests Helmholtz has done does not help us understand how the multitude of output patterns are actually translated into some of these features the ear recognizes and allows us to articulate. But nevertheless, the society of ear can and usually does start with a passive physical phenomenon. Organization (following Minsky's idea, cited above) is the key to the transformation of such passive phenomena into active consciously perceived ones.

$$\ddot{r} + a\dot{r} + b^2 r = \varphi(t), \qquad a^2 \leq 2b^2$$

$$R(\omega) = H(\omega)\Phi(\omega)$$

$$H(\omega) = \frac{1}{(b^2 - \omega^2) + ia\omega}$$

$$|H(\omega)| = \frac{1}{\sqrt{(\omega^2 - \omega_s^2)^2 + \alpha_s^2}}$$



Note the tritone frequency $\omega_t = \omega_s\sqrt{2}$. For $\omega > \omega_t$ the magnitude response is less than the response at any frequency less than the resonance frequency ω_s.

Frequency Response of the Damped Harmonic Oscillator

figure 2.2.1

2.2.2 The Societal Basis—The Active Observer.

There is clear empirical evidence to support the assertion that the ear does hear, or recognize, *organizing elements* in the signal. It is plausible to assume that these organizing elements are heard because they help the auditory processor to organize the huge amount of information exhibited in the complex mechanical response of the membrane. The process of organization and recognition is essential for the active observer. For every mechanical response pattern in the cochlea, the ear performs an active process of pattern organization and recognition. Evidence abounds that supports this assertion.

2.2.2.1 Perceptual Evidence for Organizing Elements.

2.2.2.1.1 Temporal Organizing Features.

First, consider a waveform with a well defined amplitude envelope. The ear hears the detailed fluctuation which is missing from the amplitude envelope because we would hear a different timbre if the local fluctuations change even if the amplitude envelope remains the same. At the same time, the ear also hears the amplitude envelope as a smooth description of the growth and decay characteristics of the sound—a part of the timbre. In fact, Schaeffer reports that the ear distinguishes these two perceptual features in the timbre of a sound. Furthermore, Schaeffer reports that the perception of the attack (which is part of the growth and decay characteristic of the sound's timbre) is independent of local fluctuation of the waveform.

Our empirical evidence suggests the ear does neither perceive some gross structure as the amplitude envelope alone (ignoring the local fluctuation altogether) nor perceive the waveform in detail only ("getting a bumpy ride" over the wave). It perceives both and the only way we can reasonably interpret the ear's action is through the idea of organization. It is very much like the way the eye perceives something like the eye brow of a human face. Although an eye brow consists of visually separable strands of hair, the visual processor on some perceptual scale tends to organize the elemental features into an outline and thinks of them as a continuous stroke the way it is depicted (on paper, for example) as a single feature (as evidenced by its name).

2.2.2.1.2 Spectral Organizing Features.

Second, consider how the amplitudes in a Fourier magnitude spectrum are usually organized into formant regions when the harmonic structure is such that it is possible, and consider how the formants are organized to form an overall spectral envelope. It is clear from the mechanical behavior of the membrane that the membrane response to these harmonics (when they are resolvable) is to see them as separate temporal response patterns. But if the acoustic signal is quasi-stationary, then each of these temporal response patterns must appear to be fairly regular, regular to the point of being redundant, and only the variation along the place dimension might suggest (to the ear) information content. When this is the case there is good reason to believe that the ear would stop scanning the time dimension ("until further notice") and focus only on exploring the information content in the place dimension. When it does do that, the interpolated nature of the fiber response characteristics allows organizing features in the acoustic signal to surface. Therefore, we expect that when the Fourier amplitudes tend to coalesce into some noticeable features, such as the bumps that signify what we normally refer to as formants, the ear would recognize them and identify them. In fact, speech research by Stumpf [Stumpf, 1926], Peterson and Barney [Peterson and Barney, 1952], K.N. Stevens and co-workers [Halle and Stevens, 1959, and Bell *et al*, 1961], Flanagan [Flanagan, 1972], and others, lends strong support to the idea.

2.2.2.2 Feature Extraction as Prerequisite for Pattern Recognition.

From a data processing viewpoint, SIMOR significantly increases the volume of data. Namely an N-sample sequence is converted into an amount of data of size

$$\sum_{k=1}^{M} N + L_k \sim MN,$$

where L_k is the (effective) impulse response length of the $k^{\underline{th}}$ fiber. Such an increase is however not compatible with the goal of pattern recognition. A major goal of pattern recognition is the extraction of distinctive features from the signal so as to discard unimportant ones.

2.2.2.3 Redundancy as Necessary Condition for Feature Recognition.

The SIMOR is in fact a highly redundant expression of the acoustic signal for we can surely expect to recover the input from the output function of any resonator alone by deconvolving it with the resonator's impulse response (or more simply,

plugging it into the differential equation it solves). But if recovery was the only design consideration in the ear's construction, then we would expect the ear to recover every single signal it received (given that it can deconvolve or differentiate) but not to be able to distinguish noise from a "good" signal (and all the mixtures in between) without storing the entire waveforms of all the "good" ones in the auditory memory. (And what are the "good" ones without being told by some mentor?) There would be no apparent basis for feature extraction. This is so because by definition a feature must be recognizable. It is recognizable when there are reinforcing elements to tell the ear that the feature is indeed significant, not just some noise. Only if we are able to make that decision during the process of feature analysis and extraction will we be able to make similar decisions comparing incoming data with data in memory. And it seems that redundancy is an ideal way to provide that reinforcing element during the process of feature analysis and extraction. Furthermore, smooth variation in response characteristics across the membrane provides an easy reference level for the process of organization and differentiation necessary for the analysis and extraction of features.

2.2.2.4 Input as Source of Organizing Features.

Finally, the organizing features as well as the distinguishing features must come from the input. Thus we expect the amplitude envelope that shapes the global behavior of the acoustic signal to control the global behavior of the response function of all the resonators. And the smooth variation in the response characteristics of the resonator across the membrane makes apparent that organizing element of the acoustic stimulus, *viz.*, the amplitude envelope, through the coherent global temporal pattern across the membrane. Notice that many of the local fluctuations have been smoothed over and the temporal variations over each auditory window should appear "self-similar" over a time-range scaled by the variation of resonant frequency on a constant Q basis and should be "similar" to the amplitude envelope of the signal. Similarly, the spectral features manifested as response patterns as a function of place at some adjusted constant time must be reinforced (by redundancy) over some time (most probably over the auditory window of the fundamental of a quasi-periodic signal) in order to be recognized, differentiated, and stored.

2.2.2.5 Data Reduction Consequence of Organization.

As a result, while the initial pattern is highly redundant, the resultant features can actually involve a smaller volume of data than that of the signal itself, primarily because the organizing elements in the response pattern come from the organizing elements, therefore from the distinctive features, of the acoustic signal itself. It is interesting to note that the redundant nature of the membrane response was probably more a consequence of the ear's versatility or multiple capability as parallel pitch detector, originally equipped as an efficient predator and prey detection mechanism than to guarantee robustness in signal detection. It not only achieves robustness but further plays a useful role in data reduction and feature extraction.

2.2.3 Perceptual Criteria Governing Musicality of Timbre.

Here, we are not concerned about how one timbre is judged more musical than another, or why a trained soprano's aria sounds so musical. Obviously, it has to do with the material quality of the resonating cavity, the steadiness of the energy flow, the great control over the variations of amplitude and pitch, the ability to exploit the acoustic properties of the resonator (by shifting the formant regions to coincide with the harmonic locations, i.e., to optimize the energy transfer between the excitation source and the resonance device), etc.

Rather, we are interested in discovering some fundamental *structural elements* in the class of musical timbres, if they exist, that might help us manipulate timbre without finding too many surprises.

2.2.3.1 Musicality as Balance between Predictability and Innovation.

Human psychology, in general, finds abundant evidence to support the assertion that the tasks a person enjoys doing must be something he or she can comfortably handle, i.e., something not too taxing or difficult. Minsky, in particular, points out that musical listening experience is very much like playing a game of chess. It is at the same time a little challenging to be interesting, but not so challenging or difficult that the listener cannot make sense of it. The flow of acoustic events in a piece of music must provide an element of predictability balanced by an element of surprise or innovation. He speculates that a successful composition must behave like a successful teacher leading the listener step by step through an exercise of education. Perhaps, Minsky's idea is applicable in the microtemporal scale of a single sound event that gives rise to the percept of timbre. As we have pointed out,

the perception of timbre is very much a pattern recognition exercise.

2.2.3.2 Pattern Organizability and Data Reduction as Prerequisites for Musicality.

From the communication viewpoint, timbre is no different from a message which to the receiver is necessarily a statistical problem to tackle. Therefore, the enormous dimensionality in the signal space ($2WT$, where W is the bandwidth in Hertz and T is the length of the signal in seconds) would have to be greatly reduced to be comfortably handled by the auditory processor. Periodicity provides that significant reduction. For instance, if the Nyquist rate for a waveform implies that fifty-four samples are sufficient to describe a period of vibration, then the signal space dimensionality becomes no more than fifty-four for a stationary signal.

2.2.3.3 Structural Elements for Musicality.

2.2.3.3.1 Periodicity.

If the signal can assure the ear of its content within the latter's auditory windows, especially within the one corresponding to the fundamental frequency that spans three to five periods, then the dimensionality reduction is essentially accomplished in so far as the ear is concerned. And stationarity at least in this local temporal sense frees the auditory processor to explore the features in this drastically reduced fifty-four dimensional space for information. It probably will do so by attempting to organize and discover patterns in the place direction of the space-time response pattern. Thus periodicity or local periodicity (quasi-periodicity) is an important organizing element provided by the signal to the ear to decide that it is some enjoyable acoustic event. In this sense, musicality of a timbre is a function of periodicity of the signal. This is of course consistent with Helmholtz's idea. But strict periodicity in the mathematical sense is not necessary as we will show later. The important fact is that periodicity, in some broad sense, is an integral feature of musical timbre and the notion of period trajectory is consistent with the hypothesis that the ear in many natural stimulation circumstances relies on temporal aspects of the signal for pitch detection. (This hypothesis has recently received strong experimental support [Sachs and Young, 1979].) Hence the period trajectory in the dynamic event is one of the most important physical features or correlates of a musical timbre. Naturally, as we send white noise through a bandpass filter, depending on the bandwidth, the frequency at which the quasi-random waveform revisits a cer-

tain phase (say a certain peak) in a vibration pattern spanned by the reciprocal of the bandpass filter center frequency determines the pitch strength—the degree the sound is perceived to have a pitch. But our experience tells us that pitch strength is really an aspect of timbre. Thus musicality of a timbre is tied to organizability of the acoustic signal from the ear's viewpoint (of course, we are somewhat biased by a pattern recognition model of the auditory processor—although it is difficult to believe that the ear-brain does not employ some sorts of pattern recognition, though maybe using different techniques). More specifically than the first consequence of this organizibility requirement on the signal is its possession of local periodicity as judged over the duration of the auditory window of the frequency connected with the periodicity, or the extent to which it possesses this property.

2.2.3.3.2 Exponential Periodicity.

Secondly, the manner in which the membrane detects periodicity is, in Helmholtz's words, by sympathetic vibration. The basis of such sympathetic vibration is efficient transfer of energy from the signal to the membrane by virtue of a synchronized effort. This means, as mathematics will show, that the signal does not have to be periodic in the mathematical sense as Helmholtz requires. For example, an exponentially decaying sine wave of frequency f_1, such as one coming from a percussive excitation of a tuning fork, will stimulate the membrane at the place which responds most strongly at f_1 to vibrate strongly. Similarly, an exponentially decaying superposition of harmonics of fundamental frequency f_1 will stimulate all the places on the membrane where characteristic frequencies agree with the significant harmonics of the vibration form of the signal. The constant Q property requires that the number of periods within the auditory window of each strongly stimulated fiber be constant. Therefore, the foreground temporal pattern will appear coherent and regular and the ear is free to pursue the features along the place direction, especially over the locations of this foreground response, until some new temporal event triggers a renewed scanning (by the auditory processor) over the temporal direction.

2.2.3.4 Adaptation as a Means for Organization.

The dimensionality reduction as a consequence of periodicity can actually be approached when periodicity is significantly deviated from under certain circumstances. For instance, if the non-stationarity of the acoustic waveform is recogniz-

able as a coherent geometric pattern, then that coherence is reflected in the response pattern of the resonators. If the auditory processor recognizes the coherence in some quasi-geometric sense, then it will need only take a little extra effort to follow each individual pattern in addition to that of a truly periodic one. The little extra effort refers to its ability to follow trends of movement in the recent past and adapt the state of motion to its new form. The ability to adapt comes naturally with the biological system as many other less complex systems also do. The cue can come from the redundancy of a SIMOR. That is, the entire pattern across the membrane reflects the influence of a single input.

Adaptation usually requires only a small number of parameters to transform a local pattern to the next as long as the transformation is deterministic, as a recognizable pattern of geometric coherence would be. Therefore, with a little effort following the temporal pattern, the auditory processor is again free to follow information exhibited in the frequency channel along the place direction in the space-time pattern. There are at least two kinds of adaptation in evidence. One involves scaling the pattern in the amplitude dimension, another involves time dilation (i.e., time-scale change). Amplitude scaling information comes naturally from local variations in the input's amplitude envelope. We intend to see how we may exploit this observation in the analysis and synthesis of timbre for data reduction in the physical domain in chapter III.

The kind of adaptation involving period dilation is necessary for the auditory processor to keep track of the movement of the pattern across the membrane, such as is the case with a signal that exhibits pitch glide or vibrato. The change must be slow enough in the sense of the duration of the auditory window for an interpolated pattern to emerge (as a function of time and place together) and for the ear to follow. Time scaling information comes naturally from local variations in the input's fundamental frequency trajectory or period trajectory. We intend to show how this information is important to provide the ear with "phase" clarity as well as data reduction in chapter III.

2.2.3.5 Extent of the Ear's Ability to Organize.

Finally, are there patterns that the logical part of the brain can organize that the ear-brain cannot? There seems to be evidence that the thought process and the hearing process did not evolve at the same rate. It seems that although the ear does

seem to organize information in the order of its importance and in some hierarchical sense, the pattern recognition task at the membrane space-time response level does not understand recursion, which appears to be a highly thought-oriented concept. For example, we might conjecture that a recursive pattern which does not display a periodically or locally periodically or slowly changing periodically organizable vibration pattern essentially appears as noise in the sense of white noise, where the notion of periodicity is flatly contradicted. In other words, we want to know if periodicity in the above modified sense is a fundamental structural element of the physical stimulus of a musical timbre. So we propose to study the behavior of pulse pair sequences of increasing length and complexity in the sense of their organizability in the temporal geometric sense. If we make each sequence to be locally simple and recognizable as pulse pairs of either homopolarity or antipolarity (everywhere in the sequence), but nowhere periodic in any local sense as described above, and if these sequences are conceptually simply organized such as by some simple recurrent equations, would we expect the timbre to be increasingly noisy and unpleasant?

We propose that the Rudin-Shapiro polynomials can be made to satisfy our curiosity. In chapter III, we will show how we prepare the stimulus sequence, what it sounds like, and how it appears to answer our questions as stated above.

2.2.3.6 A Notion of Musical Timbre—A Summary.

In summary, we have reached the conclusion that there are some structural elements in the class of musical timbres. One of these is the periodic structure or its generalized version under the notion of adaptation in the amplitude dimension and the time dimension, both in the form of scaling. Adaptation by scaling is possible by means of the organizing elements in the signal. One of these organizing elements is the amplitude envelope and another is the period trajectory.

2.3 Constitution of Timbre.

2.3.1 "Frames" as Constituents.

So we have argued that musical timbre possesses structural elements that help the ear to organize the complex membrane response pattern in space-time, and they include the period trajectory and the amplitude envelope in the acoustic signal. Through these organizing elements, the dimensionality of the timbre universe is drastically reduced to some relatively small number that defines the vibration form of a period, quasi-period, or local "period" in a typically non-stationary acoustic signal, through the strategy of adaptation. A consequence of the organization is the emergence of the notion of frame.

2.3.1.1 Notion of "Frames."

A frame can be initially thought of as an elemental vibration form that is known as a period, quasi-period, or local "period" in many of the acoustic signals we discussed above. It is the smallest unit of the vibration pattern in the waveform that appears to repeat itself in some approximate sense which can be made to appear more exact by scaling the amplitude-time grid over which the pattern varies according to the period trajectory and the amplitude envelope. In some cases, the adaptation resulting from such scaling will induce perceptually indistinguishable timbres from the one which involves no approximation. When this is the case, we assert that the ear prefers to approximate so that it is free to follow more carefully the detail of the vibration form in the frame. And the ear enjoys the added freedom to explore and discover the structure in a more manageable scope.

When the period organizing element does not exist, the ear must follow the long span of fluctuation in time, busily expecting new information or innovation while at the same time failing to organize too closely the information contained in the segment. The case in point is the class of noisy waveforms including the fricatives of speech. If the brain has communicated to the ear that the sound images are not really information worth decoding, it will ignore them or avoid them. It will ignore what is tolerable and avoid what exceeds a certain loudness level. (In speech or music, noise streams are typically much softer than the resonant elements which correspond to our organizable patterns of vibration. And one reason for this is the total information rate is proportional to the product of the duration, the bandwidth, and the dynamic range of the amplitude or pressure.) But in speech communication

or in listening to a melody articulated by a flute, the ear has been instructed to listen for useful messages and to decode them. So the ear must continue to follow the noisy vibration pattern. At the same time, the dimensionality of such a segment is high because no periodic structure exists. Therefore if we accept the notion of musical timbre as discussed in 2.2.3, then it is not surprising to find fricatives as noisy in the sense that the ear cannot make very good sense of their vibration pattern.

But usually, they are also articulated at such a level that they provide not only contrast but suspense for the more organizable information to come. Therefore, from a communication viewpoint it makes sense not to have a string of fricatives running together. Such a string would not be intelligible and certainly not pleasant. The musicality of a fricative preceding a vowel is like a dissonance resolving into a consonance. The larger principle here is the notion of frame.

2.3.1.2 Definition of Frame.

As we have indicated in the beginning of this section, a frame represents the smallest unit of a vibration pattern which "repeats" itself in some adaptational sense. In other words, it is a generalization of the notion of a period. But we would also like to see a frame represent some fundamental unit of organization that would include the noisy waveform patterns characteristic of such sounds as fricatives. After all, they appear in conjunction with many voiced productions. These patterns are not themselves easily organized, but there is always a transition when their function is not pure noise. And there is always a first "period" in some adaptive sense in the transitions.

We want to partition a waveform into frames and nothing else. In particular, the partition is to be such that adjacent frames are maximally similar. The rationale rests on the notion of organization in the perceptual sense, in the sense that the ear can maximally predict incoming signal behavior from context. Frames are to be some smallest ("nontrivial") units of organization which together constitute a complete waveform much as cells constitute an organism. From the perceptual viewpoint, a frame corresponds to some local timbre or microtimbre. And we will not try to make the distinction between the physical and the perceptual object (unless confusion could arise).

We should summarize by requiring that (1) a period be a frame; (2) a period

by adaptation (scaling in time or amplitude) be a frame; (3) a noisy segment be a frame (noisy to the point of not being organizable); (4) a segment from which neighboring segments can be obtained by some kind of transformation be a frame; (5) a segment which can be obtained from its adjacent neighboring segments by some form of interpolation be a frame; and (6) a smallest perceptually organizable unit of vibration be a frame. In general, we will define a frame to be a *maximal* segment of vibration without "repeating" subsegments, repeating in the organizational sense, consistent with the six conditions above; the frame boundaries are to be defined in such a way that neighboring frames are to be maximally similar in their patterns of vibration. This will be our definition of frame. (See the figures in appendix B.)

We of course will not analyze something that is not analyze able. It makes sense to define such a segment (which cannot be analyzed) as a fundamental unit. Therefore the generalized definition of a frame is both logical and useful.

Furthermore, when we must analyze a sound wave that is a superposition of several sources such as that of a chord, the frame notion provides a more economic set of data. See for example the many waveform examples of chords in Moorer [Moorer, 1975].

2.3.1.3 Consequences of the Frame Notion.

These are a few obvious consequences of the notion of frame:

(1) Timbre is but a dynamic evolution of the frames. (According to Schaeffer, timbre is a dynamic evolution of the harmonics together with the evolution of the acoustic energy, or what he calls the general dynamics.)

(2) The transition between frames within a timbre may reveal how interpolation (that is, the smooth transition from one timbre to another) may be realized.

(3) In particular, the transition from noise to periodic waveforms found in many speech timbres may reveal the secret behind the interpolation between noise and musical timbre.

(4) A frame may be a distinctive feature if other frames can be extrapolated from it or interpolated between it and another frame. We will call such frames critical frames or *breakframes*. We will consider a timbre as described by a list of breakframes together with an algorithm for recovering the intermediate frames from the breakframes. (The term "breakframes"

is formed by analogy with "breakpoints"—the points which determine the line segments in a piecewise linear function.)

(5) A frame that represents a period of a periodic waveform represents all the pattern or data the ear needs to have (although it must still repeat a few times—say three or five—within the auditory window of the fundamental frequency of the waveform to assure the ear that it is indeed the correct period before the ear is free to rest or tend to other more interesting tasks).

(6) The emergence of a critical frame would also mean that the new information or innovation would cause the auditory processor to make an update.

(7) The algorithm that determines the frame transformation must ultimately come from the content of the breakframe and the dynamics or the organizing elements of the sound.

(8) The timbre of a sound is thus in general described by three dynamic features in the acoustic signal: the amplitude envelope, the period trajectory, and the list of breakframes.

2.3.2 The Amplitude Envelope and Period Trajectory as Structural Constituents.

Although the entire waveform can be described by a concatenation of frames as defined in 2.3.1, not every frame is equally important on either a perceptual basis or an analytical basis. This is because:

(1) Many timbres involve fairly stable acoustic features. One form of stability is a periodic structure. A periodic structure is not limited to a finite harmonic series, i.e., a finite sum of sine waves. It can include, for example, a frequency-modulated signal of the form $\cos(\omega_c t + A_m \sin \omega_m t)$ as long as ω_c and ω_m are commensurate. For example, if $\omega_c = 2\pi \cdot 200\text{Hz}$ and $\omega_m = 2\pi \cdot 400\text{Hz}$, then a periodic structure exists with a period of five milliseconds. A different periodic structure exists with the same period of five milliseconds if the values of ω_c and ω_m are interchanged. Note that a periodic form arising from frequency modulation in general cannot be expressed as a *finite* harmonic series.

(2) Another form of stability is a slowly varying "periodic" structure. In

this case, the waveform is only approximately periodic even on a local basis. But as long as the "periodic" structure is such that efficient energy transfer is possible for the fibers whose characteristic frequencies agree with the spectral content of the signal, locally over their respective auditory windows, the space-time membrane response pattern will be stable.

(i) One such class of waveforms involves only slow changes in the fundamental period.

(ii) Another involves amplitude envelope changes, such as the exponential decay of a periodic structure.

In the second case, if $f(t) = e^{-\lambda t}g(t)$ where $g(t)$ is strictly periodic with period T, then $f(t+T) = e^{-\lambda T}f(t)$, which might be expressed by saying that f is periodic up to a constant scale factor of $e^{-\lambda T}$ everywhere with the same period T.

(3) Still another form of stability comes from a smooth transition from one locally periodic form to another with the transition specifiable by some form of interpolation between the two where the interpolation is not necessarily linear.

(4) Finally, a form of stability arises from general scale invariance in either the amplitude or time direction (or both) on a frame-to-frame basis. Scale invariance is accomplished by the information derived from the period trajectory or amplitude envelope. Of course, scale invariance in the amplitude envelope includes the case of the exponentially decaying periodic structure. And the scale invariance in the period trajectory includes the case of frequency modulation. (The temporal scale invariance statement in the case of inharmonic frequency modulation depends on the technical hypothesis that

$$|A_m| < |\frac{\omega_c}{\omega_m}|$$

[Hitt, private communication].)

In summary, for the waveform of a timbre, only a selected set of frames are important from the informational viewpoint. The rest can be recovered from this set of frames together with the period function and often with the amplitude envelope (or both). Therefore, the amplitude envelope and period trajectory form a kind of structural constituents for the perception of timbre.

2.3.3 The Kinematic Nature of Timbre.

From the perceptual viewpoint, the waveform that gives rise to a certain timbre is not merely a concatenation of frames. In other words, a timbre is not perceived as a concatenation of timbre frames in some static sense. The timbre of a sound is a series of local timbre frames in motion, much like animation or the frames in a movie. Of course, the kinematic nature of timbre originates in the fact that sound waves in general (except pulse-like waveforms) occur in continuous fluid motion and the fact that the auditory window is continuously sliding to track the waves. There are at least three consequences of the kinematic nature of timbre:

(1) A timbre frame is not perceived in isolation, even though one frame contains all the information of a periodic or quasi-periodic (in the sense of adaptation) signal. Timbre frames are perceived as some neighborhood of local timbre *in motion*. A good example of this is the transitions between consonants and vowels in continuous speech perception. An individual frame does not have much perceptual significance. And if isolated, it is probably not recognized unless the listener is trained in the spirit of Schaeffer's *Solfege*. Note also that such a neighborhood in motion represents formation of a distinctive feature in perception dictated by movement. A visual analogue is flapping a wing, for example. The word "flapping" is a description of such a kinematic feature, although each intermediate position must be equally real in the objective sense.

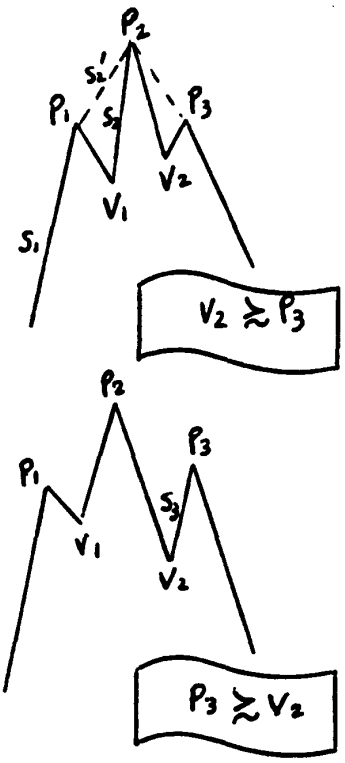
(2) From frame to frame, continuity in amplitude and rate of change of amplitude should be preserved at least to within the perceptual grid of resolution in amplitude and time. Otherwise, extraneous timbres in various forms of clicks will result. Of course, the perceptual grid is a function of context. If the waveform has been fast changing, the tolerance for discontinuity is larger. For example, experience shows that discontinuity at the peak of a vibration pattern whose spectral content is concentrated towards the low end (thus the peak is a relatively smooth and round one), like that of a human voice, is more noticeable than if it occurs at the peak with a vibration pattern whose spectral content is concentrated on the high end, i.e., one which consists of many fast oscillating peaks in one period, as in a marimba.

Also, a discontinuity occurring in a fast rising attack is more tolerable than one that occurs in a smooth decay. The contextural dependence is apparently a phenomenon of the adaptation and update characteristics of the auditory processor. The latter acting as a good-faith message receiver must treat the problem of message decoding as a statistical problem, prejudged only by context. Context, in coordination with adaptation, serves the function of pattern organization as well as improvement of the dynamic range of signal detection. As far as update is concerned, if a discontinuity comes amidst a slow information flow, then the former would trigger an update because innovation inherent in the signal and noise (or error) are one and the same thing as triggers for an update.

(3) It is therefore reasonable to assume that kinematic features in the acoustic waveform may be perceptually more important than their geometric (static) interpretation alone. For example, the rushes and slowings down in a timbre that give rise to a vibrato, produce a more pronounced timbral effect than the non-vibrato background context (that precedes the vibrato).

In this case, it is the movement rather than the extremal positions (in frequency) which is more profoundly perceived. (It is known that the listener recognizes the pitch fluctuation (and accompanying form fluctuation) but does not recognize the high and low of the pitches, but rather some average.) A similar observation applies to tremelo.

Therefore, a feature description based on perception of a static visual field may not necessarily provide comparable results in feature description for timbre.



We refer to the figure to the left. For instance, the first valley, v_1, may not be so important auditorally as the first peak, p_1, because the former does not represent the movement formed by the first slope, s_1, and the second slope, s_2, but the latter does. Although the slope change from s_2 to s_2' may be noticeable. But it may be a subtle difference. On the other hand, the second valley, v_2, may or may not be more important than the third peak, p_3, depending on the jump from the valley to the peak. If it is smaller than the context dependent perceptual grid (depending on the previous jump from v_1 to p_2 and on the position of p_2), then v_2 may be more important than p_3 because it represents part of the downward movement. Otherwise, p_3 could be more important as the jump from v_2 to p_3 represents an innovation.

In summary, the kinematic nature of the acoustic waveform may play a larger role than previously thought under the Fourier notion of sound analysis, because Fourier analysis is essentially a static notion. Note that one can Fourier analyze the architecture of St. Paul's Cathedral and obtain beautiful harmonics, amplitudes, and phases. While it may or may not have something to do with visual perception, it certainly is a different object of application than to sound waves.

2.3.4 A Dynamic Representation of the Constitution of Timbre.

When we talk about a timbre as being constituted by critical frames or break-frames, and the amplitude envelope and period trajectory as structural constituents, we mean that the intermediate frames can be recovered from the critical frames and the structural constituents. From a kinematic viewpoint, we can consider the critical frames as geometric objects, being driven dynamically across time between two successive frames by the amplitude envelope and the period trajectory that shape them and are connected by some kinematic constraint as described previously (i.e., continuity, within the context-dependent perceptual grid, of the amplitude envelope and its slope). This description takes advantage of a number of perceptual features

observed before. In passing, we note that Huggins [Huggins, 1952] suggested the utility of decomposing stimulus waveform properties into those which are "structural" or acoustic-cavity-related, and those which pertain to the excitation of the source. However, he did not attempt to justify why the ear should perceive a sound according to this type of property decomposition, other than a well adapted receiver, due to pressures of evolution, should know its sources well.

The structural constituents are organizing elements provided to the ear by the signal. They are perceptually recognized distinctive features. Furthermore, we assert that they are timbral distinctive features. For the amplitude envelope, as a timbral distinctive feature, direct evidence comes from:

(1) Schaeffer's laws of timbre perception.

(2) Schouten and Erickson's list of timbre features.

Indirect evidence comes from:

(3) Wedin and Goude's multidimensional scaling dimension identification (see [Wedin and Goude, 1972]).

(4) Charbonneau's data reduction experiment in which harmonic amplitude envelopes in a phase-vocoder analysis are found to move in synchrony suggesting that the gross amplitude envelope is behind the coherent movement.

For the period trajectory as a timbral distinctive feature, we have only indirect information from the literature. The fact that vibrato and pitch glide are often considered as timbre effects (see [Schouten, 1968], [Erickson, 1975], and [Schaeffer, 1966] for example) and the fact that their physical correlates are part of the period trajectory suggests strongly that the period trajectory itself is a physical correlate of timbre, although parts of it can become more identifiable timbre features because of their more highly "visible" collective identifiability, such as vibrato (see 2.3.2, for example).

Of course, periodicity has always been an attribute of harmonic signals and musical timbre has long been identified with harmonic or periodic signals. The degree of harmonicity often measures some degree of musicality. The noise quality of the timbre of a bandpassed white noise is directly related to the width of the filter, and inversely related to the pitch strength. Charbonneau's data reduction experiment points to the unique role the fundamental frequency trajectory plays

in the short-time Fourier transform description of timbre. He observes that all upper partials' frequency trajectories vary in ratios proportional to the fundamental trajectory. In addition, we have argued that it is a structural element the signal provides to the ear for its organization of the SIMOR pattern. And we know that it is a perceptually recognized distinctive feature. Here we assert that such a structural element *must be a timbral distinctive feature.* Now, we argue that a timbre can be described by the dynamic triple consisting of the amplitude envelope, the period trajectory, and the list of breakframes, $(\mathcal{A}, \mathcal{P}, \{F_k\}_{k=1}^{n})$. Such a description forms a dynamic representation of the constituents of a timbre.

2.4 A Hierarchical Organization of Timbre Features.

In 2.2, we have arrived at the conclusion that the period trajectory and the amplitude envelope are two structurally important organizing elements. Periodicity provides the ear the means to organize a large volume of data into a relatively small one. The amplitude envelope binds the single input multiple output response (SIMOR) through its own low-pass characteristics and the low-pass characteristics of the fiber (recall that the frequency response of a fiber is not symmetric and is large enough to "pass" frequencies lower than the characteristic frequency).

Certainly, the amplitude envelopes of the SIMOR are not identical to that of the signal, but they are determined by it. They are not identical with each other either, but they exhibit a smooth variation across the membrane that approaches the signal amplitude envelope at the high frequency end by virtue of their short damping constants. The low pass characteristics of the fibers provide an element of redundancy in the SIMOR. From the viewpoint of pattern recognition, redundancy provides robustness and stability to the message carried by the signal. (Robustness refers to the ear's ability to preserve the integrity of the signal in the presence of spurious noise; stability refers to the ear's ability to follow slowly varying temporal patterns in spite of noise).

Furthermore, the amplitude envelope provides local organizing elements to the ear when the signal is not strictly periodic. Similarly, a slowly varying period trajectory does the same thing. By scaling either in the amplitude dimension or in the time dimension appropriately (according to the local organizing elements from the amplitude envelope and the period trajectory), the ear is able to adapt the state of the current vibration form to the new one over some local time. And by so doing, the ear is again able to reduce a large volume of data into a relatively small one, not too much larger than that of a truly periodic signal. The data reduction made possible by the amplitude envelope and the period trajectory allow the ear to focus on features of more local levels such as the formant distribution or the spectral envelope.

2.4.1 Signal Organizing Elements as Processes.

The amplitude envelope and the period trajectory are processes. They are processes not only because they are in general time-varying but also because the coherence time or window of coherence is essentially determined by the auditory

windows. By the window of coherence, we mean the time span about any given instant that determines the extent over which the waveform of the signal is correlated or coherent in the eyes of the ear. The auditory windows of the fibers provide a local view of the correlatedness or coherence of the signal to the ear. The signal may be correlated over a long stretch of time as in a periodic signal. But to a fiber, the signal is correlated enough to produce a strong response if it is correlated within the fiber's auditory window. If a signal is locally periodic over a time comparable with the auditory window time of the fundamental frequency, the SIMOR will appear correlated in the sense that for each fiber which is not stimulated strongly by more than one harmonic partial, there are the same number (because of the constant-Q nature of the basilar membrane) of oscillations enveloped within the fiber's own damping window, even though their relative strengths across the place dimension may vary quite arbitrarily. Even if the signal is inharmonic, if the rate at which upper partials get out of phase with the fundamental is slow enough, a sense of coherence remains over the auditory window of the fundamental across the membrane.

2.4.2 Feature Formation.

The important point here concerns the nature of the amplitude envelope and the period trajectory as organizing elements with respect to the auditory windows. They are processes, i.e., *organizing elements in progress*. One of the consequences of these organizing processes is that the local elements that constitute the process provide the basis for hierarchical feature formation. That is, a *timbral feature derived from the amplitude envelope* may be considered as a *composition of neighboring subfeatures*. For instance, local amplitude behavior may form a line segment of the amplitude envelope with a certain slope. And several such line segments of different slopes may form what constitutes the attack of the sound. Similarly, the two characteristic decay segments of a piano tone may be considered as subfeatures that form the larger decay feature [Weinrich, 1979]. And the attack and decay of percussive instrument sounds together form a larger feature of temporal form. By the same token, the amplitude envelope segment that gives rise to tremolo via amplitude modulation may be considered as a composition of subfeatures of the high amplitude and low amplitude appearing in alternation. The ear, seeing the obvious organization, would tend to form a permanent template, regarding these highs and

lows together as one single distinctive feature. By the same argument, pitch glide and vibrato are distinctive timbral features composed of more elementary features in the period trajectory. The same reasoning can be applied to the role of the "blip" with respect to the perception of the trumpet attack or the "twang" with respect to the perception of the Indian sitar, the Chinese pipa, the Spanish guitar, or even the piano.

2.4.3 Perceptual Importance Trees for the Organizing Process.

In general, for an arbitrary timbre, we don't know which feature is more important than which other and what features form superfeatures. We need to do experiments to find out. But we can formalize these features *within each dynamic entity,* such as the amplitude envelope and the period trajectory, into a tree organized according to their importance in one direction and preserving their temporal ordinality in another. From the pattern recognition viewpoint, feature composition by means of hierarchy is an important data-reduction strategy since we don't need to search all the features at once and can stop whenever finer discrimination is unnecessary. From a data processing viewpoint, a tree organization always facilitates the search time on a statistical basis. Therefore, what we have done is essentially an attempt to concretize the speculation Erickson made concerning the way we organize timbre.

2.4.4 Importance Tree for the Breakframes.

So far, we have discussed only the hierarchical organizational nature of the amplitude envelope and the period trajectory. But there is no reason why this should not be applicable to the local timbre frame or the breakframes. A vowel may be considered as a composition of certain similar or even identical vibration forms in succession. A diphthong may be considered as a composition of a vowel followed by a transition followed by another vowel. A fricative may be considered as a composition of certain irregular forms with as yet undetermined transitions before and after. Vibrato and tremolo are distinctive features composed of alternating local vibration forms, driven by period changes or amplitude changes. A concept we call alternating timbre is another timbral feature derived solely from alternating local vibration forms *without* period length changes or amplitude changes. The attack timbre which Schaeffer reports as being distinct from the steepness of attack may be considered as a composition of local frames in the attack.

In general, if we treat timbre as composed of the dynamic triple $(\mathcal{A}, \mathcal{P}, \{F_k\})$, each dynamic entity can be organized in a perceptual importance tree—all of their features being in order of their importance. And timbre modification and interpolation can be made systematically on each feature to the level of importance desired based on one control parameter, such as the interpolation index. Alternatively, each feature can be modified to an arbitrary extent. Furthermore, data reduction may be accomplished by tree-trimming based on a perceptual criterion.

2.5 A Dynamical Relational Description of Timbre.

2.5.1 Composition of Features.

We have described timbre as local timbre frames in motion (see 2.3). We have described timbre as breakframes driven by the structural elements of the amplitude envelope and the period trajectory, i.e., as the dynamic triple $(\mathcal{A}, \mathcal{P}, \{F_k\})$. We have described each element in the dynamic triple as an importance tree (see 2.3). We have described the importance tree as a hierarchical organization of features where features are compositions of subfeatures and they in turn form superfeatures; each tree is successively analyzed into subtrees of features (see 2.4). Conversely, we can think of timbre as a composition of the three dynamic classes of features, each class of features is a feature tree which is a composition of features of the same class, i.e., it is a composition of its subfeatures defined in a recursive manner until the level where either the subfeature is too short compared with the auditory window of the appropriate fiber (in the sense of where the sound spectral energy concentrates) to be perceptually distinctive, or no further information exists to cause an update on the particular class of features the tree represents. Notice that for the class of frames, in addition to the feature tree, each breakframe that constitutes a feature on the lowest level similarly may be analyzed into a feature tree that reflects the local pattern of variation. If such a frame represents part of a quasi-periodic pattern, then it makes sense to have a feature tree in the Fourier domain where a spectral envelope is composed of formants and each formant composed of neighboring Fourier components, and so on.

The notion of a line segment approximation to an amplitude envelope or a frequency trajectory or even a waveform, as proposed in Grey's work, can be considered as a process of feature formation on the lowest level. Each irreducible line segment represents the smallest feature the ear can perceive or recognize. And then, several neighboring line segments, by virtue of the similarity in their trends, i.e., their slopes, form a larger feature. For example, in the marimba amplitude envelope, one can detect several smaller line segments governing the initial increase in the acoustic energy before a steep rise sets in. And in the middle of the steep rise, there is again a group of small line segments (smaller than the "period length" of the tone) which together provide the sharp sensation which is often described as

"noisy" [Serra, 1984] in the attack. (That is, without the non-uniform rise over a "period," the "noisy" sensation will be found missing.)

Furthermore, the transition from the initial low energy onset to the steep rise characterizes a superfeature formation in the form of the "twang"-like sound we hear so typically in percussive sounds. Finally, we can think of the entire attack envelope (albeit short for a percussive tone like that of a marimba) as being a composition of the features described above.

The timbre frames in the attack of the marimba can be described in similar terms. First, the vibration pattern of the first three frames is highly deterministic without further information from the amplitude envelope. In other words, knowing the first two frames determines the third without incurring more than a two percent error between the sample amplitudes from t to $t+P$ everywhere in the period, where P is the length of the "period," which is constant. This is not visually obvious from inspection of the waveform alone because the first frame is dominated by the fourth harmonic (actually the second normal mode of the marimba bar) but the third frame is already dominated by the ninth and tenth harmonics (actually the third and fourth normal modes, which are known to be very close, but of very different origin—see Benade [Benade, 1976]). So, these frames together by virtue of their kinematic nature describe a local timbre feature.

Next, the series of frames that follow exhibit a highly non-uniform growth whose behavior can be determined only with the aid of the amplitude envelope. The noisy character is exhibited in this highly non-stationary evolution of the frames. And perceptually, we expect them to form a feature.

Finally, the attack timbre can be thought of as a composition of the local timbral features described above and we believe it is perceived as such. Then the post-attack timbre may be thought of as composed of the local timbre corresponding to the disappearance of the third and fourth modes and the local timbre corresponding to the final disappearance of the fundamental mode. The attack timbre and the post-attack timbre are then features in the composition of the entire timbre of the marimba.

The entire marimba timbre can then be described as the composition of the growth/decay superfeatures as described above via the analysis of the amplitude envelope and the superfeature derived from the composition of local timbral features

that can be described by the breakframes or transitions between them. The period trajectory is constant, therefore it has only a trivial tree.

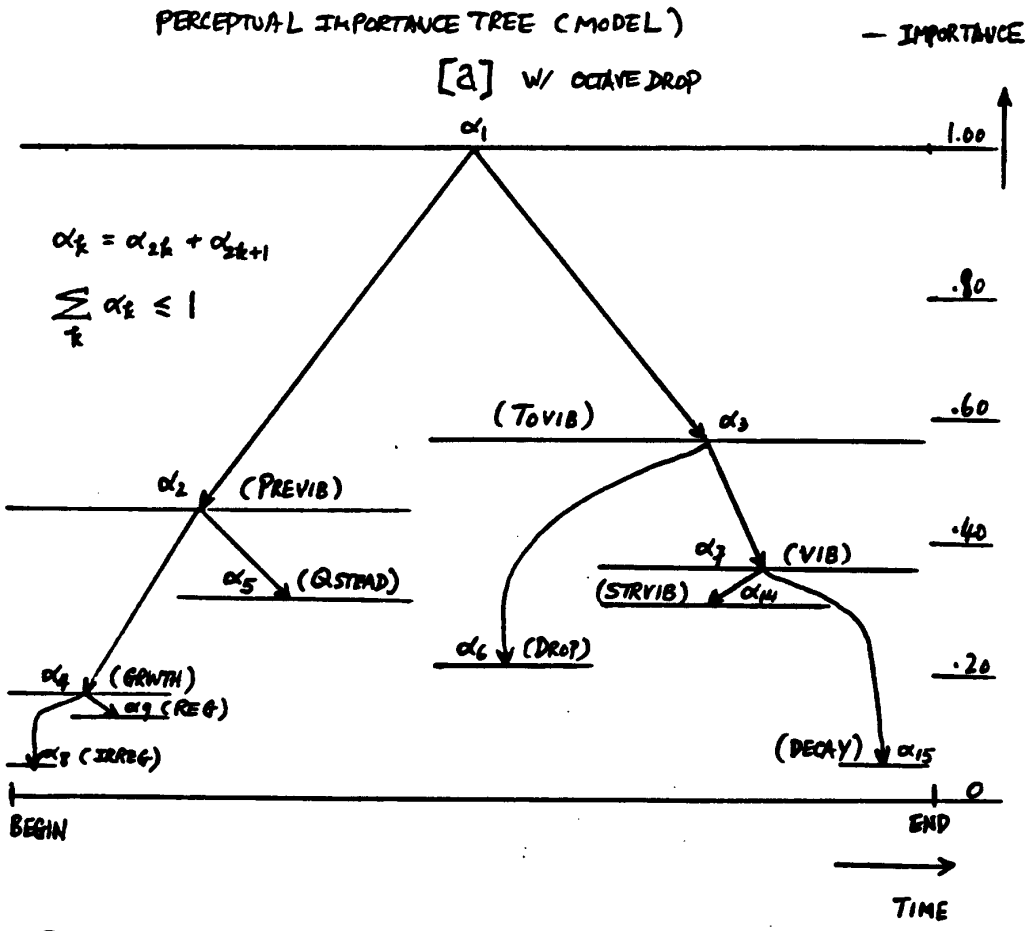
We can apply the same principles to describe the timbre of the marimba tone played backwards and see why it sounds the way it does. One important point to note is that the elemental features are no longer the same. For example, the noisiness of the original attack does not exist in the tail of the backwards run version of the marimba waveform. This is because the fiber response within its auditory window (recall that the fiber response is the convolution of the excitation function and the impulse response of the fiber, and that the latter has an exponential decay that determines its "auditory window") is different for an upward fluctuation than for a downward one. Furthermore, the temporal sequence of auditory contexts is also different, implying different auditory expectations and hence different auditory perceptions. The role of causality in the temporal sense as a universal principle thus underscores the difference between vision and audition. We are much more able to recognize that a man standing upsidedown is a man than to recognize a sound played backwards as the original sound played backwards.

We can equally well apply the same description to the timbre of an octave drop pronouncement of the vowel /a/. The timbral composition can be essentially described as follows:

The spectral evolution corresponding to the evolution of frames of vibration pattern is essentially composed of the initial non-steady timbre and the final vibrato timbre an octave lower. One might even argue that the final vibrato timbre is the most distinguishing timbral feature of the entire timbre, as Schaeffer suggests as a general rule (see 1.7) for long sustained timbres with weak growth characteristics. Then the initial quasi-steady timbre can be actually decomposed into the initial soft and breathy opening timbre and the quasi-steady-state timbre quite characteristic of a young voice. And the final vibrato timbre can be decomposed into the transition from the high-pitched region to the low and the vibrato region. (The transition region involves gradually increasing vibrato features.)

Of course, the vibrato region exhibits a more perceptually stable characteristic, therefore it is perceptually more important. Naturally, further decomposition of the vibrato region is possible and even necessary for the purpose of analysis, synthesis, and modification. But each successive step of decomposition represents a shorter

perceptual distance that one can advance as we approach a full description of the timbre in the sense of perceptual indistinguishability.



PREVIB ≡ SEGMENT BEFORE VIBRATO ;

TOVIB ≡ SEGMENT INCLUDING OCTAVE DROP FOLLOWED BY VIBRATO ;

QSTEAD ≡ QUASI-STEADY SEGMENT ;

VIB ≡ VIBRATO SEGMENT ;

STRVIB ≡ STRONG VIBRATO SEGMENT ;

DROP ≡ OCTAVE TRANSITION SEGMENT ;

GRWTH ≡ GROWTH SEGMENT ; REG ≡ THE REGULAR PART ; IRREG ≡ THE IRREGULAR PART ;

DECAY ≡ DECAY SEGMENT ;

figure 2.5.2 (a)

2.5.2 Absolute Description of Timbre by Means of Importance Trees of Timbral Features.

The issue of perceptual distance brings us to think of the perceptual importance tree organization of timbral features as not only a qualitative description but also a quantitative one *if* we provide the importance dimension not only ordinal character but metric character as well. (The temporal dimension is automatically metric if so desired.) We can assign a metric measure of importance, say between zero and one. Each subtree is given a fractional value such that they add up to the value of the tree they descend from. For example, the tree of the entire timbre assumes a value of one. Then maybe the first (in importance) subtree we designate as `ToVib`; we give it a value of say .55. The subtree right before `ToVib` we designate as `PreVib`, and we give it a value of .45. [Of course it may turn out that `PreVib` is actually more important because of temporal context. In that case $\alpha_{PreVib} := \alpha_2$, the value for `PreVib` may perhaps be .52, while $\alpha_{ToVib} := \alpha_3 = .48$. The experimental design for actually determining the values will be discussed in chapter III. It does not necessarily involve cutting the sound into pieces.] Using the indexing of figure 2.5.2 (a), we have $\alpha_j = \alpha_{2j} + \alpha_{2j+1}$. Thus, if we obliterate the irregular feature at node eight, then the entire sound would have a perceptual importance of $1 - \alpha_8 = .96$. And if we obliterate further the transition in `ToVib`, i.e., we abridge the drop by a much briefer or even abrupt transition, then perhaps $\alpha_6 = .2$ is removed from the perceptual importance and the synthesized sound would amount to only .76 of the original.

We can draw a similar tree for the marimba sound (see figure 2.5.2 (b)). In this case, perhaps the attack subtree designated by α_2 may assume a value of .8, and the post-attack tree, .2. The subtree representing the region before the steep rise in the attack subtree might take a value of .1, with the steep rise subtree taking .7, and so forth. Notice that the noisy character reflecting the non-uniform growth within the time of a "period" between two steeply rising regions may assume so much importance that it is represented with $\alpha_{21} = .5 > \alpha_{20}$ to the left and α_{11} to the right.

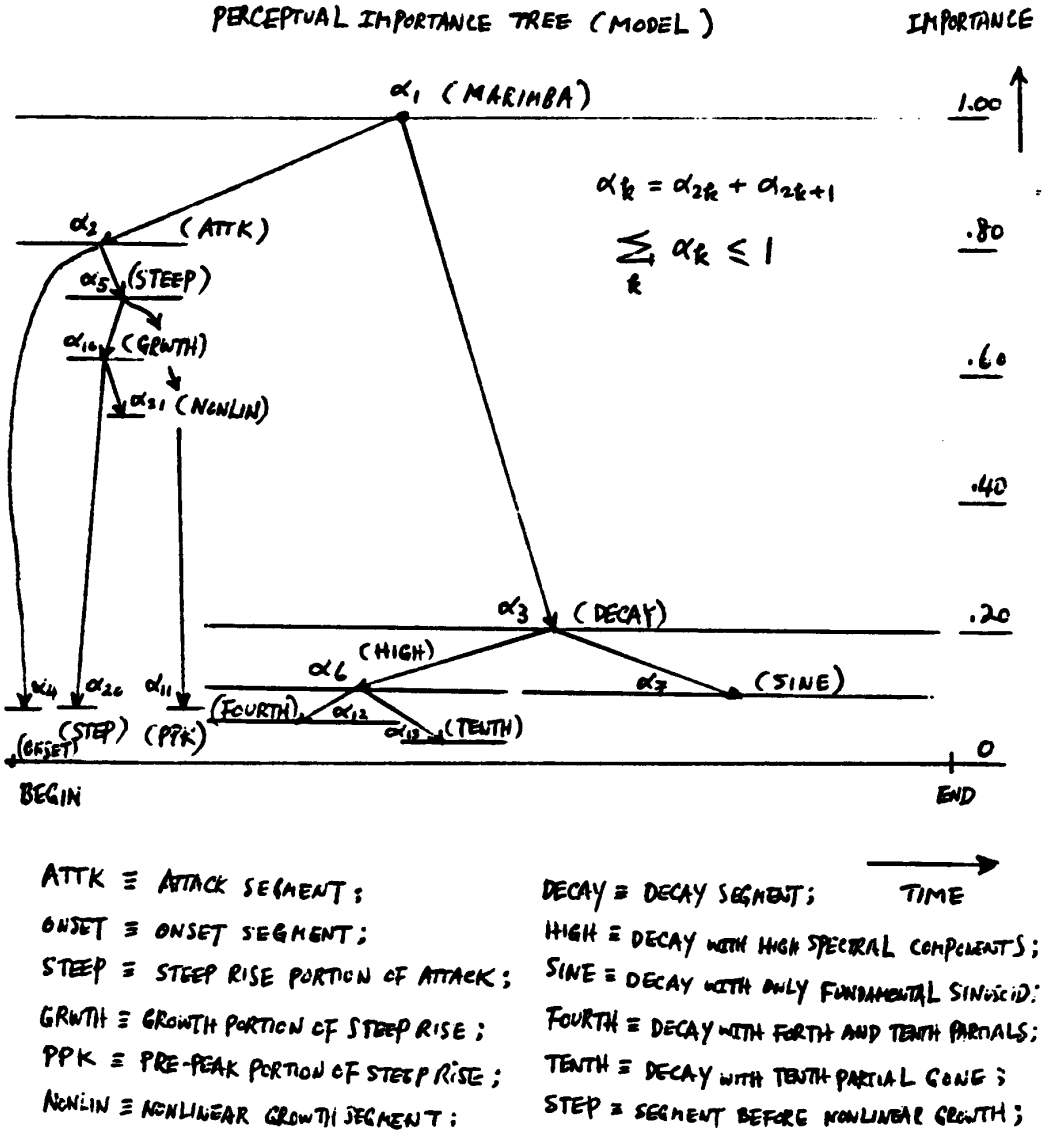


figure 2.5.2 (b)

2.5.3 Timbral Relations and Interpolation.

The trees provide an absolute description of a timbre from the notion of feature composition. This description further provides a dynamical description among relations of timbre. For example, the most important feature of the marimba timbre occurs in the first five percent of the sound duration, while that of the /a/ with an octave drop and trailing vibrato might have the most important feature in the last half of the sound. This is of course observed by Schaeffer and elucidated in one of his timbre rules. Interpolation between a pair of timbres can then be done according to the pair of importance trees on each component of the dynamic triples $(\mathcal{A}_1, \mathcal{P}_1, \{F_{1,k}\})$ and $(\mathcal{A}_2, \mathcal{P}_2, \{F_{2,k}\})$, on a level-by-level basis scaled both in the importance dimension and the temporal dimension (independently) driven by the index of interpolation. Without the tree, we would probably interpolate between fixed features such as the attacks as Grey prescribes. But if Schaeffer is correct, then Grey's prescription may have only limited application.

The tree description is appropriate both in the perceptual space and the control space. In either case, a complete transformation of timbre features from one timbre to another can be accomplished with something analogous to vector addition on the three dynamic components of the triple. This model is known as the parallelopiped model of timbre interpolation, first introduced by the author in 1986 [Lo, 1986]. The vectorial character of the control space is straightforward, but that in the perceptual space can only be surmised as upon some "vectorial" translation from one timbre point to another at this state of our knowledge. Even so, it is meant to apply only in a local basis. We do not know whether the local space is Euclidean or not. But if the ear, as a pattern recognition device, performs the task of dimensionality compacting, then continuous mapping from the connected control space to a lower dimensional perceptual space implies the perceptual subspace, defined by interpolation, should be connected also. (Shannon extensively discusses the issues of continuity for mapping between spaces of different dimensionality [Shannon, C., 1949].) In this sense, timbre may be interpolable if the ear reduces dimensionality, as most suspect. But it certainly says nothing about the existence of a vector space structure in the Euclidean sense in the perceptual space.

2.5.4 Some Thoughts about Dimension of Timbre Space.

Perhaps first we should remark that even the title of this section is misleading: dimension is a *local* concept, varying* from point to point (see figure 2.5.4), e.g., the dimension at x_1 is one and the dimension at x_2 is two. In fact, we have brought up the local nature of dimensionality of timbre space in 2.1.2.10.

So, let's put aside that point and ask the question: What is the dimension of timbre space?

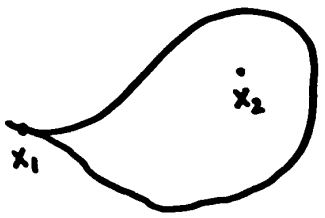


figure 2.5.4

Before we answer *that* question, we ask, why do we want to know? If we just want to get some idea about the relations the sense data have to each other, multidimensional scaling (MDS) of Kruskal might give an estimate of which $\mathbf{R}^k$ the data can be placed in most easily or justifiably with smallest possible k. (Here $\mathbf{R}$ stands for the real numbers, so that $\mathbf{R}^k$ stands for k-dimensional Euclidean space.)

If, however, we want a *constructive* idea of the dimension, in the sense of the number of parameters necessary to generate points in timbre space (so that the parameters lie in a control space), then we must have an actual method of associating a timbre to a given set of parameters. That is, in a constructive sense, it is not enough to maintain that timbers all live in a three dimensional space unless there is a method of generating a timbre given coordinates (x, y, z) or (r, ϑ, φ)—or telling unambiguously that there is no timbre at that site (because timbre space would be thought of as a *subset* of $\mathbf{R}^3$).

We should make a few more observations: any estimate of the size of timbre

---

* It may indeed be that the space is embedded in some larger space of *fixed* dimensionality. In the case of timbre, this fixed dimensionality might correspond to the dimensionality of the space the SIMOR pattern lies in, or at least the dimensionality of the space the signal lies in. On general principles, the former is greater than the latter, because of the multiplicity involved, although the actual dimensionality of the set of possible SIMOR patterns must be exactly the same as the dimensionality of the set of possible signals. And all of these dimensionalities are huge.

space necessarily gives only a *lower* bound (as more timbres are introduced, the dimension in any sense can only increase). Thus the error involved in a sixteen point MDS is that it at best gives us a very local estimate, and at worst it might be unreliable statistically.

Second, it is true that the control space may be bigger than the timbre space (as the real parameter ϑ ranges over $\mathbf{R}$, the infinite line, the image $e^{i\vartheta}$ traces out over and over the finite unit circle in the complex plane $\mathbf{C}$,—this is reminiscent of the familiar phenomenon of aliasing in digital sampling), but it would be surprising if the *dimension* of the control space were greater (this would mean that some of the parameters could (perceptually) be replaced by functions of some of the others), unless pattern organization takes place that maps the SIMOR pattern space into a smaller feature space.

Finally, we should note that not every three dimensional space is necessarily a Euclidean space, i.e., not every three dimensional space (three dimensional manifold, or three-fold) fits in $\mathbf{R}^3$ (e.g., the 3-sphere S_3 given by the locus of points in $\mathbf{R}^4$ satisfying

$$x_1^2 + x_2^2 + x_3^2 + x_4^2 = 1$$

doesn't).

It is therefore reasonable for us to advocate that until we have a much better understanding of timbre and timbral relationships, we should study timbre space from a local point of view, taking two or three timbres at a time and seeing how they are related to one another through exploration of listeners' perception of timbre sequences constructed by interpolation of their acoustic feature trees, as discussed previously. We will explore this issue in more detail in 3.15 in terms of the "vector addition" or parallelepiped model of timbre.

Chapter III: Analysis, Synthesis, and Tests of Theory

3.0 Introduction.

At the beginning of chapter I, we stated that the goal of this thesis is to formulate a dynamic theory of timbre in such a way that:

(1) Timbre is described in a way consistent with the dynamic character of the sound waves that enter the ear.

(2) A universal language is formed that enables us to describe a diverse collection of timbres.

(3) A timbral operating environment emerges as a result of (1) and (2) which provides more precise control over the timbres we want to generate and allows us to generate them efficiently.

By the end of chapter II, we developed a treatment of timbre that appears to be more consistent with the dynamic character of sounds as we perceive them and that permits us to view timbres and their relationships from a unified perspective. In particular, we have offered a precise formal description of timbre that articulates relationships among timbres in terms of the internal dynamics of their features.

In this chapter, we will develope an analysis and synthesis approach from this new treatment of timbre. In other words, we want to have an analysis program that will extract the features and the featural dynamics prescribed by the formal description. When we modify timbre, we want to modify these features and their dynamics, or, in short, the composition of timbral features. And when we synthesize a timbre, we want to synthesize the sound, or its equivalent sampled waveform, from these features and their composition. In a word, we are looking for a set of analysis/synthesis algorithms that is consistent with the auditory requirements for the perception of timbre and timbral relationships. Figure 3.0 shows a schematic of such an analysis/synthesis approach.

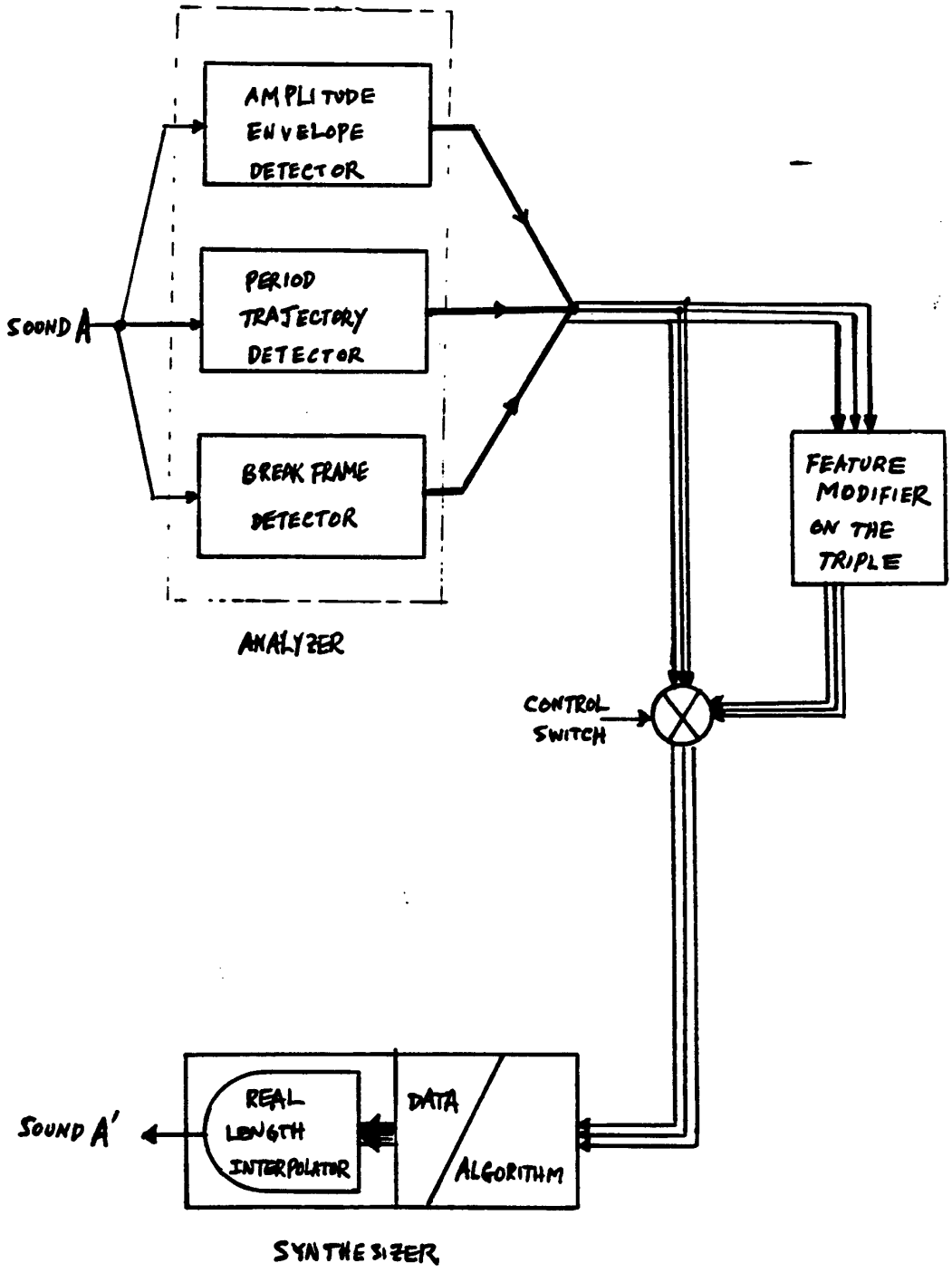

AMPLITUDE ENVELOPE DETECTOR
PERIOD TRAJECTORY DETECTOR
BREAK FRAME DETECTOR
SOUND A
ANALYZER
FEATURE MODIFIER ON THE TRIPLE
CONTROL SWITCH
SOUND A'
REAL LENGTH INTERPOLATOR
DATA
ALGORITHM
SYNTHESIZER

figure 3.0

3.1 Analysis/Synthesis Criteria.

The formal description of a timbre is its dynamic triple $(\mathcal{A}, \mathcal{P}, \{F_k\})$, or, specifically, the dynamic triple of the corresponding importance trees. Here, the acoustic waveform has amplitude envelope $\mathcal{A}$, period trajectory $\mathcal{P}$, and list of breakframes $\{F_k\}$. Thus the first requirement for the analysis of timbre is the extraction of these three dynamic quantities.

3.1.1 Analysis as Extraction of the Triples.

In chapter II, we argued from psychoacoustic principles that the amplitude envelope, the period trajectory, and the list of breakframes are fundamental acoustic elements for the perception and constitution of timbre. In other words, these acoustic elements are both necessary and sufficient to completely define a timbre, necessary in the sense of irreducibility, sufficient in the sense of completeness. These elements define a timbre in the sense of perception as well as of synthesis of the corresponding waveform. In other words, these elements are physical correlates of timbral features or distinctive acoustic features of timbre. When we perceive a timbre, we perceive a composition of these three dynamic entities.

When we perceive any of these entities, we are really perceiving, according to our proposition, the importance tree of acoustic features associated with the entity in question. From the viewpoint of analysis of timbre, we need to first obtain the whole of each of these three dynamic entities before we analyze them further into their respective importance trees. By the same token, we need complete information regarding each of these three dynamic entities before we can complete the synthesis of the timbre.

3.1.1.1 The Amplitude Envelope as a Fundamental Timbral Feature.

In chapter II, we argued from pattern recognition principles that the amplitude envelope serves as a structural constituent of timbre because of the organizing role it provides to the collective behavior of the highly redundant temporal pattern of the single input multiple output response (SIMOR) induced by the acoustic signal (which carries the amplitude envelope). It is so because of the broad-based response characteristics of the fibers to an arbitrary signal, especially over transient regimes and because of the asymmetric shape of the frequency response curves of the individual fibers that "pass" low frequencies coupled with the "low-pass" or slowly varying characteristic of the amplitude envelope.

As a result of the amplitude envelope's central role in organizing an acoustic signal's SIMOR pattern, amplitude changes in the vibration pattern from one "period" to the next or even more locally oriented amplitude structures (such as those in a fast attack) in the signal, will perceptually coalesce and become distinctive features for the timbre. Thus, the attack, the fine structure of the attack, the decay, the fine structure of the decay, tremolo, and other perceptually notable features derived from the amplitude envelope are perceived as distinctive timbral features.

Therefore, except in the case where it is flat, and hence lacking in new information or innovation, the amplitude envelope is in general perceived as an acoustic element that provides important information about timbral features. This view has in fact, in one respect or another, been advanced by Schaeffer, Erickson, Schouten, and others, and is strongly implied by Charbonneau's perceptual data reduction study (at least for the class of timbres he studied). It also agrees with the experience of many musicians as well as with the practice in many non-analysis based synthesis approaches. These approaches include Alles' synthesizer, Chowning's FM synthesis, and hence the Yamaha DX series. Even additive synthesis (non-analysis based) practices this approach. So do many less sophisticated synthesis techniques using specialized waveforms, such as pulse trains. Yet, the (gross) amplitude envelope has not been a part of the analysis parameter set for any existing sound analysis that is also aimed at resynthesis. These analysis techniques include Fourier-based methods such as the phase vocoder and various time-varying spectrum analyzers. They also include such signal analysis methods as the Wigner transform and Linear Prediction.

3.1.1.2 The Period Trajectory as a Fundamental Timbral Feature.

In chapter II, we argued from pattern recognition principles that the period trajectory serves as a structural constituent of timbre because it helps the ear to decide whether the single input multiple output response (SIMOR) pattern can be organized, to what degree it can be organized, and hence to what extent the timbre is considered musical. A periodic, or almost periodic, signal may not sound expressive or lifelike, but it is usually heard as musical in an isolated context. This is the class of timbre Helmholtz referred to as musical. Such signals sound musical apparently because the temporal dimension of the SIMOR pattern is highly organized, resulting in most of the acoustic information concentrated in the "place" dimension, a much

smaller pattern space for the ear-brain to wrestle with and hence more explorable in the sense of Minski [Minski, 1981].

We have discussed this view in detail in chapter II. Exponential periodicity (i.e., that quality which a periodic signal multiplied by an exponential decay has) provides structure to the ear in a similar manner. When the period length changes or when the period structure changes, as long as they change in a perceptually organizable fashion, a balance of innovation and predictability provides further lifelikeness and expressiveness to the timbre's musicality. Here we use the term musicality in the narrow sense of a timbral quality, not in the sense of belongingness (as in, a "click" does not belong in the sound of a soprano's aria). In this (narrow) sense, we have asserted in chapter II that a timbre is musical if and only if the ear can organize the acoustic signal's SIMOR image. Thus band-passed white noise exhibits varying degrees of musicality or provides a varying measure of timbral quality. In chapter II, we further suggested that organizability to the ear is quite different from organizability to the mind. We will show this with the sound sequence generated by the Rudin-Shapiro polynomials [Schroeder, 1984].

As a result of the period trajectory's central role in organizing an acoustic signal's SIMOR pattern and its movement across the membrane, (1) the time scale change across several "periods," or (2) the time scale change from one period to the next, or (3) even time scale changes in the acoutic features on a finer level within a period, or (4) the changes in these changes from period to period (which we will call secondary features from now on) will perceptually coalesce and become distinctive features for the timbre. Thus, pitch glide, vibrato, and their time-varying characteristics derived from the period trajectory are perceived as distinctive timbral features.

Therefore, except in the case where the signal is rigidly periodic, and hence lacking in new information or innovation, the period trajectory is in general perceived as an acoustic element that provides important information about timbral features. This view has in fact, in one respect or another, been advanced by Schaeffer, Erickson, Schouten, and others, and is clearly evident in Charbonneau's perceptual data reduction study (at least for the class of timbre he studied). (On a more subtle level, pitch glide during the attack was found to contribute to the perception of the trumpet timbre [Risset and Matthews, 1969].) It also agrees with the experience of

many musicians as well as with the practice of many synthesis approaches, including some of the more sophisticated ones cited in 3.1.1.1. From an analysis point of view, the period trajectory is often identical to the fundamental frequency trajectory of a short-time Fourier transform analysis. However, the notion of frequency loses its *operative* meaning when the "period" length, or the "period" structure changes too fast, because of the (receiver's) time-frequency uncertainty principle. When this condition becomes true, different frequency transition trajectories between two stable frequencies can be equally acceptable because the receiver cannot tell them apart. But the period trajectory defines the evolution of local patterns of variation which we called frames in chapter II. Its definition is therefore unique and contributes to a unique timbre. In this sense, period trajectory is a timbral acoustic feature transcending the traditional notion of a pitch trajectory or a fundamental frequency trajectory. This distinction prevents the short-time Fourier transform or similar transform techniques from being generally useful analysis tools for timbral analysis and synthesis. In fact, it is a standard assumption among various short-time Fourier transform approaches that the signal be quasi-stationary, say over a fifteen to thirty millisecond time span. And it has been reported [Strawn, 1982] that the phase vocoder method does not work well for signals with drastic changes in their fundamental frequency.

Linear prediction generally uses a separate pitch detection method. Even then, we will see that a typical minimum-mean-squared-error (MMSE) approach does not necessarily provide accurate detection of the period trajectory. Thus this aspect of the timbral analysis is by no means trivial and we will show how more precise pattern matching than provided for by the maximum likelihood estimate algorithm is necessary for a perceptually meaningful reconstruction of timbre in the worst cases.

3.1.1.3 Breakframes and Transitions as Fundamental Features.

In chapter II, we defined a frame as a maximal segment of a vibration form that has no repeating subsegment. And we described a breakframe as a distinctive feature the ear stores in its permanent memory. Furthermore, we defined a breakframe as a frame from which other selected frames, not stored in permament memory, can be predicted or interpolated, with the aid of neighboring breakframes. In general, we observe that the continguous list of frames that make up the waveform together

with their kinematic nature trivially constitute the acoustic basis for the perception of timbre. We call this description a kinematic model of timbre.

In a stationary signal, any period is by definition a frame and, in this special case, can function as a breakframe because other periods can be trivially extrapolated from it (by an identiy operation together with a time shift of the period length P). Therefore, for such signals, a single frame forms the entire acoustic basis for the perception of timbre, i.e., any frame constitutes the entire list of breakframes. Under this condition, the discrete Fourier transform (DFT) of the entire signal has the same shape as the DFT of a single breakframe. The amplitude distribution displayed along the place dimension of the basilar membrane provides meaningful information about the sound in terms of spectral feature composition. Our frame description is in this case identical with Helmholtz's Fourier description.

In an exponentially decaying periodic signal with time constant α, any "period" can also be a breakframe and the other frames can be extrapolated from it by simple scaling. The scale factor for a frame N periods from the breakframe is $(e^{-\alpha T})^{\pm N}$ where the sign is chosen depending on whether the frame for extrapolation is in the past or the future of the breakframe (as long as it falls within the decay regime). A Fourier description continues to make sense given that the coherence provided by the decay envelope and the initial condition for the decay are maintained. That is, the phase information defined by the first period of the decay must be preserved.

For many natural timbres, the condition of quasi-stationarity is satisfied. Their waveforms usually consist of one or more stable neighborhoods of frames or locally "periodic" regimes. In the stable regimes, frames don't change very much and they can be "derived" from certain pivotal frames. These pivotal frames are therefore breakframes. In many instances, the intermediate frames can be "derived" from the breakframes by linear interpolation. One the one hand, these breakframes are acoustic distinctive features for the perception of timbre. On the other hand, they provide us with a way to simplify our description of timbre and its synthesis. For example, a graph of formant trajectories representing the loci of local maxima of time varying harmonic spectra of a speech sound can be thought of as a compact representation of the evolution of the breakframes of the speech timbre occurring at the time points where there is a breakpoint or at least one of the loci.

A syllable consisting of a consonant followed by a vowel is usually a composition

of a segment of irregular form gradually changing into the locally periodic form of the vowel. The transition is generally considered as an important timbral feature and is heard as perceptually interpolated. We can think of the acoustic transition as a succession of vibration forms, each one emerging as the transformation of the previous. The simplest case is linear interpolation. Other types of transformation may include some small random length and amplitude change in addition to some form of interpolation, not necessarily linear. We expect the transformation to be fixed, or at least slowly changing, for each frame, since usually the production mechanism cannot change very much in a short time. The frame that contains the irregular form is a breakframe because it cannot be derived from other frames but instead contains a pattern related to the transition. At the end of the transition, a frame representing the locally periodic vowel is also a breakframe, because it also contains a pattern not foreshadowed by the transition.

In this example, we suggest that a syllable with a leading consonant may be viewed as consisting of the initial segment of irregular form as breakframe number 1, a frame in the locally periodic regime of the vowel as breakframe number 2, and a frame in the trailing sound as breakframe number 3. The intermediate frames are derived by iterative transformation of these breakframes. Or the intermediate frames may be though of as a function of the breakframes and a parameter (say changing from zero to one). In the simplest case, the transformation is linear interpolation. In more complicated cases, information from the amplitude envelope and the period trajectory may be needed to guide the interpolation, as in the /a/ or the marimba. Note that we consider such a consonant-vowel syllable to be a single timbre, consistent with our dynamic theory of timbre, instead of adopting the phoneticist's phonemic division of the syllable.

In 3.1.1.2, we indicated that one purpose of detecting the period trajectory is to identify the local pattern of variation and its evolution. In other words, successful detection of the period trajectory amounts to marking out the frames. But in order to discover the nature of transformation from one breakframe to another, we sometimes need to figure out secondary and tertiary features. Here we have an example of trading algorithmic complexity for data volume. In using secondary or tertiary features to transform or interpolate frames, we increase algorithmic complexity in the analysis-synthesis procedure, but at the same time, reduce the data volume on

the intermediate frames. But more importantly, being able to translate the data into a transformation algorithm improves our knowledge of the internal dynamics and provides a more meaningful way for timbral interpolation and modification. It is always nicer to modify or interpolate the processes than the actual product of the processes. We haven't demonstrated how breakframes are actually determined except in broad terms, but we will present a model process to make it precise in 3.7.

In general, we can think of the breakframes as the (temporally) local acoustic variation which provides such an important perceptual imprint as to be retained in our permanent auditory memory and in such a way that we can recall or even vocalize the timbre by recalling these frames and their temporal relationships, i.e., the timings of the breakframes. The transitions are then resynthesized during the recall process.

3.1.2 Other Analysis/Synthesis Criteria.

In listing analysis criteria, the first thing is the analysis of the waveform into its dynamic triple. We have discussed an outline of the analysis above. A detailed description of the analysis we have used in our laboratory experiments will be given later.

3.1.2.1 Analysis.

3.1.2.1.1 Analysis of Importance Trees.

After demanding that it yield the triple, the next obvious criterion of an analysis method is that it decompose each component of the triple into a feature, or perceptual importance, tree. There are obviously obstacles at present to arrive at a complete metric description of the tree even if one wants to take the trouble to do it. First, the task involves no less than a psychoacoustic experiment involving a large number of perceptual distance measurements together with all the necessary preparation of stimuli. Second, it would, as we will show, involve the process of timbre interpolation, which is well known to be difficult. We will go into a more detailed discussion of these issues when we present experimental design(s) for these types of experiment.

On the other hand, we assume that a user of the analysis/synthesis operating environment whose purpose is oriented towards modifying sound and making music will be able to use his or her ear to organize the features into an *ordinal* tree. Such

a task involves no obvious technical problem other than careful listening (probably time windowed listening, *cf.* Helmholtz's frequency-windowed listening using the famous Helmholtz resonators) and work involving feature modification synthesis. The extent of it also depends on the user's application, specifically, the information rate of the music.

3.1.2.1.2 Non-destructiveness.

Next, in order to be sure that an arbitrary degree of analysis refinement can be achieved, the analysis technique should be non-destructive so that when we don't need all of the information, we know exactly what acoustic features to obliterate. We will show how this criterion is applied to the amplitude envelope detection algorithm and period trajectory algorithm in 3.5 and 3.6.

3.1.2.1.3 Locality.

Finally, the analysis should be local. The main reason is the dynamic nature of timbre perception. But it is also beneficial because any data processing is some function of the length of the data. The longer it is, the more inconvenient the analysis. In the case of linear predictive coding or the phase vocoder method, the data precessing requirement is $O(n^2)$ or $O(n \log n)$.

Next, we turn to synthesis criteria.

3.1.2.2 Synthesis.

From our point of view, synthesis can mean two things. If we are given the triple as a complete geometric object, then we can simply put the items together (see below). If we are given the feature importance trees, then we must first synthesize the corresponding components of the triple. From the triple, we synthesize the waveform on a frame by frame basis. Synthesis of a frame, from the digital point of view, means the synthesis of samples according to a certain sampling rate and a certain quantization width. A frame corresponding to a breakframe is simply retrieved from memory. A frame which is not a breakframe is synthesized based on the data provided by the relevant breakframes, the amplitude envelope and period trajectory for the regime, together with the algorithm of transformation. We will go into the details of the synthesis algorithm in 3.17; here, we shall simply bring forth some of the requirements necessary for a successful synthesis of timbre.

3.1.2.2.1 Elasticity.

First, because of the dynamic nature of timbre, it is critical that the synthesis algorithm possess the property of elasticity. By elasticity, we mean the ability to insert breakframes of broad characteristics so that timbral characteristics can be freely altered.

Consider the issue of dimensionality again, as discussed in 2.5.4. We may think of a short wood block sound say a few milliseconds long. Naturally, the sound is highly nonstationary and its waveform consists of only a few frames. Its dimensionality (more precisely, the dimensionality of timbre space at its location), i.e., the number of degrees of freedom or parameters necessary to define it, is necessarily smaller than that of a sound which is say of the order of one second long—provided the long sound is not restricted to be periodic, as the marimba (approximately, locally) is. We can imagine that x_1 in figure 2.5.4 represents the timbre of the woodblock and x_2 represents the marimba. In order to establish a transition between the two timbres, it is clear that the number of degrees of freedom will gradually increase from x_1 to x_2 or decrease from x_2 to x_1. (Of course, in the class of periodic signals with a given period ξ, the dimensionality is fixed, independent of the duration that a signal is played.) The way to realize this increase in degrees of freedom or dimensionality is to have a synthesis approach that is elastic enough to permit insertion of new breakframes and provide some kind of transitions to bring about the new timbral element. This property is important not only in the area of interpolating two timbres of different dimensionalities, but also in the area of providing new forms of timbre. For example, one can start with a marimba timbre, insert a clarinet local timbre frame at some point (prolonging the sound if necessary), then followed by another insertion of an /a/ frame, on and on, with each insertion followed by some kind of transition to ensure that the insertion can be *heard* and satisfy the kinematic property of timbre (see 2.3.3).

3.1.2.2.2 Geometricity.

It is clear from our description of timbre that acoustic distinctive features are highly geometric in nature, whether they are vibratos, tremolos, attacks, decays, local patterns of variation, or evolutions of variation, even when there are equivalent Fourier descriptions. Since humans are highly adapted to graphical representations of knowledge, it is important to take advantage of the geometric property of these

timbre acoustic features. For example, viewing the amplitude envelope as a concatenation of linear segments provides a much more powerful way of "understanding" the nature of attack and decay of a timbre than doing it with a list of pairs of numbers. A list of numbers does not easily illustrate the local cohesiveness of a subpattern or the "movement" by which subpatterns form patterns. Furthermore, it is important to be able to *see* that features that do not *belong* are actually absent. For example, waveform synthesis can often be marred by sudden large changes in sample amplitudes in the form of "clicks." But we can sort things out based on the local geometric character of the acoustic features if we have a graphic representation. Similarly, being able to see each frame we synthesize and compare it with what comes before and after is important to ensure a smooth or desired synthesis of the transition. Of course, the geometric nature of acoustic distinctive features also allows us to interpolate or modify these features more easily with desired results. With a strongly graphic-oriented environment, we can anticipate one or more being considered, viewed, and selected (with or without modification) as breakframes; then we could proceed to make a new waveform with the aid of an amplitude envelope and a period trajectory to drive the interpolation. This would allow the user to quickly hear the result from what "material" he or she chose.

3.1.2.2.3 Self-Consistency by Composition.

Since a feature is a composition of local (in the temporal sense) subfeatures which are in turn compositions of subfeatures, it is advantageous to be able to synthesize each component of the triple by composition of features. In Fourier synthesis, users are accustomed to this idea with respect to the partial amplitude envelope functions (see [Strawn, 1981]), but that is the extent of feature composition in timbral synthesis.

Actually, the most efficient synthesis of a sine wave can be viewed in the same way. First, we synthesize the samples for the first quarter cycle. Then, we retrograde the previously synthesized pattern and *form* the first half cycle. Finally, we invert (reflect) the previously synthesized pattern and *form* the full cycle of the wave.

For our synthesis approach, insertion of timbre regimes can be accomplished by this approach and fairly arbitrarily varied results can be generated with minimal effort. And this is one of the advantages of thinking of timbre as a composition of features. In general, if we have the trees, then we can modify them first (for

example, changing a particular subtree or even its importance) before the frame-by-frame waveform synthesis actually begins. And this can be efficiently done when we have a library of trees which we can call from long term computer storage.

Alternatively, if we have a library of acoustic features, e.g., breakframes, then we can select them, form a tree, and then tell the synthesis program to perform the waveform synthesis according to the selected trees from the triple.

3.1.2.2.4 Organicity by Autogeneration.

Organicity refers to the state or quality of being organic; in this case, organic in the sense that the analysis of a waveform proceeds in terms of patterns of vibration most natural to it, instead of in terms of, e.g., some fixed "basis" set such as the trigonometric functions.

Consider the diversity of timbre, or the range of acoustic waveforms in nature that give rise to timbre, and consider the general nature of a frame, as we have defined it. It would seem reasonable to think of the breakframes as the basic building blocks of waveform synthesis, which vary from sound to sound. For example, a "square" wave is a very simple geometric pattern requiring three parameters (amplitude, period, and duty cycle) to completely specify it. Therefore, it is unnecessary to describe it in terms of a Fourier series with a large number of parameters to specify the Fourier coefficients. Even if we are to interpolate the square wave with a sine wave, it is much easier to add sidebands to the spectrum of the sine wave, then back transform it to form a frame (or cycle), as a way to approach the square wave, than to start with a large Fourier representation and decimate it as a way to approach the sine wave. There is still another way which totally avoids the Fourier complexity. That is, we can interpolate the two geometric graphs in a polar coordinate system whose origin is midway within each half-cycle on the abscissa, based on the distance between the points of intersection for each fixed angle. As the wave moves away from the square shape, it is a more smooth curve, i.e., better approximated by a polygon of higher degree. Of course, we will still have to listen to it to see if the transition is perceptually smooth. But in general, the kinematic nature of the acoustic features of timbre dictates that the frames in a neighborhood of a breakframe must be similar. As a result, the data necessary to specify all of these frames could simply be the breakframe plus parameters describing the small distortions or the transformation that accounts for such small distortions. In fact,

our marimba synthesis using this strategy results in a 98% savings of data.

3.1.2.2.5 Rate-Distortion Criterion.

Rate refers to the amount of acoustic feature data necessary to synthesize a timbre. Distortion refers to the perceptual *distance* between the desired timbre and the actually synthesized one. The idea is borrowed from Shannon's rate-distortion theory in communication (see [Shannon, 1959] and [Berger, 1971]). In this theory, a message is coded at the source with enough information (or bits) so that the final recovered message at the receiver (distorted by lack of noise immunity) would not exceed a certain tolerance threshold. In our case, our concern is not noise interference, but in a similar spirit, that the sound should be synthesized with no more acoustic features than are required by the total information the ear is being fed by the stream of acoustic events. For example, in a busy musical passage, few listeners would be able to tell that the percussive synthesizer accompaniment is actually ahead of the live flute soloist on every beat due to what we call the Gordon effect—the phenomenon in which a melody composed of an arbitrary selection of timbres sounds out of synchrony with the intended rhythmic pattern—after Gordon's investigation of perceptual attack time [Gordon, 1984] (see 2.1.6.3). So there is no point investing precious resources to handle the Gordon effect in such a passage. But when we come to a slow soulful passage, we might want to invest our resources into correcting the Gordon effect now that the total information rate is low enough that such an effect is musically undesirable. Similarly, with the perceptual importance tree, we can decide which features to ignore with a measured loss of fidelity.

Finally, an ideal synthesis method must have the ability to make perfect duplication when desired and approach it with increasing fidelity with either higher data volume or higher algorithmic complexity or both. That is, it must satisfy a rate-distortion trade-off criterion.

3.2 Limitations of Current Sound Analysis Methods.

Each analysis method to be discussed has been found useful and even advantageous in extracting certain information about the acoustic signal under study. Each has certain limitations and each performs better than others in certain respects. Although every reasonable sound analysis technique requires that it be validated by the ear in the sense that the analysis parameters be capable of regenerating the original sound, not every analysis technique is receiver-based or ear-based. In particular, most existing sound analysis techniques do not take advantage of the fact that the ear organizes the multitude of acoustic information it receives. Those few which do, however, do not attempt to provide complete acoustic information for synthesis of arbitrarily high fidelity.

3.2.1 Linear Prediction.

Linear prediction is heavily biased towards source modelling. Although one might argue that the conjugate pole pairs provide a description of the attack, decay, and resonance characteristics of the system response, and hence their counterpart in the waveform that enters the ear, this argument is not valid in general. The logical leap from the system response to the waveform that enters the ear cannot always be justified because of the effect of transmissions (room acoustics) and because of various interactions that actually take place between the excitation and the response of the production system. Secondly, the pole pairs are not generally interpolable in the perceptual sense, making internal dynamics difficult to describe or synthesize. Thirdly, for the same reason, linear prediction parameters are not generally suitable for interpolation of timbre or description of timbral relationships.

These shortcomings appear to be characteristic of current physical (i.e., source) modelling approaches because the parameter set seems to vary significantly from source to source, making them useless for interpolation of timbres. Of course, there are all kinds of difficulties involving the issues of windowing which render the analysis data method-dependent at best and unreliable in general. Numerical instability inherent in solving recursive equations and inverting matrices is another problem. Although excitation can be in principle extracted from the *innovation* of the signal, i.e., the difference between the state and its estimate, in practice it is a separate analysis task. The choice of filter order, the choice of when to update the coefficients, and other problems have been discussed in detail by Moorer [Moorer,

1979] and Rabiner and Schafer [Rabiner and Schafer, 1978].

3.2.2 The Wigner Transform.

The Wigner transform is principally a general signal analysis technique. It possesses a number of properties that are considered superior to other transform-based signal processing algorithms (see [Claasen and Mecklenbräuker, 1980]). An important property pertains to its ability to avoid the problem of smearing in time or frequency due to the effect of analyzing windows found in time-varying filter bank analysis. As a result, the usual assumption of quasi-stationarity required by these data-smearing windowing transform analyses can at least in principle be relaxed. But it is not without drawbacks.

First, as an analysis tool aimed at discovering the instantaneous energy distribution of the signal, it is non-linear and requires massive computation power. Second, the analysis data cannot be readily interpreted: negative "energy densities" can occur, as well as cross terms for multi component signals, for example. Third, the massive amount of data in the time-frequency distribution mandates further processing to extract a meaningful "profile." But there is neither a known working processing algorithm, nor meaningful criteria to formulate one, which would do the further processing. Applications have not succeeded beyond simple frequency component signals, thus rendering it quite useless for timbre analysis except in very special cases such as those corresponding to simple chirp signals. The principal drawback is of course that the analysis does not take the ear into account; therefore, there is no reason for the analysis data to correspond to any timbral distinctive features. Similar lines of reasoning can be applied to Gabor's signal expansion in terms of logarithmic or Gaussian elementary signals (see [Bastiaans, 1980]). It is not aimed at timbral feature analysis.

3.2.3 The Short-Time Fourier Transform.

Fourier transform techniques at first seem to agree with Helmholtz's model of the ear as a frequency analyzer. But as we have discussed in chapter I, under Helmholtz, the agreement is at best on a formal level. The basilar membrane is a dynamic response system whereas Fourier analysis is naturally designed to solve stationary problems. The time-variant generalization of Fourier analysis introduces a number of signal processing problems. The first one is the effect of the analyzing window to smear data in both the time and frequency dimensions as discussed above. And

then there is the fundamental limitation imposed by the time-frequency product relation. Finally, partitioning the frequency domain into discrete bins introduces channel cross talk in the form of beating or microfluctuations that are artifacts of the analysis. While it is possible to recover the original signal if numerical stability is under control, since Fourier transforms are invertible transformations and the effect of analyzing windows can be mathematically accounted for, the analysis data is highly dependent on the analysis method itself, whose accuracy may or may not approach the theoretical limit attainable under physical law.

These signal processing limitations also prevent meaningful modification of the acoustic signal. But the most important drawback is that it is at best a restricted receiver-based analysis technique. First, the frequency response of the ear is logarithmically scaled. Second, it is constant-Q. Third, acting under physical law, it responds broadly to a wide range of frequencies. In particular, the response is wide-band for non-stationary signals, consistent with Heisenberg's principle of uncertainty in measurement. Fourth, the ear's form of signal detection employs a high degree of redundancy, a property advantageous for a receiver, but Fourier analysis is parsimonious in the sense that for any fixed time, the information among channels is orthogonal. Fifth, the ear responds on the basis of efficient energy transfer, which depends strongly on timing, whereas Fourier analysis is based on minimization of mean squared error. Although a minimum-mean-squared-error (MMSE) criterion can have an energy interpretation, it does not take timing into account. When the signal is dynamic in character, such as having a fast varying period trajectory, the ear's response is strongly dynamic, i.e., a fiber's response is strongly dependent on the timing between its own natural response to earlier excitation and the signal's current state in its changing dynamic. We will show this to be relevant in our analysis and synthesis of the /a/ sound which has a large change in its period trajectory.

Finally, the ear is an active observer which organizes its data, but Fourier analysis does not take advantage of this fact. In chapter II, we saw that timbre is not necessarily perceived in terms of Fourier components. Fourier techniques make the most sense when the sound is highly stationary or slowly varying in spectral content. In that case, the most interesting kind of information appears across the place dimension of the membrane. Yet many naturally occurring sounds do not

satisfy this requirement. Serra and others have shown that Fourier analysis alone does not make a very good analysis of the marimba or marimba-like tones. In this case, the amplitude envelope is a distinctive timbral feature and contains critical information for the synthesis of the marimba timbre. We will show this to be the case with the marimba analysis (done according to the method that we are proposing).

As an alternative to short-time Fourier transformation with constant partition width, Gambardella [Gambardella, 1971], Altes [Altes, 1978], Kajiya [Kajiya, 1979], Youngberg [Youngberg, 1979], Teaney *et al* [Teaney *et al* 1980], Schwede [Schwede, 1983], Petersen *et al* [Petersen, *et al* 1983], and Kashima [Kashima, 1984] have developed versions of constant-Q short-time spectral analysis. These efforts take into account the constant-Q and logarithmic scaling properties of the ear. Furthermore, they are linear and formulated in such a way that perfect recovery of the signal is possible, i.e., like short-time Fourier analysis, these constant-Q spectral analyses admit an inverse integral transform, or they exist in an analysis-synthesis transform pair. However, in addition to the fact the transformation property has an added level of complexity, it lacks the low-pass characteristic of Helmholtz's resonator bank model. But the most severe reservation we have with them in so far as timbre analysis and synthesis are concerned is the fact that they, like other analysis aproaches we have discussed above, do not take advantage of the ear's organizing property, therefore do not directly address the issues of timbral feature analysis and synthesis as we presented them in chapter II.

3.3 A Coordinated Analysis-Synthesis Strategy.

Grey correctly pointed out the importance of having a coordinated analysis-synthesis strategy by which data from analysis permits resynthesis perfect in the perceptual sense, i.e., indistinguishable from the original. The idea is that once one is able to do that, then one can throw away various parts of the analysis data and see which parts induce more damage from not being included. The idea of an analysis-synthesis transform pair has been an integral part of all serious analysis approaches in modern times. For example, short-time Fourier transformation, linear predictive coding, Wigner transformation, and some of the constant-Q versions of short-time Fourier transformation all exhibit this property on a mathematical basis (see 3.2).

However, a coordinated analysis-synthesis strategy requires more than the existence of a mathematically perfect analysis-synthesis transform pair. For if we want to be able to have controlled generation of timbre from the original, such as in timbre interpolation, we want to know that the acoustic control parameters being modified are timbral distinctive features; we also need to know exactly how the timbre will change as these acoustic features change to an arbitrary degree. Therefore, the analysis goal must be to discover the distinctive features and their relations, described in an importance tree, and the synthesis must be carried out on the basis of manipulating these distinctive features of timbre. The fact that the ear acts as a resonator bank does not mean that we can afford to look at the Fourier component magnitudes alone. Consider π, the ratio of the circumference to the diameter of a circle. Analysis of π into a series of digits such as 3.14159 does not mean that 3.04159 has a smaller error than 3.14150, as a naive expectation based on a distance function independent of decimal place might dictate. Of course, the reason is that each place has a weight or quantitative importance. In fact, we can arrange the infinite decimal expansion of π into a tree. In binary, it is obvious. For our purposes, the distinctive features are also ordered. Furthermore, in a representation that preserves the temporal measure, the hierarchy of importance, i.e., the place a certain kind of feature occupies, is dynamic, changing from timbre to timbre. Therefore, a coordinated analysis-synthesis strategy must not only be able to reproduce the original from the analysis data, it must involve distinctive feature analysis and synthesis. In this sense, mathematically generated analysis-synthesis

transform pairs do not automatically provide control over perceptual quality as we modify the parameters of the analysis.

Furthermore, such a coordinated analysis-synthesis of distinctive features permits a rate-distortion trade on the synthesis. And if we have adequate knowledge of the timbre, we can simultaneously simplify the analysis on the same basis, that is, we analyze just enough to serve the needs of the synthesis. So if the data rate of the distinctive features is high enough, the perceptual distortion will be sufficiently small as to be unnoticeable. But otherwise, a systematic inverse relationship exists that is advantageous to a user of the system.

In our approach, we analyze a waveform into its triple and synthesize it back to a waveform whose timbre is perceptually identical to that of the original. Note that our analysis-synthesis strategy is such that we do not *a priori* specify any basis functions from which a waveform is to be composed. We take the pattern from the waveform. Such an approach obviously does not provide an explicit transform pair. But the analysis data is cells or atoms from the waveform, the shaping functions are parts of the waveform itself (and also part of the analysis data), so enough knowledge about how they are transformed should provide the original. However, we do try to exploit a few facts from perception theory: the perceptual grid has a coarse nature, most timbres are of an interpolating nature from regime to regime, and the ear is an adaptive device. So if there are small deviations, the ear cannot tell. Our analysis-synthesis is aimed to take advantage of this fact. The algorithm does not form a mathematical transform pair in the usual sense, but, because of the perceptual stability of the distinctive features, modification of these features will more likely lead to a stable result than with other methods.

3.4 Choice of Analysis/Synthesis Test Material—Worst Cases.

Because of our stated goals, it is important to test our analysis-synthesis approach with worst case materials, i.e., with acoustic stimuli whose timbral features can severely test the validity of the method. The first requirement is that we have a severe test for both of the dynamic shaping functions. For the amplitude envelope, we are looking for a highly percussive timbre. For the period trajectory, we are looking for a natural timbre with a drastic pitch change.

The second requirement is that we be able to demonstrate that both dynamic functions are timbral distinctive features. We want to be able to hear a smooth but unmistakeable perceptual change in these features as they are gradually changed as by interpolation between two very different amplitude envelopes and two very different period trajectories.

It turns out that a hard-mallet-excited marimba tone at 260 Hertz has a very steep attack but a flat period trajectory. And it has been shown [Serra, 1986] that phase vocoder analysis data is unable to produce the sharp "noisy" quality of the attack timbre. So we think it is an appropriate *worst case* choice for the testing of the amplitude envelope as a distinctive timbral feature as well as the testing of the capability of our analysis/synthesis method, especially the amplitude envelope detection algorithm. Also, a female voiced **/a/**, with a pitch drop of slightly more than an octave, possesses a mild amplitude envelope, but a drastically changing period trajectory with vibrato (and amplitude modulation) towards the end, is, we think, an appropriate *worst case* choice for the period trajectory as a distinctive timbral feature, as well as a test of our period trajectory and frame marking algorithm.

The interpolation between the amplitude envelopes of these two timbres should provide an unmistakable test of the amplitude envelope as a distinctive feature of timbre. Similarly, the interpolation between the period trajectory of these two timbres should provide a similar test of the period trajectory as a timbral distinctive feature. Furthermore, the timbres interpolated along each feature according to the "vector addition" or parallelepiped model provide a further test of the analysis-synthesis method's ability to handle timbres of intermediate amplitude envelope and period trajectory characteristics. For example, we will see that applying the amplitude envelope detection algorithm to the waveforms synthesized by interpolated

amplitude envelopes using the same control parameters derived from the detection of the marimba envelope achieves similar success. Therefore, the amplitude envelope analysis is not simply successful on the marimba and the /a/ but on a series of timbres generated from them. In fact, the algorithmic complexity for the amplitude envelope seems to be dictated (or bounded) by the requirements of the worst case, namely the marimba envelope. This fact is encouraging in terms of what the algorithm can do for other timbres. The detailed analyses, and discussions of the observed properties of the timbres synthesized,will be presented in later sections.

3.5 Period Asynchronous Analysis for the Amplitude Envelope.

The main goal of amplitude envelope detection is to discover, from the waveform, a trend by which a local vibration form is perceived to transform on a frame to frame basis, in a collective manner, along the amplitude dimension, that is not part of the local fluctuations or their *relative* changes from frame to frame. For example, that part of the ebbs and flows of the harmonic partials of a sound that appears as local fluctuation or relative dynamics but not as a coherent overall amplitude change over some neighborhood in time that can be accounted for by transition from one breakframe to another should not become part of an amplitude envelope. More specifically, the first frame of the marimba tone we have analyzed is composed of four peaks representing the dominance of the second vibration mode of the marimba bar (whose eigenfrequency is approximately four times that of the fundamental mode of vibration) whereas the third frame is composed largely of ten peaks representing the dominance of the third or fourth vibration mode of the bar (whose eigenfrequencies are very close to being ten times that of the fundamental mode). The second frame has a local fluctuation pattern that can be approximated to within one or two percent on a sample by sample basis as half-way (.5 interpolation) between the first and the third. These local fluctuations and their relative dynamics do not form a coherent pattern as an overall change in amplitude over any neighborhood in time. Therefore, the marimba amplitude envelope over this time segment should not exhibit these local fluctuations.

It is clear from chapter II that insofar as timbre perception is concerned, the amplitude envelope of a sound represents a pattern (or trend) from the waveform that the ear can take advantage of in order to organize the multitude of fluctuations exhibited in the highly redundant single input multiple output response (SIMOR) pattern. Therefore, an amplitude envelope is primarily a geometric object for pattern recognition in the cochlea, not necessarily the output of a mathematical device designed for the logical part of the brain. We have discussed this aspect in terms of an auditory language analogous to the one in the visual perception of geometric forms. Perception of geometric forms in vision is a pattern recognition task involving feature detection and composition. We can devise algorithms to approximate the performance of the eye, or, in our case, the ear, as superb pattern recognizers.

But it might be premature to write down all the admissible features for an amplitude envelope. Pattern recognition is essentially a statistical estimation problem similar to the task of extracting a message from the bit-stream of a physical signal corrupted with noise in communication. The error criterion of our estimation is essentially one that must be tested perceptually. Specifically, an amplitude envelope is successfully extracted if it enables one to duplicate the original timbre through our synthesis strategy using the triple. However, we can list some of the more obvious constraints on a successful detection of the amplitude envelope.

(1) It is generally not enough to extract an amplitude envelope from a period synchronous analysis where one amplitude per "period" is calculated. This is because during a fast attack, a coherent amplitude growth pattern can appear as a composition of two or more coherent subpatterns. In other words, during a "period" of vibration, there is a non-uniform, or perhaps exponential, growth in amplitude, which requires more than one point or value to completely specify the pattern of the amplitude envelope in this neighborhood. This is in fact observed in the marimba attack envelope.

(2) It is not enough to look for all the local maxima in a waveform and interpolate them. This is because some of these local maxima are part of the local fluctuation in a frame that does not form a pattern for the overall amplitude change in any neighborhood in time. For example, the transition from a breakframe consisting of the first and second modes to a breakframe consisting of the first, second, and third modes would include frames in which the peaks of the waves representing the third mode would appear near or at the valleys of the waves representing the first or second modes. It would be a mistake to include these local maxima as part of the amplitude envelope because the transition frames would be perceptually maximally smooth when all fine structure that can be removed from the amplitude envelope point set is removed. However, the amplitude envelope point set should include a minimum (irreducible) set of features that must be part of the amplitude envelope or else a perfect duplication of the original would not be possible even though a smoother transition may be possible. On the other hand, if the amplitude envelope includes more than a minimal set of features, then the local pattern of variation will be

spuriously modified, so it should actually *be* an irreducible set of features.

(3) Moving average (MA) detection or auto-regressive moving average (ARMA) detection of the amplitude envelope may not serve our needs. First, an amplitude envelope function obtained from moving average type detection is usually smoother than the graph which strings the peaks of the waveform together by interpolation. Second, it lags in geometric shape behind this graph. In other words, the amplitude envelope reaches a certain value after the gross pattern of the waveform has done so. Third, the MA or ARMA approach involves a window of processing usually long enough to smear the positions of the peaks. If the processing window is not long enough, the output is rough. In any case, the approach is data-destructive in the sense that it produces a set of numbers which are not part of the sample sequence of the waveform. This new set of numbers may make the attack sound milder than it should. It may skip certain short-lived but nevertheless critical amplitude patterns necessary for a more perfect detection.

(4) A typical envelope detector using a rectifier followed by an RC (*R*esistor *C*apicitor) low-pass filter commonly used in signal demodulation using the amplitude modulation (AM) method is also inadequate because in the AM application, the carrier structure is regular, but a typical waveform for a natural timbre the ear receives does not have such a regular underlying structure. It is more often like the case of the marimba waveform we described in (2), above.

In general, the detection must be made adaptive to work for our purposes. In order to cope with the diversity of acoustic waveforms at hand and be able to maximally separate out the coherent amplitude change from the local amplitude fluctuation, we need:

(1) An adaptive strategy;

(2) A detection mechanism asynchronous with respect to the frame evolution or period length change;

(3) A non-destructive analysis, non-destructive in the sense that the output data set be a subset of the sample sequence (in fact, a subset of the local maxima), so that the amplitude envelope evolves synchronously with the

waveform; and

(4) (Because of (3) above), an operation on a kind of exclusion principle under which local maxima from the waveform are excluded if they do not form a coherent overall amplitude change, i.e., if they bear significance only to the local pattern of variation.

The essentials of the algorithm are:

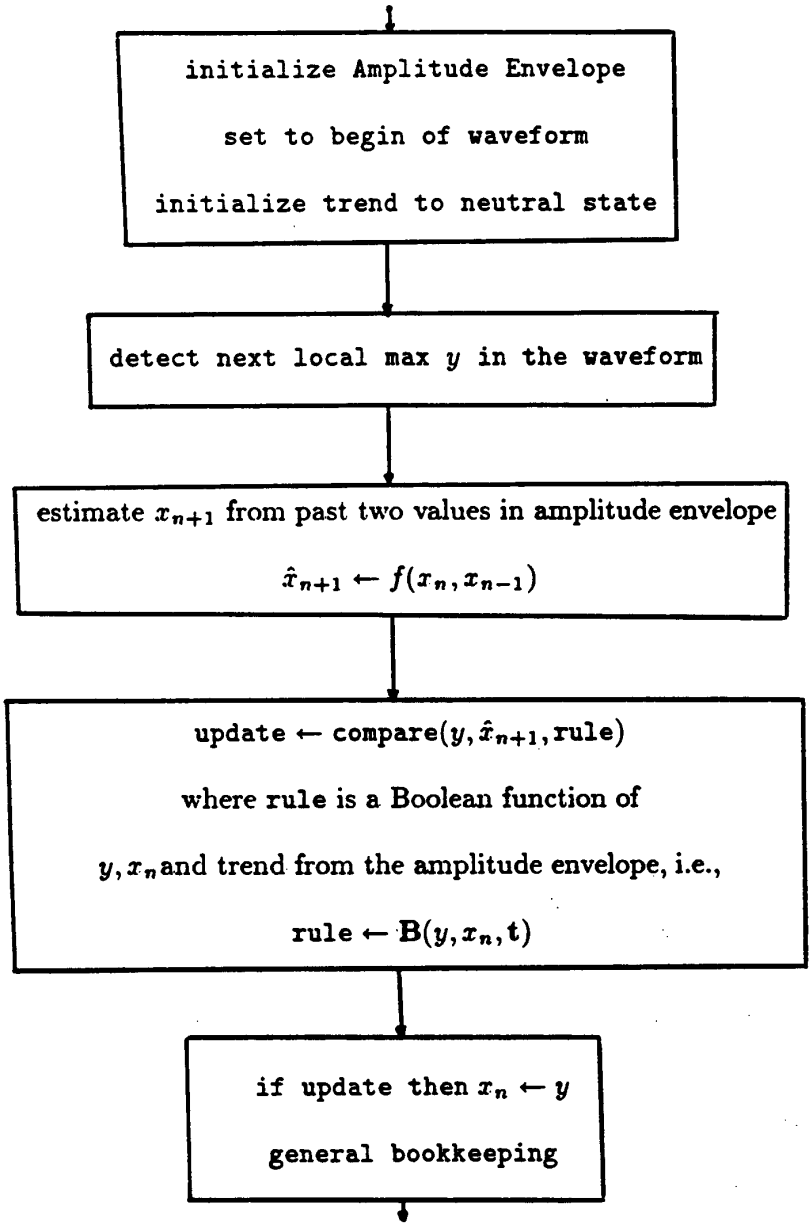


figure 3.5

Here f is a function of the ratio of its variables to imitate the ear's logarithmic scaling of the amplitude; **B** is empirically determined to make sure, for example, that the transition from attack to decay is not confused with local fluctuation. The former is observed to be milder partly because of the more rapidly varying nature of

the local fluctuation and partly because the gross amplitude evolution from attack to decay occurs over a time scale that is much greater than that of local fluctuation, e.g., ten times in the case of the marimba waveform. As a result, the normalized amplitude change is much smaller. Thus, by experimenting within the width of some band of tolerance, one can arrive at a certain optimal criterion. Therefore, the rule function that triggers an update on the amplitude envelope depends on y and x_n as well as the trend **t**. For example if $y < x_n$, and if the trend has been clearly descending, the the rule is to keep y from being a part of $\{x_k\}$ if it falls below the estimate $\hat{x}_{n+1}$ by amount ε_{DD}. ("DD" for "descending descending"—one descending for the trend and one descending for local behavior.) But if $y < x_n$ and the trend has been clearly ascending, then the rule is to keep y from being a part of $\{x_k\}$ if it falls below the estimate $\hat{x}_{n+1}$ by amount ε_{DA} ("descending ascending"), where ε_{DA} is designed to be more stringent than ε_{DD} since the relationship of y and x_n is against the trend of $\{x_k\}$. The update clause works for example like this:

```
if descending AND maxval < lastMaxval*exp(αε) then update ← false
```

where α is a tweaking parameter found to be between 1.05 and 1.1. This means $\log \texttt{maxval} - \log \texttt{lastMaxval} < \varepsilon$ and

$$\exp(\varepsilon) \equiv \left(\frac{\texttt{lastMaxval}}{\text{nextLastMaxval}}\right)^{\gamma}$$

or

$$\varepsilon = \gamma(\log \texttt{lastMaxval} - \log \texttt{nextLastMaxval})$$

and

$$\gamma \equiv \frac{\texttt{maxloc} - \texttt{lastMaxloc}}{\texttt{lastMaxloc} - \texttt{nextLastMaxloc}}.$$

Another feature of the algorithm is the ability to *backtrack*. For example, if the gross trend has been ascending but the last event is a downturn, and if the current event is an upturn again, then it is likely that the last event was an estimation error. So we erase the last event and in its place we put the current event. This is a feature that allows the trend to control the exclusion process while at the same time preventing the trend from being blind to the possibility of going over a peak on a global scale. We might imagine the "agents" in the society of ear, modeling after Minski's *Society of Mind* (see [Minsky, 1986]) performing in a

similar way: always favoring a majority-ruled pattern but alert enough to be good observers (receivers) to notice changes or new information. Of course, the algorithm must keep a counter for the Boolean trend variables and allow the user to supply threshold counts. Furthermore, for pragmatic reasons, it keeps the exclusion process from throwing away too many local maxima in a row. Our experience shows that a threshold count of eight or ten maxima works well both for the marimba waveform and the /a/ waveform. Plots for the amplitude envelopes detected for the marimba waveform and the /a/ waveform are given in figure 3.5.

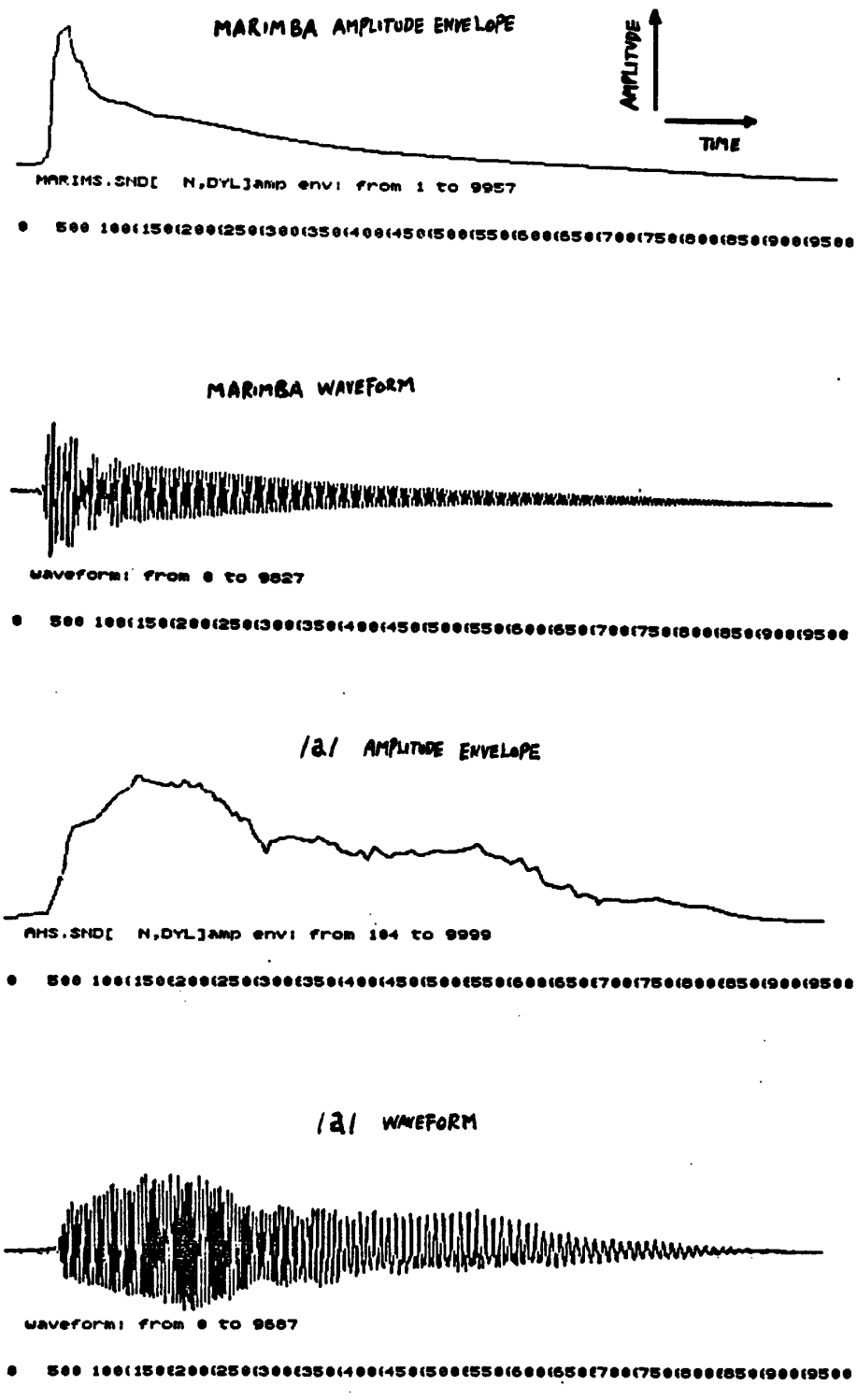

MARIMBA AMPLITUDE ENVELOPE
AMPLITUDE
TIME
MARIMS.SND[N,DYL]amp env: from 1 to 9957
0 500 100(150(200(250(300(350(400(450(500(550(600(650(700(750(800(850(900(9500
MARIMBA WAVEFORM
waveform: from 0 to 9827
0 500 100(150(200(250(300(350(400(450(500(550(600(650(700(750(800(850(900(9500
/a/ AMPLITUDE ENVELOPE
AHS.SND[N,DYL]amp env: from 104 to 9999
0 500 100(150(200(250(300(350(400(450(500(550(600(650(700(750(800(850(900(9500
/a/ WAVEFORM
waveform: from 0 to 9887
0 500 100(150(200(250(300(350(400(450(500(550(600(650(700(750(800(850(900(9500

figure 3.5

3.6 Adaptive Analysis of the Period Trajectory.

In order to mark the frames correctly, the single most important criterion is to produce a series of frames that are maximally similar from one frame to the next. The extremal condition (i.e., the condition of maximal similarity) is of course dictated by the signal itself. Its existence is implied by the observation that acoustic patterns in naturally occurring sounds do not change too abruptly (see 1.4.2.2.3 and 2.3.3). But the ear tends to organize the single input multiple output response (SIMOR) pattern locally. That means that the similarity criterion is most meaningful when it is applied to adjacent frames (or patterns).

There exist algorithms in the digital signal processing literature that appear to serve our purpose. They are generally known as pitch detection algorithms. Often their abilities rest on the assumption that the fundamental frequency (or period) does not change very rapidly, i.e., the signal must be stationary over a time of fifteen to thirty milliseconds for a fundamental frequency range centered between one-hundred and one-thousand Hertz. For our application, the worst case scenario involves rapid pitch drops or fluctuations in the frame lengths where "frame" of course refers to a segment of vibration pattern maximally similar to its adjacent neighbors. Therefore, the following criteria for the detection algorithm emerge:

(1) It must be a local analysis whereby the estimation of current frame boundaries should ideally be based on the knowledge of nearest neighbor frames or the past few neighbor frames; that is,

$$\hat{\mathbf{f}}_{n+1} = G(\mathbf{f}_{n-1}, \mathbf{f}_n),$$

where the circumflex in $\hat{\mathbf{f}}_{n+1}$ means it is an estimate for the frame $\mathbf{f}_{n+1}$. This means that the analysis method cannot be based on global transforms or long-time correlation strategies.

(2) The algorithm must be adaptive and adapt fast enough to follow the dynamic character of the period trajectory of some of the worst-case timbres. From a signal processing viewpoint, the algorithm must be able to develop strategies based on the current state of the signal. For example, when the detection is deep in a quasi-stationary regime, then the data length necessary to estimate the next frame boundary should be small so

as to minimize the computation intensity. Furthermore, under this condition, the estimation can be entirely based on knowledge of the past. However, during a rapid transition, or at the beginning of a waveform, the data length might be quite long and we must allow the algorithm to spend more time to do the job. Furthermore, under these conditions, i.e., when it is "groping," it might need future information to improve the quality of detection, hoping that things will get better. This strategy amounts to a delayed decision in a pattern recognition task with a certain fidelity criterion. Usually, things do get better, considering transitions from one quasi-stationary regime to another or from noise to a periodic regime.

(3) The algorithm should not be entirely dependent on a minimum-mean-squared-error (MMSE) criterion. Rather, the final decision should be based on maximal matching of neighboring patterns in the acoustic waveform. In other words, the criterion must be primarily a geometric one in the Euclidean sense. On the other hand, obtaining a MMSE is primarily an algebraic feat. In many cases, it makes no difference when determining a MMSE is done on a sample-by-sample basis, but our experience with the /a/ period trajectory analysis shows that being at the MMSE produces a perceptually inferior quality timbre when compared with one that takes into account the maximal pattern matching requirement.

(4) The frame marking should be peak-based because the ear is peak-based in its signal analysis character. The reasons are that noise interference makes zero crossing detection of frame boundaries unreliable, that the ear's rectifying characteristic also makes valley detection unreliable, and finally that peaks provide the strongest interaction between the membrane and the cilia through which neural transduction is initiated.

(5) The frame marking should be such that a frame detected should contain no repeating subframes (subsegments).

Criterion (1) rules out transform methods such as the cepstrum method and its close relatives (see [Noll, 1969]). The minimum variance method (also known as the optimal comb method [Moorer, 1974]) does not satisfy (1) very well. Our experience with the method suggests that it tends to get stuck when it is faced with a noisy segment, it picks the smallest segment possible, and it does not lend itself

to strategy switching very well. However, its tendency to pick the smallest segment is useful for criterion (5). We will see how it might be used in a hybrid fashion.

In search of an ideal pitch detection algorithm in the literature, we looked at the maximum likelihood estimation (MLE) algorithm. Regarding this method, Noll [Noll, 1969, p.17] observed, "But clearly there should be an 'optimum' method for fundamental-frequency determination, and standard signal processing techniques of analysis should give the method. This motivation resulted in a maximum likelihood estimate of the period of a periodic signal. The mathematical derivation of the maximum likelihood estimate of the periodic signal was performed by Dr. David Slepian; ... ."

However, this method also has a number of drawbacks: (1) It tends to over-estimate the period length so that a frame marked by it tends to contain several repeating subframes. (2) It uses the MMSE idea. (3) It does not care about peak picking. But its tendency to pick a longer frame than is correct is useful from a computational point of view. In other words, even if it does not do the job right, it will do it fast. For this reason, and for the reason that philosophically, it is close to the pattern recognition task at hand, we use it as the starting point.

We will not go into the details of the algorithm or its justification, for the interested reader can find this information from Noll's paper. We will simply give a description of the idea and proceed to describe how we modify it to suit our purposes. The basic idea of MLE has a geometric content. If we take a fixed length of samples in the sampled waveform, divide it up into contiguous segments, and then add the corresponding samples in each segment together, square them, and then add up all of these squared sums as a function of the sample location in each segment, then ideally the one partition which yields maximally similar segments should yield the largest sum in the summation above. This partition defines (identifies, delineates) a frame, and the process will continue, starting with the end of this frame.

We modify this algorithm in a number of ways.

First, the range of estimation R is made initially large. For example, if the estimated length is x, then we provide a range R between 0 and $2x$. But once the estimation stabilizes, R can be narrowed according to the estimate $\hat{x}$ and the initial value $\hat{x}'$. In general, R is a function of $\hat{x} - \hat{x}'$.

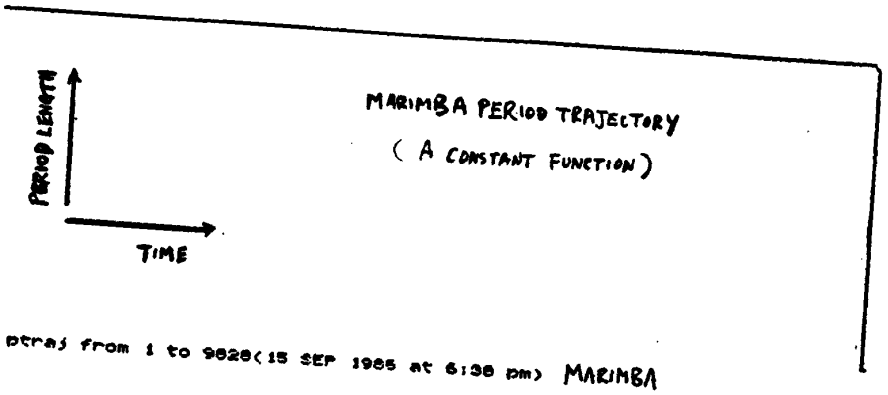
Second, when the estimation process is unstable, depending on whether it starts

out unstable or becomes unstable, the algorithm would choose either future data or past data for the estimation. For example, during a transition from noise to a vowel, future data may be more relevant, but during a transition between two stable regions, the decision depends on which part of the transition the current estimator is working on.

Third, a minimum variance option is available so that when it is activated, the estimation process would take the MLE and send it through the minimum variance length detector. This hybrid combination seems to have worked quite well with the /a/ analysis. (The marimba analysis is easy because the fundamental frequency is constant.) Note that the detection starts at the beginning of the sound and is complete after a one pass process that does not involve normalization.

Fourth, the frame boundaries are forced to coincide with the nearest peaks around. This process has been found to greatly improve the perceptual quality of the /a/ sound.

A reprint of the code, followed by the list of frame boundary locations in sample indices, is in appendix A. Plots of the period trajectories detected for the marimba waveform and the /a/ waveform are given in figure 3.6.



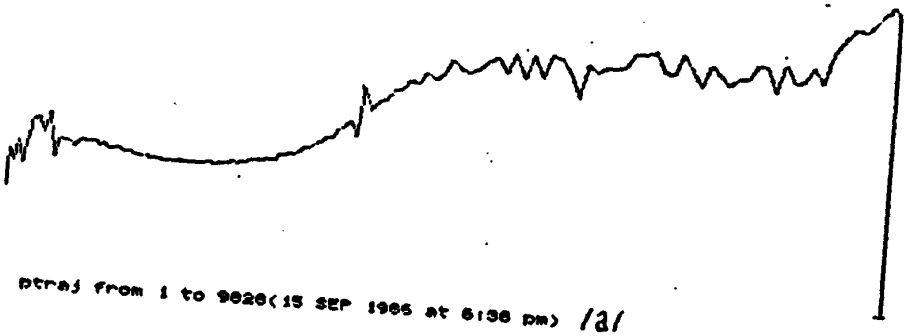
ptraj from 1 to 9828(15 SEP 1986 at 6:38 pm) MARIMBA

MARIMBA WAVEFORM

waveform: from 0 to 9827

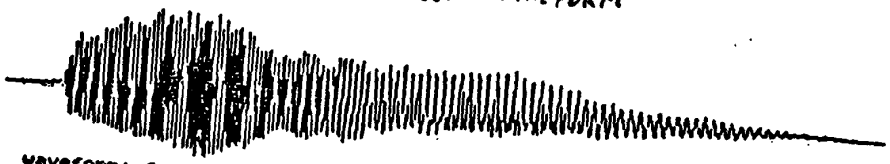
0 500 1000(1500(2000(2500(3000(3500(4000(4500(5000(5500(6000(6500(7000(7500(8000(8500(9000(9500

/a/ PERIOD TRAJECTORY



ptraj from 1 to 9828(15 SEP 1986 at 6:38 pm) /a/

/a/ WAVEFORM



waveform: from 0 to 9827

0 500 1000(1500(2000(2500(3000(3500(4000(4500(5000(5500(6000(6500(7000(7500(8000(8500(9000(9500

figure 3.6

3.7 Analysis of Frames and their Evolution.

The purpose of frame analysis is to determine the sort of breakframes necessary to describe and hence to resynthesize the timbre. The list of breakframes can be thought of as a distinctive acoustic feature of the timbre that is stored in memory and can be recalled or even used to help one vocalize one's perception of that timbre, in the sense of simulating the timbre by one's own voice. We can model the ear's frame analysis in terms of a processor, a frame buffer, and a frame memory. The frame memory stores the breakframes in the sequence they occur in a timbre. The frame buffer stores the frames that are currently mirrored in the auditory windows of the membrane fibers in the cochlea. The analysis process can be summarized into the flow diagram on the next page.

In Pascal-like code, the algorithm is:

```
last ← 0;
curr ← 1;
while not EOS(current) do
  begin
    next ← curr + 1;
    get(F(next));
    F̂(next) ← Estimate(F(last),F(curr),AE,PT);
    if not compare(F̂(next),F(next),ε) then
      begin
        M[loc] ← F(curr);
        loc ← loc + 1;
      end;
    Comment: M is organized in such a way that each
    loc points to a record structure that stores the
    entire frame of vibration pattern
      else begin
        last ← curr;
        curr ← next;
      end;
  end;
```

Note that $\mathcal{F}$ is a buffer with a structure similar to $\mathcal{M}$. A block diagram representation has been presented elsewhere (see figure 3 of [Lo, 1986]).

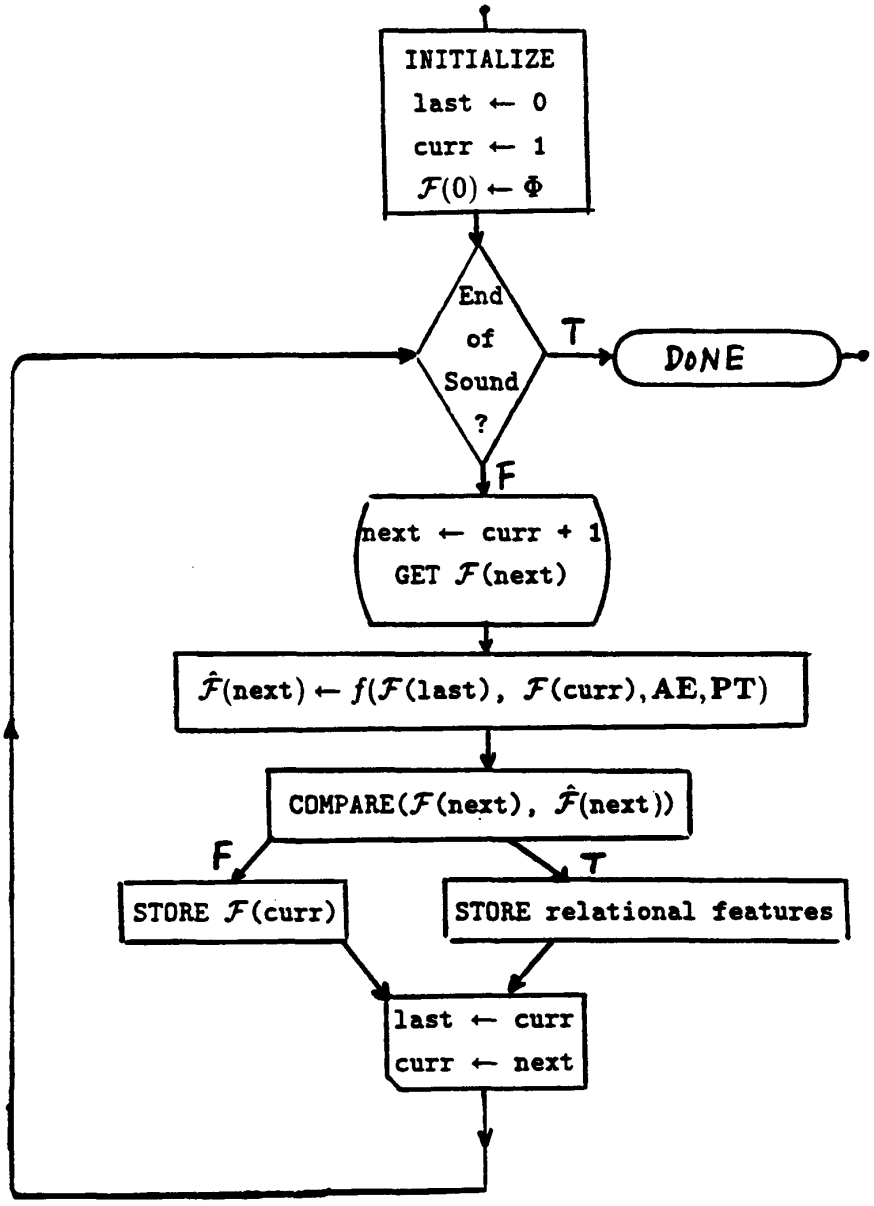


figure 3.7

Here, Φ is the silence state. The last, current, and next frames are denoted by `last`, `curr`, and `next`. The estimate of the next frame is denoted by $\hat{\mathcal{F}}$(`next`). The frames are presumed already marked by the detection of the period trajectory. In actuality, the detection of the amplitude envelope, the period trajectory, and the list of breakframes goes on simultaneously. In computer language, these detectors have a shared data memory. In the future, perhaps we will work out a model of such parallel processing. But for now, we are satisfied to break up the analysis in such a way that the analysis of frames into breakframes and transitions is subsequent to the amplitude envelope detection and period trajectory detection, and they make use of the data obtained from these.

The analysis of frames also involves discovering the nature of the transition between any two breakframes. `STORE relational features` describes the amplitude-scale transformation and time-scale transformation that is necessary to produce a successful comparison between $\mathcal{F}$(`next`) and $\hat{\mathcal{F}}$(`next`). Successful transformation from $\mathcal{F}$(`curr`) and $\mathcal{F}$(`last`) to $\hat{\mathcal{F}}$(`next`) may involve non-uniform scaling within each frame in time, amplitude, or both. For example, the KDSI synthesis of the marimba demonstrates that successful synthesis of the attack using only the first and peak frames must be accompanied by non-uniform amplitude-scaling dictated by the amplitude envelope. For the time dimension, non-uniform scaling means extracting secondary or tertiary features. Secondary features are of course acoustic events of secondary perceptual importance, such as peaks other than the frame-marking peaks.

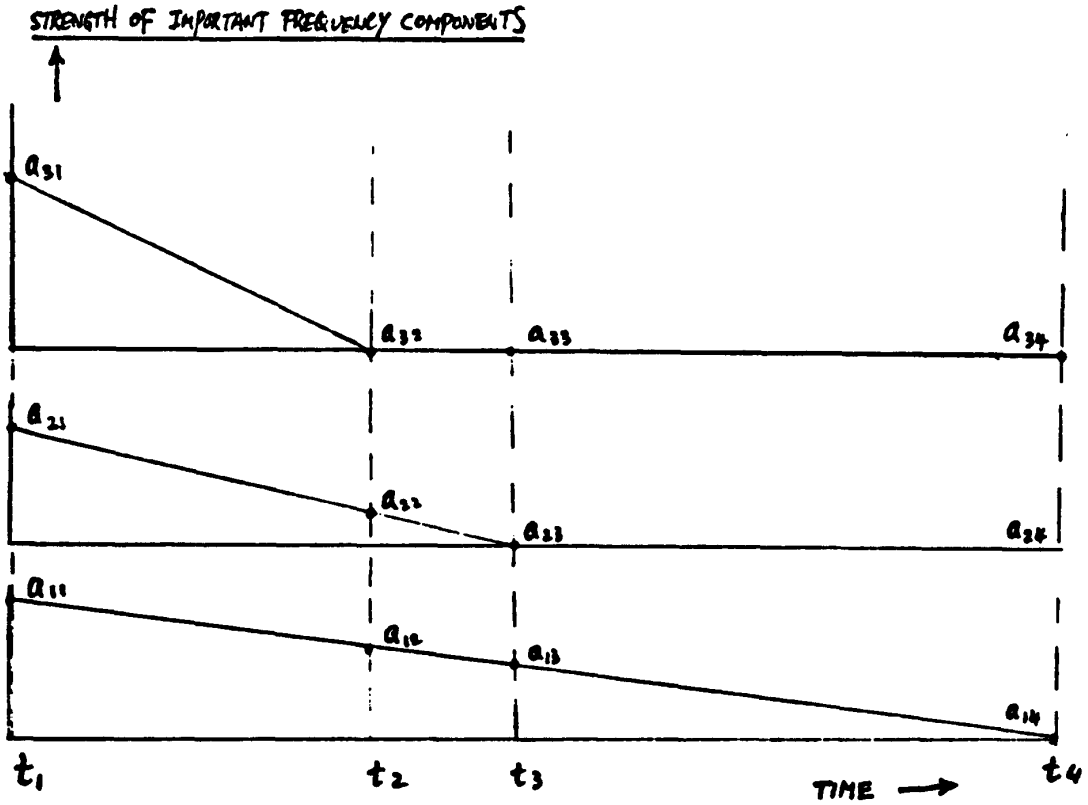
These relational features will be used later for synthesis of the transitions. Notice that the `COMPARE` procedure includes a parameter ε which is a tolerance band for the frame comparison. This tolerance parameter is believed to be a function of the local fluctuation similar to the ε's in the amplitude envelope detection (see 3.5). Although this model has not been implemented on a computer, we believe that by experimenting with different ε's applied systematically to a ℓ_1, ℓ_2, or ℓ_∞ error criterion on the frame data, one may be able to automate the frame analysis process. The selected list of breakframes can then be checked directly against the waveform and further perceptually checked by comparing the original with the resynthesis using the triple whose list of breakframes is chosen this way.

3.8 Synthesis by the Triples and Linear Interpolation.

If the amplitude envelope can be approximated by a line segment between two successive breakframes and if the period trajectory is flat between them, then synthesis of the timbre can be realized by rescaling the breakframes and linearly interpolating the transition beteen the breakframes, without even the need for the amplitude envelope and the period trajectory. This assertion is consistent with the concept of amplitude envelope and period trajectory as timbral features. In this case, these two elements do not have added information to contribute as additional features for the timbre. Consider the marimba decay as composed of three partials with straight line declining amplitude envelopes (see figure 3.8(a)). Suppose the decay begins at t_1, and partial number 3 goes below the threshold of hearing at t_2, partial number 2 does so at t_3, and partial number 1 at t_4. We then have three linearly interpolable regimes defined by the four partition points t_1, t_2, t_3, and t_4. The distribution at t_1 is (a_{11}, a_{21}, a_{31}), at t_2 it is (a_{12}, a_{22}, a_{32}), at t_3, (a_{13}, a_{23}, a_{33}), and at t_4, (a_{14}, a_{24}, a_{34}). In other words, the evolution of the timbre for the decay can be represented by the matrix

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{pmatrix},$$

where each column represents the partial magnitude distribution at a certain instant. Therefore, from the frame viewpoint, we can think of these columns as breakframes and the simultaneous transitions of the partials as transitions between the breakframes. It is true that the magnitudes, unlike the real and imaginary parts of the Fourier transform, are not linearly related to the waveform. Therefore, physically, they are different objects and perceptually they may or may not be the same, depending on the degree of stationarity and the auditory attention. In general, however, our waveform approach tends to preserve the phase relationships among the partials provided that the period trajectory is correct and the transition exhibits a certain interpolated property in the local pattern of fluctuation from frame to frame. This approach of course provides elasticity for inserting a frame of arbitrary spectral composition, which cannot be conveniently done in the Fourier representation because of the phase issue, the number of partials to contend with, and the difficulty of turning oscillators on and off in the middle of an acoustic event



$a_{ij} \equiv$ strength of the ith component at time t_j. (SEE 3.8)

PARTIAL MAGNITUDE DISTRIBUTION AS A FUNCTION OF TIME

FIGURE 3.8a

in a precisely controlled manner, i.e., to the sample time level. The phase issue refers to the general unavailability of the phase information on the partials and the numerical instability in their time derivatives for a typical phase vocoder type analyses.

We speculate that the idea of breakframes and their interpolation as a means of resynthesis when the amplitude envelope and period trajectory are treated as secondary timbral features may be applied to obtain a coarse synthesis of speech sounds, given their formant trajectories and their fundamental frequencies. For example, if there are three formants with each formant trajectory approximated by a number of line segments, then if n is the number of time points which mark at least one breakpoint in the line segment collection, then n breakframes whose waveforms are inverse discrete Fourier transforms (DFTs) to the magnitude distribution matrix (similar to the one given above) may be capable of reproducing the speech timbre, provided that all the breakframes agree in their phase relationships (i.e., if the frames always start with a peak, for example).

The synthesis algorithm essentially consists of looping through the frame synthesis N times, where N is the number of frames chosen to make the transition. The frame synthesis consists of producing a sequence of samples of length L determined by the period trajectory at a certain point in time. If L is the same as the length of the end frames, then the sequence of samples is simply a weighted average of the corresponding samples in the end frames, where the weight is determined by the distance to the frame being synthesized, normalized by the distance between the end frames. If L is not the same as the lengths of the end frames, and if it is not too different, then simple length interpolation at constant sampling frequency must be performed first. That is, the end frames must be interpolated to the desired length of the frame being synthesized before they are mixed according to the distance-controlled weight. Furthermore, length changes may become non-integral and the accumulation of fractions can produce perceptually adverse effects (such as clicks), unless the length change interpolation is done on a real length basis with the off-set properly acknowledged by the next frame to be synthesized. For example, if the frame is cosine-phased and if the offset amounts to $\pi/6$, then the next frame must not start with the same cosine-phase, but with a lag of $\pi/6$.

When the amplitude envelope or period trajectory is not constant, or cannot

be approximated by a line segment between consecutive breakframes, then the variations in the amplitude envelope or period trajectory become acoustic features for the timbre, and the synthesis must be controlled by these components of the triple as well. We have already explained how length changes may be handled. Amplitude changes are done by first normalizing the sample amplitudes of the amplitude envelope. After real length interpolation on these frames, they are mixed according to the distance controlled interpolation factor and then scaled to the amplitude-segment function local to the frame. A schematic of the synthesis algorithm is shown in figure 3.8(b).

FRAME-GENERATION

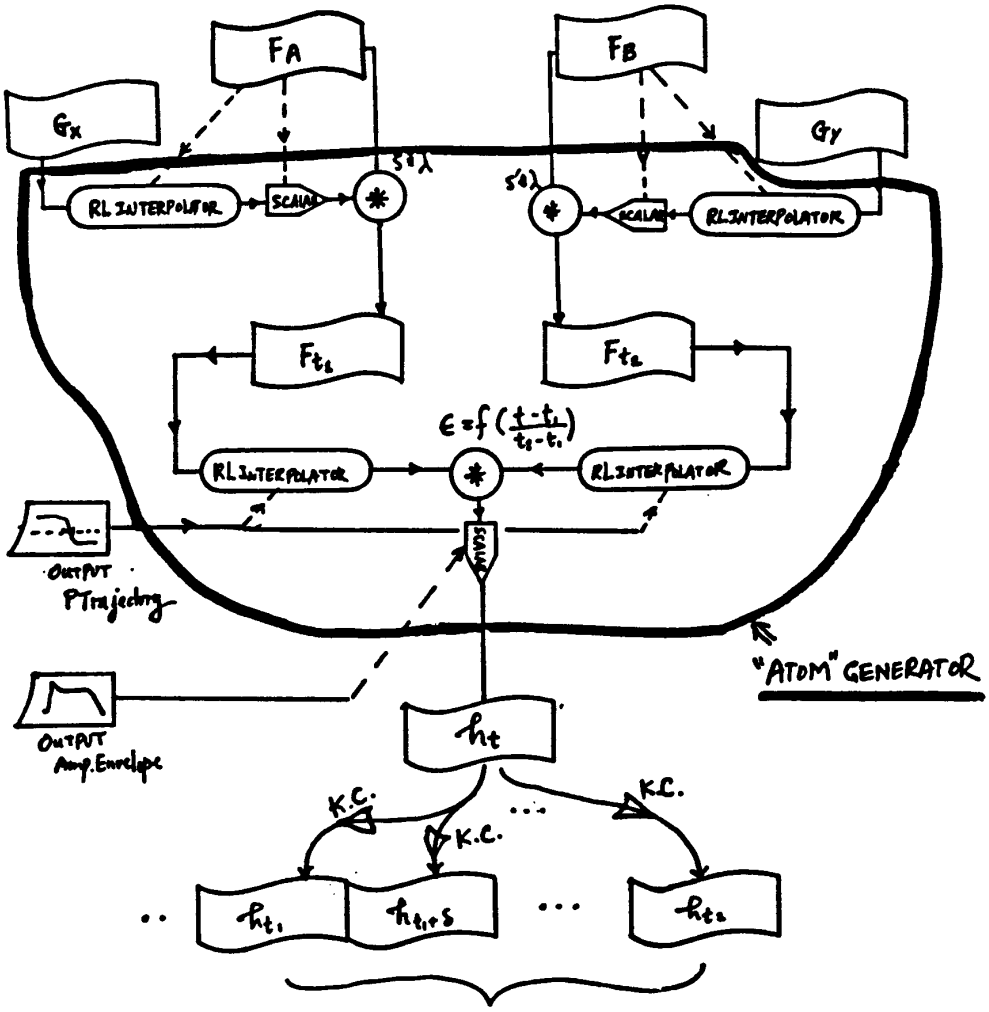


FIGURE 3.8(b) OUTPUT WAVEFORM SEGMENT $[t_1 : t_2]$

3.9 More General Synthesis Approaches.

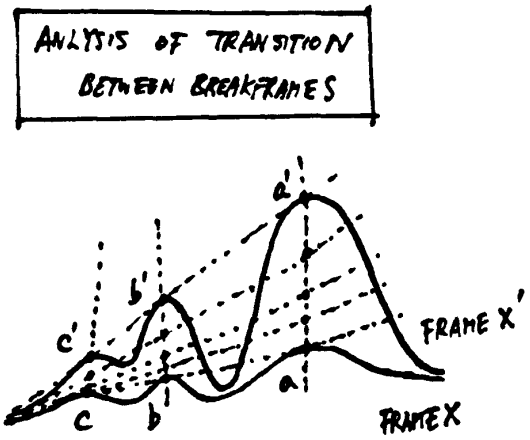
A timbre, according to our theory, has been alternately described as a composition of features or as a triple $(\mathcal{A}, \mathcal{P}, \{F_k\})$. (Of course, each component of the triple is also a composition of features in its own class.) It is therefore possible to synthesize a timbre using its triple. In other words, a timbre can be synthesized according to its data structure and a fixed algorithm applicable to all timbres. Specifically, at least in principle, the amplitude envelope possesses all the amplitude-scaling information and the period trajectory contains all the time scaling information necessary to resynthesize all the intermediate frames between the breakframes selected in the manner described in 3.7. In practice, however, prominent features such as vibrato or percussive attack occur in so many timbres that it might be desirable to have an algorithm to *describe* the synthesis of these features based on a much reduced data set. For example, while it is possible to recover the attack timbre of a marimba tone from the six breakframes defining the lowest level segments of the subtree called `attk` in the model tree of figure 2.5.2(b), it might be better if there existed an algorithm to synthesize the attack using fewer breakframes. We will show that this is indeed possible with a non-linear method using $\mathcal{A}$ as a key information element of the interpolation (see 3.10 on kinematic synthesis by dynamic interpolation). Similarly, instead of using two breakframes for each cycle of a vibrato (this can be compared with sampling a sinusoid at least twice per cycle), and hence needing $2f_V T$ breakframes for a vibrato of frequency f_V lasting a duration T, we might want to find an algorithm which takes only two frames and two parameters, f_V and T.

This approach might apply to continuous speech sounds as well. It has been observed that in speech, a systematic shortening of peak-to-peak distances in a sine-like vibration pattern, similar to a segment of a vibrato, within the duration of a fundamental period, is common [Shannon, R.]. Furthermore, this type of transient is not in general within the capability of existing transform-type methods to analyze [Shannon, R.]. On the other hand, transitions involving these "rushes" appear to be describable by a systematic transformation from one frame to the next. In other words, there is enough regularity in the transition pattern to suggest that a frame to frame transformation can be found to regenerate the series of frames in this type of transition. Still another way to describe the situation is to say that the frames

visually appear to be interpolable from the breakframes that delimit the transition although the nature of the interpolation is certainly not merely linear.

Transitions in naturally occurring timbres almost always appear to be interpolated because, as Winckel pointed out (see 1.4.2.2.3), source response to changes must go through a transition because of the mechanical inertia inherent in the source. For example, the brain might tell the voice production system to make a sound that starts with a high-pitched, high intensity /a/ and ends with a slightly nasalized, low-pitched, low intensity /a/. High-pitched, high intensity voiced sounds usually involve high air volume velocity and low vocal cord duty factor; low-pitched, low intensity sounds involve just the opposite things [Flanagan, 1972].

These velocity transients, coupled with the opening and closing of the velum (the membrane that controls air passage between the nasal tract and the vocal tract), create momentary spatial inhomogeniety in the air density. The law of flow conservation translates this type of spatial inhomogeniety into temporal inhomogeniety in the pressure variation at a particular point in space (see [Sommerfeld, 1947]). It is quite possible that the type of "rushes" Shannon observed, and the fairly irregular nasalized vibrato appearing in the /a/ with a drop of slightly more than an octave that we have worked with, may be described by the above physical mechanism since the transient occurs within the length of a fundamental period. It is therefore useful to study the geometric aspect of these organizable and yet little understood patterns of transition discussed above. It is particularly useful because, as we pointed out above, transform techniques cannot in general cope with this type of transient. And it is quite possible to discover a transformation that can be iteratively applied to a frame to generate a whole series of frames that constitute a certain transition. This kind of knowledge can be obtained by computer simulation of the fluid flow in a voice production system that includes the nasal and oral cavities and is put under these excitation and transition conditions.



(SEE TEXT TO THE RIGHT)

Alternatively, we can approach the problem as one in signal processing. For example, we can superpose the frames and trace out the loci of the acoustic features. Suppose we can do this on an interactive graphic basis—say with a mouse or a light pen. Suppose software exists to save these loci. Then during resynthesis, all we need is to recall the breakframe and these loci. And the transition can be regenerated on a frame-by-frame basis using these loci and the breakframe.

In general, we look at the patterns of any two consecutive breakframes as boundary conditions between which the transition frame patterns "move" from one boundary to the other. The transformation may be quasi-conformal, i.e., almost angle preserving, between the families of equipotential curves of the two independent variables describing these patterns.

This idea has been applied in biology by D'Arcy Thompson, a British biologist (and mathematician). In his book on growth and form [Thompson, 1961], he first studied in detail the biological forms in nature from cells through tissues to skeletons and then derived transformations that successfully correlated the shapes of different species of fish and birds, etc. Apparently, the famed artist Albrecht Dürer (1471-1528) had succeeded in applying the principle of coordinate transformation to the study of proportion and shape of human figures. According to Thompson, Dürer had fully described and put in practice this method in his *Geometry* and especially in his *Treatise on Proportion* (see [Thompson, 1961, p. 290]). Thompson's own work on fish can be found in [Thompson, 1961, p.301]. Although these are spatial figures while our patterns are temporal figures, there is reason to believe that there are fundamental similarities. For the law of flow conservation does relate pressure as a function of time to its spatial variation and the volume velocity. And our ear

was once on the fish's belly, interacting with the fluid flow all the fish's life.

If we speculate that Thompson's species transformation could be explained by change of fluid condition in the living environment, then those changes in fluid condition on a much more local scale of time may in fact be part of the transition of vibrato forms observed in speech and other sounds. [Nature, according to A.C.Clarke (inventor of the communications satellite), is extremely economical in her use of forms, using the same idea for the shape of the galaxy and the shape of the whirl of the water down the drain (see [Clarke, 1982]).]

The techniques of Thompson and Dürer are essentially empirical. And it should be the same for our purposes. The coordinate transformation properties should be inherent in the data structure. But if we have enough samples (in the sense of examples) for a particular class of features, a pattern may emerge from the collection of coordinate transformations that would allow us to describe a feature common to all the samples as an algorithm. When this is the case, one may associate an algorithm with the composition of a distinctive feature, a feature salient to a particular class of timbre. Such a feature may be a fortissimo attack in a certain class of percussive sounds, a vibrato of a female voice, or a particular type of transition in a continuous speech sound.

An advantage of replacing a part of a data structure with an algorithm is simplicity. The advantage of simplicity can be gained in either duplication of a timbre or creation of a new one with the particular feature the algorithm is associated with. An important example in the literature are the trumpet studies described in 1.5.1.2 and 1.5.1.3 which enabled Beauchamp to correlate the brass-like quality with a non-linear synthesis algorithm in which a voltage-controlled linear predictive filter's cutoff frequency is a monotone increasing function of the sound amplitude [Risset and Wessel, 1982]. The most famous example, however, is the Yamaha DX series' approach to digital synthesis of timbre based on Chowning's frequency modulation technique. The synthesis is algorithmically driven. Each algorithm is a logical combination of functional units, or oscillators, designed to facilitate the creation of a particular class of timbre. What we have proposed to do in this section is to achieve a similar synthesis objective from a systematic analysis approach. The purpose of analysis is of course to cope with the wide class of timbres or timbral features which are not already found in the frequency-modulation based timbre universe.

Another advantage of replacing a part of a data structure with an algorithm may be the possibility of discovering how the auditory processsor works in regard to how it perceives, remembers, and recalls the features. We have seen how perceptual considerations figured in our analysis/synthesis procedure (see 3.5, 3.6, and 3.7). A properly formulated algorithm may be a model of an agency in the society of ear. The discovery of each algorithm may well represent another perceptual agency the society of ear has developed in response to a particular class of signal stimulation. If Schubert is right about the versatility and adaptivity of the ear as a successful receiver (see 2.2.1), then an all-purpose, simple-minded and passive observer type effort, such as those of the Fourier transform class of methods, may not be what the ear prefers to have as its only tool. The Fourier transform may be just one of the many analysis agencies in the society of ear. The Fourier tool handles the job of stationary or quasi-stationary stimulation excellently, and this is where its place in the society of ear should be. If we take this view, perhaps a successful timbre operating environment will eventually emerge as we learn to treat different signal situations, i.e., to synthesize different classes of timbral features, with some of the finesse and dexterity of the ear.

3.10 KSDI and the Trade-Off between Data and Algorithmic Complexity.

An example, a simple one that works, to demonstrate the use of an algorithm to replace a part of the timbre data structure for a successful timbre synthesis, is the kinematic synthesis of dynamic interpolation (KSDI) of the hard mallet marimba tone, demonstrated at the ICMC '86 by the author. The brilliant attack timbre of the tone is the most distinctive feature of the tone's entire timbre. At the same time, Serra, who successfully duplicated the tone perceptually by a hybrid synthesis technique using additive noise to simulate the attack timbre, reported that the data from the Fourier based phase vocoder analysis was not sufficient for the duplication, without the additive noise. Therefore, it appears to be a challenge to see if our analysis/synthesis can obtain the salient acoustic features for perfect duplication.

First, we observe that only two of the three components in the triple provide the attack features because the period trajectory is flat, and therefore uninformative beyond organizing the waveform into frames of fixed length. Second, we observe that the attack is so sharp that interpolation based on temporal distance between breakframes in the attack could not capture the exponential or non-linear growth characteristic of the timbre. Third, we analyze the change in the pattern of fluctuation on a frame-by-frame basis for the first seven or eight frames, and confirm that while the first three are well related with one another linearly, the following frames cannot be obtained from one another, even with the normalization of a single amplitude scalar for each frame. On the other hand, if some of these frames are scaled internally by different amplitude scalars based on the local values of the amplitude envelope, then the interpolation again is numerically meaningful. Thus, when we try to interpolate corresponding samples between two frames, the first one and the one that contains the peak, the result is good to within the same accuracy as the first three. Thus we decide to try synthesizing the marimba with the algorithm

$$S(k) \leftarrow (1-\varepsilon)S_1(k) + \varepsilon S_2(k),$$

where

$$\varepsilon = \frac{\mathbf{AE}(d)}{\mathbf{AE}(d_0)}$$

is the weight of interpolation derived from the amplitude envelope, instead of being equal to d/d_0, where d is the temporal distance of $S(0)$ from $S_1(0)$, and d_0 is the temporal distance of $S_2(0)$ from $S_1(0)$; the interpolation is valid over a monotonic

segment of the amplitude envelope **AE**. For the decay, ε is obtained in the same way. Monotonicity guarantees $0 \leq \varepsilon \leq 1$ (see figure 3.10).

First, we remark that the attack timbre was synthesized in this manner with only the first and eleventh frames (the peak being in the eleventh frame), with the decay frames simply attached from the original. The stimulus was presented to five listeners* in an A-B forced-choice, randomly-ordered paradigm, where one event was the original and the other the synthetic. The response was unanimously random-guess, with the additional comment from the listeners that they could not tell the difference between the two stimuli.

Later, we synthesized the whole marimba tone with breakframes from the decay taken from the last frame (number 166) only, from number 166 and 80, from number 166, 80, and 40, and so on. Their distance was chosen to correspond to the decay characteristic; the attack portion was synthesized as described above. We presented a convergence test to several listeners **. The convergence test was to play pairs of tones. In each pair, one was the original and the other was synthesized from n breakframes, for n = 3, 4, 5, 6, and 7. Pairs involving different number of breakframes in the synthesized tones were played, in random order, and the listener was asked which of the two pairs sounded closer together. The response was unanimous; all the listeners found the pairs getting closer and closer together as the number of frames involved increased. All the listeners found the tones made with six or seven frames to be identical to the original, and most found the tones made with five frames identical to the original. Note that the frames in the five-frame configuration corresponded closely to where the three prominent partials change (or vanish). These examples were later presented at the ICMC '86 as well as in front of a local audience with a similar response.

The importance of this demonstration is:

(1) It shows for the first time that kinematic synthesis works.

(2) It shows the operational meaning of the amplitude envelope, and it is a measure of the success of the amplitude envelope analysis (which was by no means trivial, see the discussion in 3.5).

(3) It displays the practice of the notion of a coordinated strategy of anal-

* DAJ, MHF, DYL, JDH, and AN

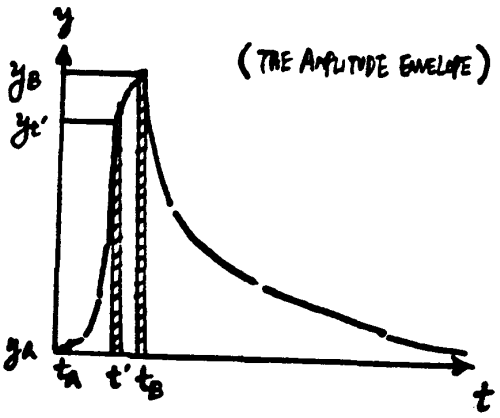
** DK, XJS, JDH, DYL, STK, and others

ysis and synthesis of timbre in terms of distinctive acoustic features, what they are, and how to use them.

(4) It shows the utility of a data versus algorithmic complexity trade. In this case, the amount of complexity increase is minimal.

(5) It provides a model for how the ear might organize the data it receives. It provides a model of the ear's update mechanism. Furthermore, it provides a logic for meaningful data reduction in the control space for synthesis.

(6) It provides an alternative view to Serra's noise interpretation of the marimba attack timbre. However it is perceived, the acoustic distinctive feature for the sharp or "noise-like" quality can be derived from the signal's organization. It further proves that the limitations encountered when one uses Fourier analysis, as reported by Serra [Serra, 1986], can be avoided when the acoustic data in the waveform is properly exploited.



Kinematic Synthesis by Dynamic Interpolation of Marimba Tone

Normally, kinematic synthesis of a frame of samples is driven by the weight

$$\frac{t' - t_A}{t_B - t_A}.$$

With dynamic interpolation, the weight is determined by the amplitude envelope over regions where it is monotonic; the weight is then

$$\frac{y_{t'} - y_A}{y_B - y_A}.$$

Using this technique, the attack timbre can be duplicated, i.e., generated with no perceptual distortion from the original, using only the first and the peak frames (the peak frame is the one containing the peak of the amplitude envelope). Furthermore, six or seven frames are all it takes to recover some 188 frames of the entire waveform and a sound which is perceptually indistinguishable from the original.

figure 3.10

3.11 Phase Preserving Frequency Resampling and Reshuffling.

Before we learn how to extract the detailed time-scale information from the waveform for a full period trajectory detection analysis, it might be beneficial to look into the discrete Fourier transform (DFT) method for transformation of frames with substantial length changes. For example, the /a/ with an octave drop retains the same vowel character in the course of the drop. Our knowledge in the spectral domain tells us that the formant structure is invariant throughout. Therefore, it is advantageous to take a frame in the high-pitched quasi-stationary regime, DFT it, and resample (at constant sampling frequency) the real and imaginary parts of the transform to the new length in such a way that the "envelope" is kept constant. The inverse transform will be the desired frame of the given length.

Although formant invariant frequency resampling with changes in the fundamental frequency is an idealization, and results in significant changes in timbre for large changes in fundamental frequency, the vowel quality is preserved in our studies. Doing a DFT to a period on the order of a hundred samples is not a computational concern. The phase necessary for successful frame to frame transition is preserved in frequency resampling. However, this is not a general property for frequency reshuffling. Moving DFT components around can destroy the phase of continuity critical in a kinematic synthesis. Frequency resampling for an octave drop also has a simple time domain interpretation: it amounts to an amplitude modulation by a cosine-type wave of twice the frame length. We have applied the simple time domain technique* to the marimba and /a/ timbres. The results can be described as distinctly different timbres but with similar characteristics, presumably those of the spectral envelopes.

An interesting question arises as to whether these new tones are closer to the original than those obtained by simply stretching the time-scale of the frames, or vice-versa. It is quite obvious, after a hearing, that most people would correlate the one with frequency resampling to the /a/ more than with the time-scale changed one. But different listeners give different responses to the tones derived from the marimba. The controversy arises probably from the fact that the frequency distribution in the marimba is sparse; therefore, a spectral envelope is not very meaningful.

* This is as opposed to other more complicated techniques (e.g., frequency resampling) that involve discrete Fourier transformation and inversion.

As to the question of which one is closer, the modified timbres are probably in different regions ("dimensions") of the timbre space. That is, they may be two points on a sphere (or higher dimensional analogue) with the natural marimba timbre at the center. The only rigorous way to find out is by some kind of perceptual distance measurement using interpolation.

Frequency resampling can be accomplished by (in pseudo-Pascal code):

```
DFT(length1, Input, RealPartTransIn, ImagPartTransIn);
Interpolate(length1, length2, RealPartTransIn, RealPartTransOut);
Interpolate(length1, length2, ImagPartTransIn, ImagPartTransOut);
IDFT(length2, Output, RealPartTransOut, ImagPartTransOut);
```

Here, `DFT` does the discrete Fourier transformation on `Input`, storing the results in `RealPartTransIn` and `RealPartTransOut`, `IDFT` does the inverse transform, storing the results in `Output`, and the code for `Interpolate` is something like:

```
procedure Interpolate(lengthIn, lengthOut, In, Out);
...
  begin
    for count:=0 to lengthOut - 1 do
      begin
        RealIndex:=count * scale;
        index:= trunc(RealIndex);
        fraction:= RealIndex - index;
        Out[count]:=In[index]*(1-fraction)+In[index+1]*fraction;
      end;
  end;
```

Frequency reshuffling is an important technique, especially for the purpose of timbre interpolation, that needs to be studied rigorously. It is important for timbre interpolation because if the ear retains a strong imprint about certain "place" characteristics in the basilar membrane corresponding to a particularly frequency distribution important to the ear (note that the membrane is known to exhibit greater sensitivity in the frequency range from one to four kiloHertz), then these characteristics become dominant features which are strongly retained. That is, modification of other features may not receive sufficient attention to have an effect in the overall perception of the timbre as it is changed in this way, and this remains

the case even when the strength of this acoustic feature is proportionately reduced. In the theory of dynamics, the "place" characteristic is termed an attractive basin (or said to be in an attractive basin—see [Thom, 1975]). If the modification is in the form of interpolation, the ear is likely to be continually pulled to this dominant attractive basin as other features are being modified. And the result is considered categorized. But of course, until we have tried modification which has a real effect of "pulling" out from the attractive basin, we don't know whether it is possible to interpolate for this highly resistant timbre to another timbre which is not in this attractive basin. One way to "pull" out from the attractive basin is not only to reduce the strength of the feature, but also to populate features in a growing neighborhood of the basin with gradually increasing strength. This transition may prevent the impression of "jumps" in the sequence of interpolated timbres.

The problem is a critical one insofar as timbre interpolation is concerned, because it has been implied [Risset and Wessel, 1982] that the problem of interpolation in the manner Grey approaches it has been solved. Grey's approach not only took very little account of the dynamic nature of timbral relationships (and therefore always used the attack, and in particular the epoch of the attack as an anchor point), but also used a technique which does not dynamically manipulate the frequency distribution of the signal. Specifically, the frequencies being populated are the combined totalities of the partial frequencies of the end tones. The transition is accomplished essentially by fade-in and fade-out of these partials. The partials existing strongly in both sounds will have additional shifts in their epoch positions. Those existing in only one or the other will simply fade out as the interpolation moves away from the sound that contains them. There is no repopulation of energy in the frequency domain as described above.

While we are convinced that the problem of interpolation is in general more difficult than is implied by Risset and Wessel, we are also aware of the possibilities for circuimventing the difficulties. One such was is by frequency reshuffling. The technical problem that needs to be studied extensively is to determine under what conditions reshuffling will preserve the phase of the transform, and thereby preserve the continuity of frame evolution. At present, such knowledge is scanty.

3.12 Proposed Test of Perceptual Importance Trees.

In chapter II, we developed the notion of feature composition as a way to describe timbre in terms of its features in absolute terms. The recursive nature of the composition process lends itself to a hierarchical description by way of importance trees.

An importance tree is essentially a data structure for a particular class of perceptually relevant acoustic features represented in a two dimensional cartesian coordinate system. In other words, a tree for us is a two dimensional graph whose nodes are either points or lines, depending on the application, such that the ordinate denotes the order of the node's importance in the class of features being delineated; the abscissa, in the case of a point, denotes the relative time at which the feature event actually appears in the timbre; and the length, in the case of a line, denotes the duration of the acoustic feature. The tree can be ordinal or metric. An ordinal tree provides a convenient way for a user to manipulate and synthesize timbres from recorded sources. It acts as a road map for the user in these processes. The user need only rely on his or her ear. On the other hand, a metric tree must be established by psychoacoustic experiments.

In 2.5.2, we presented a couple of model trees in which a node is represented as a line denoting the feature's duration. Theoretically, we might want to deduce an importance tree for each component of the triple. Furthermore, we could even have an importance tree for each breakframe pertaining to a quasi-stationary regime. Such an importance tree would normally represent spectral feature composition, or feature composition along the "place" dimension of the space-time single input multiple output response (SIMOR) pattern. In practice, however, we believe it is more pragmatic to first analyze a timbre into a tree of feature segments just as they are represented in the figures in 2.5.2 before we set out to determine the α_k's.

The important question is of course how do we determine the α_k's? Since the idea is so new, we don't have other models to follow. So we propose the following. Take a subtree at node k. If we want to measure the relative contribution of nodes $2k$ and $2k+1$ to the tree, we want to obliterate the effect of the feature at node $2k$ and measure the (perceptual) distance between the timbre A and its modified version A'. We want to do the same thing at the feature $2k+1$. Suppose we can measure these perceptual distances and put them into numbers, say α'_{2k} and α'_{2k+1}.

Then we can obtain α_{2k} and α_{2k+1} by normalizing the α''s with $\alpha'_{2k} + \alpha'_{2k+1}$.

Before we address the issue of distance measurement, we will first discuss the issue of feature obliteration. If we were to measure the importance trees on a component by component basis, obliteration would amount to simplification of the geometric form that gives rise to the perceptual feature. For example, if we were to measure the importance of the non-linear growth segment to the attack, we would simply replace the octave jump over the duration by a single line segment. But if we were measuring a tree such as in figure 2.5.2, then the first order of business would be to find out which component of the triple was contributing to the feature. In the case of the marimba, if we can decide that the amplitude envelope is responsible and the other members of the triple are not, then the problem is reduced to obliterating the feature on the amplitude envelope in this region. If we could not decide, then we could experiment with keeping the features in the amplitude envelope and simplifying the features in the other components in the same region.

Sometimes, isolated change is not possible. For example, flattening the local peaks in the frame amounts to flattening the amplitude envelope feature at hand. But then we could tell that the feature originated (or regard the feature as originating) from the organizing element of the timbre, namely the amplitude envelope. So modification of the amplitude envelope would account for the distance measured and hence the importance value derived later. In another situation, a vibrato is normally reflected by period length fluctuation together with amplitude modulation. Thus obliteration would mean equalizing the time-scaling as well as the amplitude-scaling of the acoustic features involved. The importance value derived may then be equally assigned to the appropriate node of the individual trees and to the combined tree we are mainly concerned with.

But how do we measure these distances? We suggest the possibility of interpolation. If we can generate an interpolated sequence between A and A', and A and A'', with imperceptible steps, then the minimum number of such steps will constitute a timbre distance between each pair, in some sense.

One might then question the approach on the grounds that the distance obtained might be path-dependent, therefore laboratory dependent. But path-dependence can happen only if we switch algorithms for interpolation. As long as it is done consistently, laboratory-dependence is not an issue. Usually, coherence

or the interpolated nature of changes in an acoustic waveform suggests that some compromise algorithm can be used both for the left subtree and the right subtree and throughout the entire tree.

Finally, we might want to start from the bottom of a tree for control reasons. The reasoning is based on the observation that Grey's work on interpolation is successful largely because of the proximity between the stimuli being interpolated. The absolute value of the α_k's of course can be obtained by normalization after the measurement is completed.

3.13 Test of Musicality.

We usually associate musicality of a timbre on two levels. On a lower level, we judge a timbre by itself, i.e., in an absolute sense, for its quality. This is essentially the way Helmholtz defined musical timbre. On a higher level, we judge a timbre against a timbral context, i.e., in a relative sense, for its belongingness. A typical example is in the music of William Schottstaedt. The timbre of a typical acoustic event in many of his works has a certain amount of a noisy quality that would not have qualified the event to be in Helmholtz's class of musical timbre. Nevertheless, the author finds these works musical, as do many other listeners.

In chapter II, we discussed musicality on both levels and saw them as consequences of the same psychoacoustic principle. This principle addresses the ear's ability to organize the single input multiple output response (SIMOR) pattern of external acoustic events. On a more local level, "clicks" and "pops" are microevents that do not "belong." We say that the ear cannot organize the images of these "clicks" and "pops" into the rest of the SIMOR patterns. In order to test this hypothesis correlating organizability of the acoustic signal to musicality, we proposed the generation of a sequence of timbres by means of arranging a sequence of pulses of 0° or 180° phase according to the (coefficients of the) Rudin-Shapiro polynomials. These polynomials are defined via a recursion:

$$p_0(z) = q_0(z) = 1$$

and recursively

$$p_{n+1}(z) = p_n(z) + z^{2^n} q_n(z)$$

and

$$q_{n+1}(z) = p_n(z) - z^{2^n} q_n(z).$$

The first few polynomials are given by

$$p_0(z) = 1, \qquad q_0(z) = 1,$$
$$p_1(z) = 1 + z, \qquad q_1(z) = 1 - z,$$
$$p_2(z) = 1 + z + z^2 - z^3, \qquad q_2(z) = 1 + z - z^2 + z^3,$$

and

$$p_3(z) = 1 + z + z^2 - z^3 + z^4 + z^5 - z^6 + z^7,$$
$$q_3(z) = 1 + z + z^2 - z^3 - z^4 - z^5 + z^6 - z^7.$$

We are interested, of course, in the coefficients of the polynomials so defined. For p_1, we have two terms with coefficents +1 and +1. The corresponding pulse pattern is a pulse of 0° phase followed by another pulse of the same phase. For q_1, we have two terms with coefficients of +1 and -1. The corresponding pulse pattern is pulse of 0° phase followed by a pulse of 180°. Both the pattern corresponding to p_1 and the pattern corresponding to q_1 are faintly pitched click-tones if the pulse separations are below six to seven milliseconds. Furthermore, the pitches can be clearly heard in context. That is, we can form a melody using well controlled pulse separation over the range of validity. Each pulse is of course wide-band, therefore the Fourier transform of any pulse pattern generated according to the p_n's and q_n's as above is of course wide-band. The interesting point is that, although the local appearance of any of these patterns is that of a pulse pair (which one can easily isolate and play without incurring extra timbre typical in similar operations on continuous waveforms—this is the reason we chose a pulse sequence for the test), the global pattern becomes increasingly complex and unorganizable to the eye. The question is, while the pulse pairs themselves can be musical, is the global pattern organizable to the ear?

We notice that the patterns are perfectly organizable to the logical part of the brain because we can summarize the pattern behavior elegantly in the form of a recursion (with initial conditions). We can just as easily apply this organization to generate the pulse pattern resonse as a function of n using simply shift, scale, and add operations. In this application, scale simply means scaling with +1 or -1. However, the increasing complexity in the pulse pattern is also accompanied by a rapid change in the timbre in such a way that by the time the index n has a value of six or seven, the timbre becomes totally obnoxious and repulsive. We have interviewed at least three listeners on their reactions to these timbre sequences. Their reactions were unanimous in the way described above. One listener suggested that the sequence was accompanied by "increasing bandwidth of noise character." Of course, the width of the frequency support is uniformly as high as the cutoff of the low-pass filter, although the spectrum certainly changes.

From the ear's point of view, whether it is a single pulse, a pulse pair, or an infinitely more complex pulse pattern, a broad response across the basilar membrane will be excited. But the temporal character is clearly different. The potential

dimensionality (or number of degrees of freedom) of a waveform is a linear function of the length of the waveform (compare the discussion in 2.5.4). Although the signal is known to the source—the generator, it is a statistical process to the ear—the receiver. If the receiver can organize the SIMOR pattern, it is equivalent to saying that the receiver is able to localize the timbre space in which the stimulus "lives." In the case of the Rudin-Shapiro polynomials, the ear is apparently unable to translate the logical organization of the pattern into an organization expressible in the auditory language. The lack of amplitude- or time-scale adaptable organizability typical in naturally occurring timbre, i.e., the interpolated nature of transitions in naturally occurring waveforms, and the lack of a semblance of periodicity, removes two timbral structural constituents, the amplitude envelope and the period trajectory, as organizing elements essential to the perception of the timbre as musical, according to our theory. It also suggests that the frames in these patterns of the polynomials of high degree are evolving too rapidly. This experiment suggests that the stability of our perception of a timbre is inversely "proportional" to the speed of frame variation

From a synthesis viewpoint, this experiment tells us that the breakframes selected for a timbre synthesis must be bridged by transitions of reasonable lengths. From a psychoacoustic viewpoint, this experiment shows that although acoustic events are fundamentally phenomena of fluid fluctuation, and timbre is the receiver's measure of the fluctuation and its evolution, and although the receiver is equipped with fibers to measure rapid fluctuation locally, the receiver cannot tolerate very rapid change in the overall SIMOR pattern. This tolerance or the lack of it sets a kind of qualitative limit on what timbre can be appropriately used as musical material.

3.14 Test of Timbral Features.

There are at lesat two ways to test whether a certain acoustic feature (or signal feature) is indeed a timbre feature. The first way is to show that it is essential to a perfect duplication of the original. This way is consistent with the spirit of Grey's timbre research requirement of having an approach for analysis and synthesis of timbre in terms of distinctive acoustic features.

A second way is to modify these acoustic features in a controlled manner and see if the timbres generated from such modification change smoothly, as expected. It is in accordance with these two methods that we test the components of the triple as predicted by our theory.

3.14.1 Amplitude Envelope.

To show that the amplitude envelope is a fundamental timbre feature, we first detect two very differently shaped amplitude envelopes, one from the marimba, as shown in figure 3.5(a), and another from the /a/ with an octave drop as shown in figure 3.14(a). These envelopes are applied to the marimba frames. The timbre associated with the attack and the decay are clearly changing smoothly even though the frames remain the same. Its indispensable role in the synthesis of a perceptually identical marimba tone has been discussed in 3.10 and other sections.

3.14.2 Period Trajectory.

To show that the period trajectory is a fundamental timbre feature, we interpolated a sequence of period trajectories based on those of the /a/ with an octave drop and the flat marimba trajectory. These were then applied to the marimba frames with constant (i.e., the same) amplitude envelope and to the /a/ frames with variable amplitude envelope. In either case, the timbre transition is smooth and the characteristic change due to the period trajectory change is clearly evident. The fluid texture usually ascribed to hybrid syntheses of this type (i.e., varying fundamental frequency with percussive attack) is striking in the marimba derivative. The crucial role the period trajectory plays in the duplication of the /a/ sound is discussed in detail in 3.11.

3.14.3 Frames and Alternating Timbre.

The critical role of breakframes in the resynthesis of the marimba, /a/, and derivative sounds is evident. Using the same /a/ frame, we are able to generate a series of /a/-related timbres, all with the same vowel quality. Furthermore, the role

of a frame is dramatized in the synthesis of the alternating timbres.

An alternating timbre is made from alternating two breakframes for a number of times, or for a fixed duration at a given rate of alternation. The frames in between are synthesized by interpolating between the two breakframes. In our demonstration, we have synthesized an alternating marimba timbre with the alternating breakframes derived from the eleventh frame (the peak frame) and the one-hundredth frame (or its neighbors). The alternation rate is $\sim$ ten cycles per second.

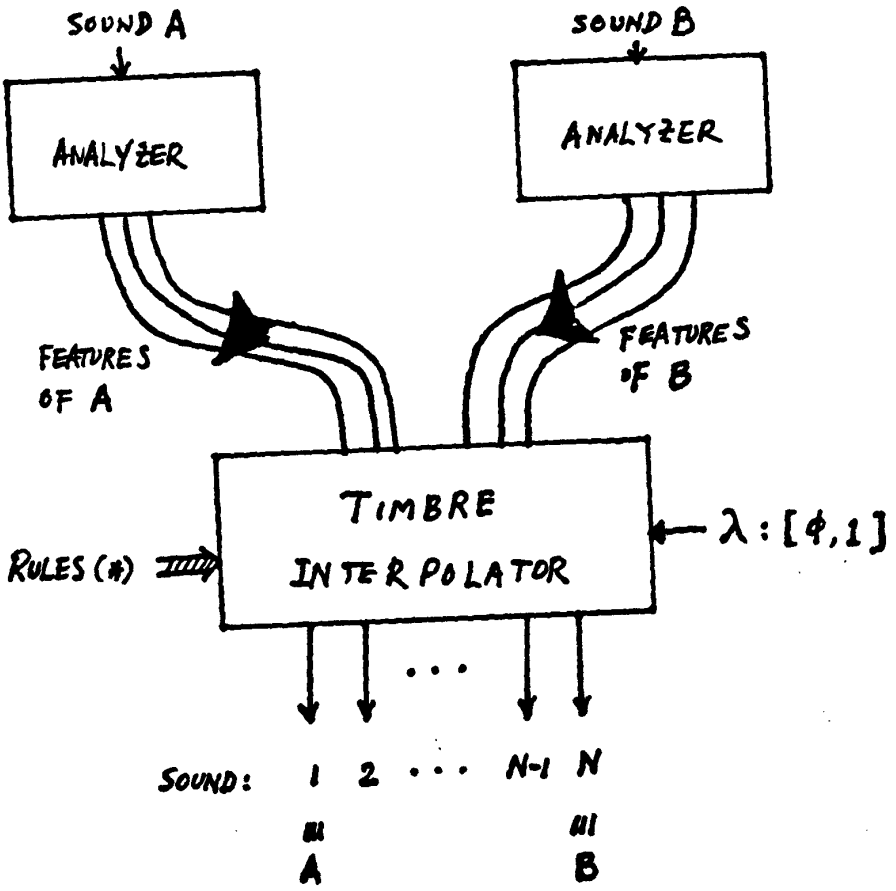
Another set of alternating timbres are generated from alternating the eleventh frame of the marimba and a frame representing the quasi-stationary region at the high part of the fundamental frequency trajectory of the /a/ at rates of five and ten units-per-second. The amplitude envelope and period trajectories are interpolations between those of the marimba and the /a/. An interesting feature of the first set of examples is the evidence of the dull marimba decay timbre which casual listening to the marimba tone would have certainly missed, and with which even careful listening would still have had difficulty. The second set of examples suggests that there might be a separation-time limit below which the timbre of different breakframes would be blurred. It is reasonable to expect that the less different the breakframes are, the higher the time limit.

The idea of an alternating timbre is actually a prototype for a more complicated composition of "spectral" timbre under rather arbitrary amplitude envelope and period trajectory conditions.

3.15 Test of the Parallelepiped Model of Interpolation.

Rodet, the author of the *Chant* program, in a seminar given at CCRMA in 1984, raised the question of whether there are holes in timbre space. Risset and Wessel [Risset and Wessel, 1982], on the other hand, strongly implied that timbres are in general interpolable. (These points are of course independent: if timbre space is shaped like a (higher dimensional) torus, then timbres can be interpolated; it is also conceivable that timbre space is in disconnected pieces, in which case timbres (at least timbres from separate components) cannot be interpolated.) Incidentally, Rodet, Risset, and Wessel were all at IRCAM (the French research conglomerate for computer applications to music and acoustics) in the early 1980s. Risset and Wessel's claim rested largely on Grey's work. We had (in 1984) informally repeated Grey's work (using Grey's approach but our own synthesis software) with an added stimulus—the timbre of a marimba tone softened up by reducing the intensity of of the tenth partial. The interest of this addition is to widen the range of timbres available for interpolation. Grey's library did not include a percussive timbre. But otherwise, the two sets of stimuli, including the tone qualities, are virtually identical by direct aural comparison. Therefore, we feel confident in making inferences from Grey's interpolation work whenever it is appropriate, based on our experience alone. In presenting the interpolation sequences to a group of listeners and obtaining their responses, we found that while the listeners in general found them musically interesting, some of them suggested the shortness of tones, $\sim \frac{1}{3}$ seconds long, made judgments about order difficult. In general, Grey's stimuli, like most of ours at the time (1984), are too short and too similar to really afford a clear picture on the issue of timbre interpolability. As a result, we don't know very much about the nature of the timbre space that they live in, even in their immediate neighborhoods. Thus there is indeed a controversy on the question of whether timbre is in general interpolable.

Before we can pursue this question further, the first order of business is to look at the general issues timbre interpolation is involved with. In our paper "Techniques for Timbral Interpolation" [Lo, 1986], we suggested the following blackbox view on the issue (see figure 3.15).



(*) RULES : (i) IN BETWEEN-NESS
(ii) SMOOTHNESS OF TRANSITION
(iii) WELL-BLENDEDNESS

figure 3.15

TIMBRE INTERPOLATOR

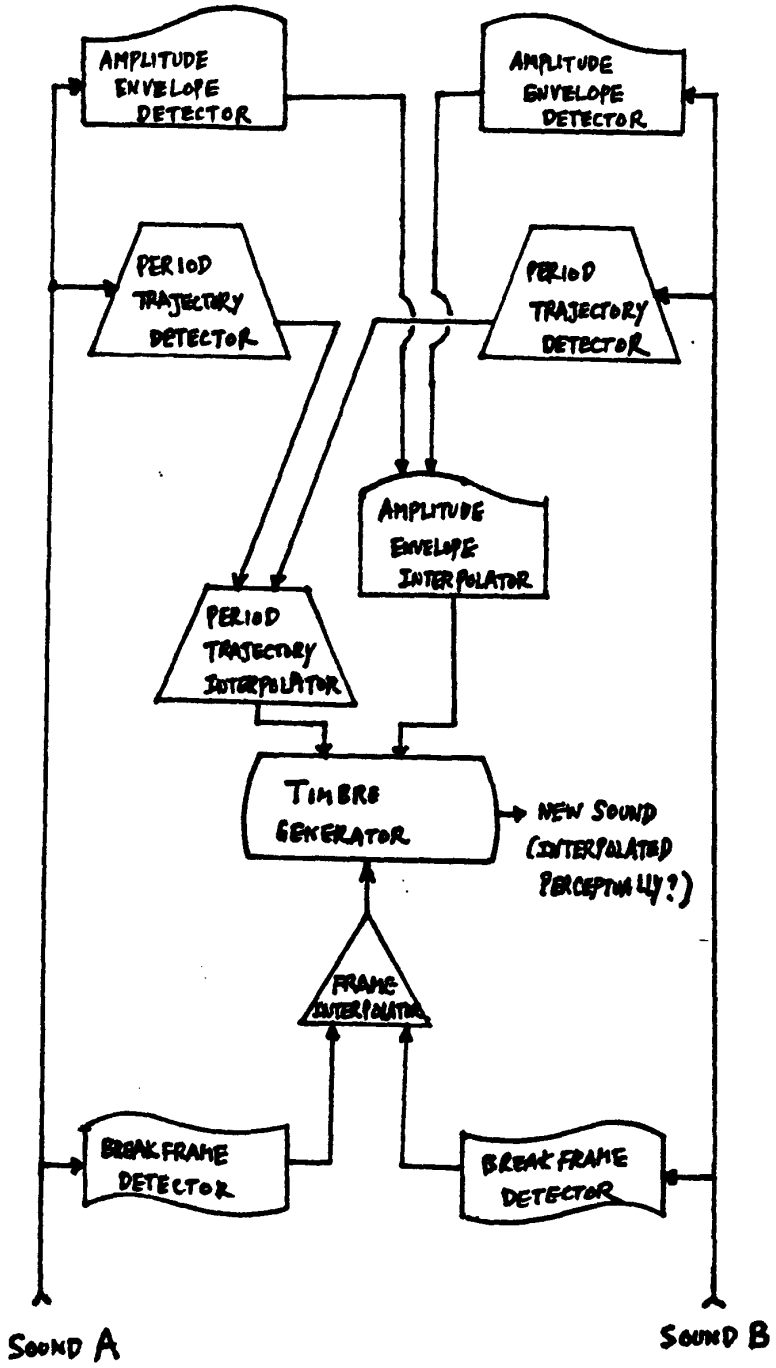


figure 3.15a

The blackbox is the interpolator. The input consists of two or more sounds called end sounds or anchors. The output is a sequence of new sounds that satisfy certain perceptual criteria. The interpolator includes an analysis box, a modification box, and a synthesis box. The analysis box takes the raw physical input data and generates knowledge about the timbral features of the end timbres, and ideally their relative importance and the perceptual distances between the features of one timbre and those of the others in the same class. The modification box, controlled by an interpolation index, generates modified features from features of the end timbres. In a highly developed system, we may expect the analysis box to supply the modification box with the algorithm of interpolation best suited for the given end timbres. The synthesis box takes the original and modified features and makes new timbres from them.

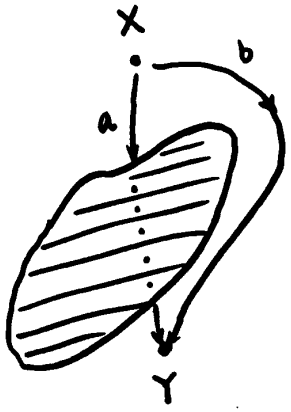
The perceptual criteria which govern timbre interpolation are:

(1) The betweenness criterion, which insures orderability by similarity judgment.

(2) The smoothness criterion, which insures an even spread over the distance between the end timbres. (This is usually the criterion to judge whether our perception is categorical in the sense that timbres are stored in a discrete cognitive map such that timbres generated from in between acoustic features are perceived as strongly similar to either side even though they still satisfy the betweenness criterion.)

(3) The well-blendedness criterion, which requires that each tone generated sounds fused in the sense that it does not sound like a superposition of sounds from different sources.

There are many good reasons to study interpolation. One primary reason is to allow the composer to efficiently use a few timbres of diverse characteristics for composition use. To be effective, we must have the ability to slowly (or subtly) *shift* the timbral character in some preferred direction; we must also be able to create timbres of *measured* contrast. In other words, to be able to use a few timbres of diverse characteristics for composition, one must have a good *constructive* understanding of the local timbre space containing the anchors.



(SEE TEXT AT RIGHT)

But then we must know whether timbre space is multiply connected (i.e., has holes) as Rodet suggests. From the mapping argument following Shannon (see 2.5.3), timbre space should be at least connected, even if there are holes (i.e., even if it is not *simply* connected). This means that an interpolated path may be long but it should be nevertheless interpolated.

But we should also understand that not every path is an interpolated path. For example, we might have a situation like that represented to the left. In this case, interpolated acoustic features inducing path a by a certain algorithm will not produce an interpolated sequence of timbres. In general, we have to search for a path like b.

Having a perceptual importance tree for X and Y helps us identify the features according to their relative importance in the timbre. As a result, bridging features of similar importance in similar temporal regimes (e.g., a feature which is a descendant of the left subtree of timbre X does not mix with one which is a descendant of the right subtree of timbre Y at the same level) in both the importance dimension and the temporal dimension would be the primary objective of timbre interpolation.

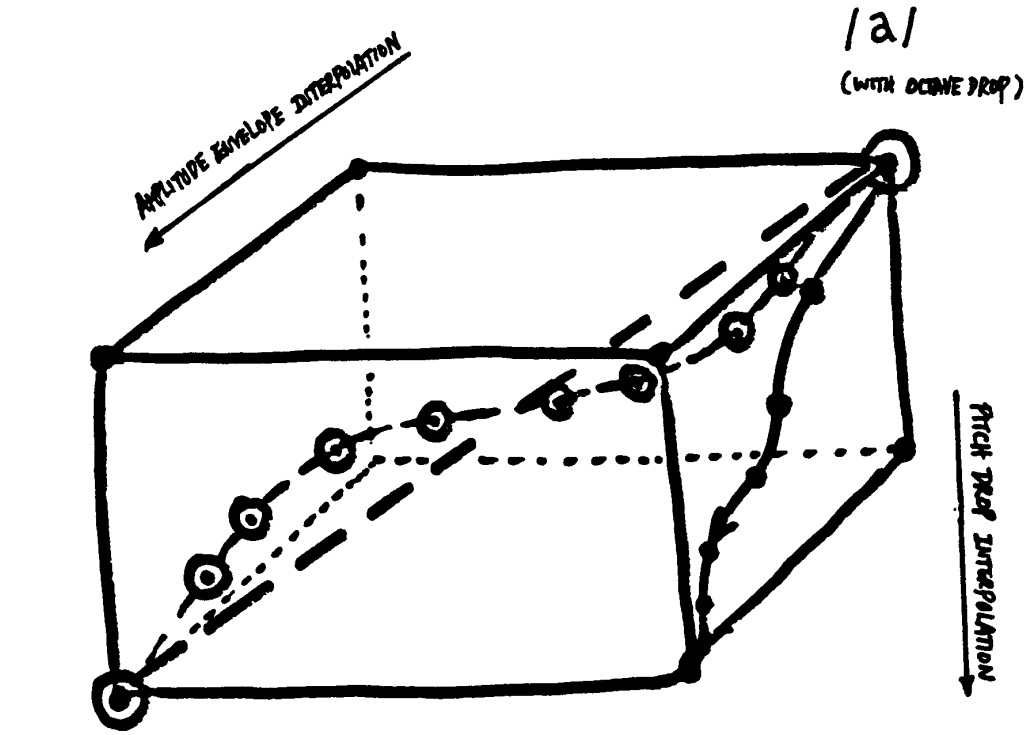
Bridging means more than simply weakening one feature and strengthening another in some isolated sense, as we have seen in 3.11. It might mean bridging over some neighborhood in the single input multiple output response (SIMOR) domain. We have not had a chance to test this idea, which includes perfecting the entire framework as we have presented it piece by piece in this theory. But the suggestion should provide a signpost for future work.

Next, we will examine the parallelepiped model of interpolation proposed earlier [Lo, 1986]. The model suggests that an interpolated timbre sequence can be obtained by comparing the interpolated components of the triple. Each component of the triples for the end timbres is interpolated individually and then a kinematic synthesis is performed according to a triple obtained from the interpolated components. Customarily, one would expect interpolated features of the same interpolation pa-

rameter to combine. But our experience is so limited, there is no way to know what perceptual effects will arise as we experiment with different combinations. Different combinations may lead to significantly different paths.

Of course, we have generated interpolated sequences along the amplitude envelope dimensions and the period trajectory dimensions for the purpose of demonstrating that these components are distinctive features (see figure 3.15(b)). They apparently satisfy the criteria listed above. In general, fusion is not a problem when either the amplitude envelope or the period trajectory is modified and wholly applied to the synthesis. But if timbres are generated with more than one organizing element in the period trajectory or the amplitude envelope, fusion often fails. The real test of interpolation, i.e., one that involves all three components of the triple, seems more difficult. Although the first and third criteria are met, there are some questions concerning the smoothness of the transition, even after great effort. The reason, we believe, is largely rooted in the strong attractive characteristic of the energy at 2500 Hertz for the marimba as we mentioned in 3.11. Our approach, similar to that of Grey's discussed above, does not take care of repopulation of the energy in an increasingly larger neighborhood of 2500 Hertz as it moves towards the /a/.

Another technical difficulty is a lack of a more manipulable tool to transform frames for the /a/. As tools are further developed, we will be better able to test the parallelepiped model of interpolation.



IDEAL INTERPOLATION PATH : PERCEPTUALLY SMOOTH AND DIRECT

INTERPOLATION PATH WITH STRONG PULLS TOWARD ANCHOR TIMBRES

INTERPOLATION WITH CONSTANT LIST OF BREAKFRAMES .

"VECTOR-ADDITION" MODEL OF TIMBRE INTERPOLATION

figure 3.15b

3.16 Conclusion.

We have presented a dynamic theory of timbre in the form of a framework. This framework can be summarized by the tetrahedral block diagram below.

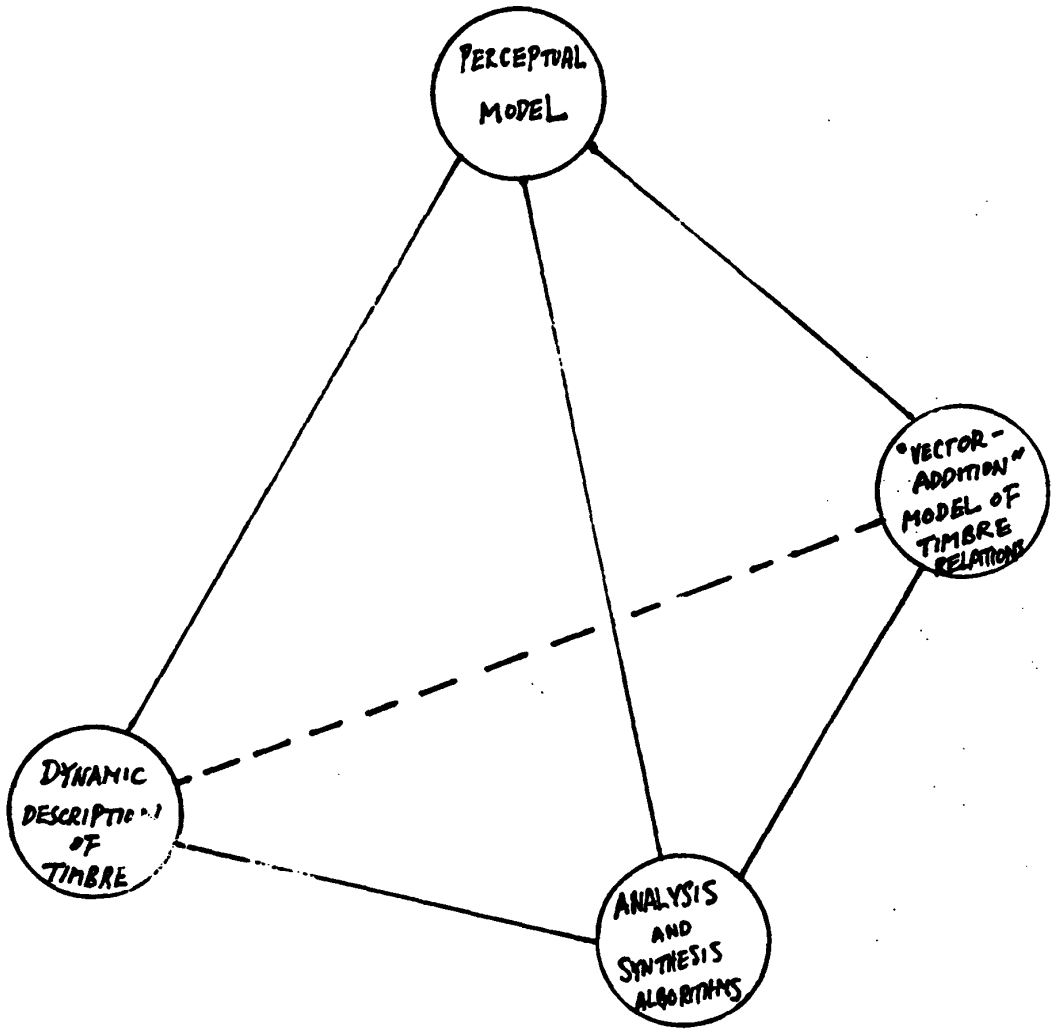


figure 3.16

The perceptual model essentially consists of the well-connected single input multiple output response (SIMOR) pattern based on Helmholtz's mechanical model of the basilar membrane superposed with an active observer that is able to adapt, organize, and do pattern recognition on the SIMOR. This perceptual model leads us to arrive at the physical features in the acoustic waveform that we can analyze directly as timbral distinctive features, i.e., acoustic features distinctive for timbre. This method is copied from Helmholtz's passive observer model of timbre perception. The derivation establishes the timbre language we sought, that is, a language which describes (1) a direct correlation between timbre features and acoustic features, and (2) a direct correlation between the composition of timbre features and the composition of acoustic features. In particular, we concluded from our perceptual model that the amplitude envelope, the period trajectory, and the breakframes play a fundamental role.

We have also developed a quantitative way to describe timbral feature composition in terms of an importance tree whose structure is determined by, and can be modified as a result of, experiment. Such a tree is easy to think about and is directly applicable to the problem of analysis and synthesis of distinctive features of timbre. The development of a comprehensive approach to an analysis/synthesis process in terms of distinctive features is directly dependent on the dynamic description of timbre and the active observer model of perception. The implementation of the analysis and synthesis algorithm has shown some promise as we discussed in previous sections and the accompanying demonstration.

The twenty-five to thirty timbres (duplicated and modified) from the analysis of two digitally recorded timbres shows that there is some validity in both the idea and the implementation of the analysis/synthesis approach. We have validated via demonstration the description by triple, the idea of data versus algorithmic trade, and a rational way to achieve data reduction in control space consistent with perceptual principles. Finally, the perceptual model, the dynamic description, and the analysis/synthesis all contribute by way of the parallelepiped model of interpolation to present a reasonably consistent way of describing timbre relationships, in terms of internal timbral dynamics. In describing these relationships, the feature trees of the triples and the parallelepiped or "vector-addition" model provide the road map, the direction, and the strategies.

Much remains to be done. The conditions for phase-preserving frequency reshuffling and processing must be worked out. It is essentially an applied mathematics or signal processing problem. It is believed to be important for certain timbre interpolation problems.

Timbre interpolability remains an unresolved question in psychoacoustics. It is of great interest to see if our theory would further our insight into this question. Interpolation is also an important feature in defining a metric importance tree. Therefore, tests for some aspects of the theory depend on an affirmative answer to the question of timbre interpolability. Given our current belief that the dimensionality of timbre space is much smaller than that of the physical control space, we believe the interpolability issue will be resolved in the affirmative once we have a good understanding of the nature of timbral distinctive features. This thesis of course sets out to obtain some understanding of these features, and we believe it is a matter of perfecting the analysis and synthesis techniques to make the understanding (more nearly) complete. We also believe much progress can be made in this understanding just using the ordinal information from the importance trees.

On the analysis front, there are two points that must be addressed. First, in order to discover all the organizing features related to scaling in time, we must have the capacity to perform a more detailed analysis of the time-scale changes in the acoustic features as part of the period trajectory analysis. We have referred to these very local but nevertheless vital features as secondary. Second, we have proposed a breakpoint selection algorithm in 3.7, but we must discover suitable error criteria so that we can automate the detection of breakframes. This is a decidedly non-trivial proposition, and of course much work must be done before any definitive conclusions can be drawn on the nature of these error criteria.

A logical extension to our theory would result from discovering the rules that govern the generative aspects of timbre, based on the dynamic triple description. In other words, under what general conditions will combinations of elements in the triple lead to desired timbral qualities? For example, a certain amplitude envelope combined with a certain local pattern of vibration (i.e., a certain spectral content) leads to surprising effects (experience has shown that a highly percussive amplitude envelope produces hollow sounding timbres when applied to slowly changing local

patterns of vibrations, such as that of the human voice—does this have a broader implication in that the amplitude envelope and period trajectory as organizing elements must remain fairly steady with respect to at least some of the spectral components so that some fibers in the basilar membrane would get a chance to ring sufficiently to feed that sensation which we describe as musical?). These effects can work in combination with other timbres to achieve certain musical functions much as certain chords do in a harmonic progression. But they must occur at "correct" places. Therefore, a natural extension to our theory is examination of the issue of "musical grammar" on the micro-temporal level. We expect that a good understanding of this type of musical grammar would lead to meaningful attempts in the direction of timbre composition as described below.

Finally, an important contribution a theory of this kind can make is to provide a timbral operating environment suitable for timbre composition. As we have pointed out in 1.1, Schönberg's much cherished dream of timbral composition remains unfulfilled. In 1.6.1, we have offered a conjecture for why this has been the case. We suggested that at least part of the problem stems from the fact that timbre perception is a dynamic process. Therefore, a timbre environment which does not take into proper account the diversity of timbre and the dynamical nature of timbre relationship, which relies on a fixed set of tools, and which is computationally intensive, will probably discourage interested composers from making attempts in this direction, i.e., attempts that exploit timbre patterns as driving or shaping elements of a composition.

Now that we have gone to great lengths to justify, from perceptual considerations, a dynamic description and have laid a foundation for a presumably more adequate analysis and synthesis strategy, it is appropriate to suggest that the next step in any effort of this kind might be for someone interested to try to make compositions in which timbre functions as a moving force, much as pitch has in western music, that shapes the development of the composition and at the same time holding the piece structurally together.

We pointed out in 1.6.1 that certain ancient Chinese poems appears to be a genre of timbre composition in the sense described above. It would therefore be a worthwhile project for someone who is interested to analyze some of these poems in the sense of abstracting the relationships among the triples of neighboring

timbres, one, two or more steps removed, and summarizing them in the matric forms of 3.8 or their graphical equivalents. These timbres, embodied in monosyllabic words are full of timbre nuances. These nuances in the form of stresses, intonations and diphthongs are reflected in the amplitude envelope, fundamental frequency trajectory and spectral evolution of the timbres. Hopefully, new music may result from such analyses. And hopefully, some of these pieces of music will bring us closer to the dream of Schönberg.

APPENDIX A

A listing of the frame boundary marking algorithm, implemented in the Stanford Artificial Intelligence Language (SAIL), follows.

```
comment mle(groping,index,window,winopt←2,alpha,beta,delta,gamma,pest,pdevu,pdevl,endseq,ttystr);

boolean procedure mle( reference boolean groping;
                       integer ref,window,winopt;
                       real alpha,beta,delta,gamma;
                       reference integer pestio,pdevu,pdevl;
                       reference real array endbuf;
                       reference string errmes);

begin "MLE"

    integer low,high,j,k,Pest,Nest,Rest,first,MLEloc,perr,pestin,dummy,ra,a,qm,
            minsubhar,minsubharindex,nsubhar;
    integer array subharloc[1:128];
    real array subharval[1:128];
    real MLEval,r,s,tmp,minsubharval;
    real array CBuf[0:511];
    real array RJ[0:511];

    pestin←pestio;
    low←(pestin-pdevl) max 2;   high←(pestin+pdevu) min (2*pestin);

    if mledebug then begin
        print(crlf,"pestin",tab,"pdevl",tab,"pdevu",tab,"low",tab,"high");
        print(crlf,pestin,tab,pdevl,tab,pdevu,tab,low,tab,high);
    end;

    case winopt of
    begin "list"
        [0] first←ref-(window div 2);
        [1] first←ref;
        [2] first←ref-window;
        else                        ! ignore the rest;
    end "list";

    if groping then first←ref;      ! use future data hoping that things will get
                                      better when you are groping, in D.H.'s words;

    if first<bosloc then first←bosloc;

    if (window div pestio)<2 then begin "warning"
        errmes ← "window segment to cycle period < 2: "&cvs(window)&"/"&cvs(pestio);
        return(false);
    end "warning";
    if (window div pestio)>4 then begin "warning"
        errmes ← "window segment to cycle period > 4: "&cvs(window)&"/"&cvs(pestio);
        return(false);
    end "warning";

    for pest ← low thru high do begin "compute estimate"

        arrclr(CBuf);
        Nest ← window div Pest; Rest ← window mod Pest;
        if debug then begin "print"
            print(crlf,"Ref",tab,"Window",tab,"Pest",tab,"Nest",tab,"Rest");
            print(crlf,Ref,tab,Window,tab,Pest,tab,Nest,tab,Rest);
        end "print";
        for k ← 0 thru ( Window - 1 ) do begin "add aliases"
            j ← k mod Pest;
            CBuf[j] ← CBuf[j] + SndBuf[First+k];
        end "add aliases";

        if mledebug and debug then begin "print"
            print(crlf,"SUM OF ALIASES CBUF() FOLLOWS:",crlf);
            for j ← 0 thru ( Pest - 1 ) do print(CBuf[j],tab);
```

```
        end "print";

        e ← 0;
        r ← 0;
        if rest=0 then begin "get norm for imperfect cycle"
            for k←0 thru rest-1 do e←e+CBuf[k]↑2;
            if nest>0 then begin "at least 1 whole cycle"
                for k←rest thru Pest-1 do r←r+CBuf[k]↑2;
                e←e/(nest+1);
                r←r/nest;
            end "at least 1 whole cycle";
        end "get norm for imperfect cycle"
        else begin "get norm for perfect cycle";
            for k←0 thru Pest-1 do e←e+CBuf[k]↑2;
            e←e/nest;
        end "get norm for perfect cycle";
        RJ[pest-low] ← e + r;
        if mledebug then print(crlf,"MLE function(",pest,") = ",RJ[pest-low]);

    end "compute estimate";

    ifc false thenc begin

    @ ==========================================================================;

        e ← 0;
        for k ← 0 thru ( high - low ) do begin "subtract floor"
            e ← ( e * k + Rj[k] ) / ( k + 1 );
            Rj[k] ← Rj[k] - e;
        end "subtract floor";

    @ ==========================================================================;

    endc;

    if mledebug and debug then begin
        print(crlf,"THE NORMALIZED J(tau) FOLLOWS:",crlf);
        for k ← 0 thru ( high - low ) do print(Rj[k],tab);
    end;

    MLEval←MaxAmp(0,high-low,0,MLEloc,Rj);

    if fatmax(0,high-low,MLEloc,MLEval,alpha,beta,Rj) then begin
        pestio ← Moment(1,low,0,high-low,Rj);
        print(crlf,tab,"=== too fat ===");
        print(crlf,"from Moment: ",pestio);
    end else begin
        pestio←MLEloc+low;
        if debug then print(crlf,"from Peak: ",pestio);
    end;

    if groping and minV then begin "minV"

        print(crlf,"minvar!");
        m←1;
        dummy←low+MLEloc;
        while dummy≥(low max 4) do begin "loop"
            tmp←0;
            for k←ref thru ref+dummy-1 do begin
                tmp←tmp+(endbuf[k+dummy]-endbuf[k])↑2;
            end "for";
            tmp←tmp/dummy;
            subharloc[m]←dummy;
            subharval[m]←tmp;
            if mledebug then print(crlf,m,tab,dummy,tab,tmp,tab,tmp/dummy);
```

```
            m←m+1;
            dummy←(low+MLE[loc)/m;
        end "loop";
        nsubhar←m-1;
        if mledebug then print(crlf,"number subharmonics: ",nsubhar);
        minsubharindex←1;
        minsubharval←subharval[minsubharindex];
        if nsubhar>1 then begin "find max subhar"
            for k←1 thru nsubhar do begin "k"
                if subharval[k]<minsubharval then begin "<"
                    minsubharval←subharval[k];
                    minsubharindex←k;
                end "<";
            end "k";
        end "find max subhar";
        minsubhar←subharloc[minsubharindex];
        print(crlf,"minsubharindex",tab,"minsubhar",tab,"minsubharval");
        print(crlf,minsubharindex,tab,minsubhar,tab,minsubharval);

        if nsubhar>1 then begin "pick"

            for k←1 thru nsubhar do if subharloc[k]=pestlo then done;
            print(crlf,"pestlo",tab,"k",tab,
                    "subharloc[k]",tab,"subharval[k]",tab,"minsubharval");
            print(crlf,pestlo,tab,k,tab,
                    subharloc[k],tab,subharval[k],tab,minsubharval);
            if k>nsubhar then begin
                print("========== error in MLE_minv ==========");
                exit;
            end else begin "ok"
                if abs(minsubharval-subharval[k])>(tmp←gamma*minsubharval) then begin
                    print(crlf,
                    "minvar says pestlo should be ",minsubhar," instead of ",pestlo);
                    print(crlf,"minsubharval",tab,"minsubharindex",tab,"k",
                            tab,"subharval[k] (for pestlo)","gamma*minsubharval");
                    print(crlf,minsubharval,tab,minsubharindex,
                            tab,k,tab,subharval[k],tab,tmp);
                    pestlo←minsubhar;
                    groping←true;
                end else groping←false;
            end "ok";

        end "pick";

    end "minV";

    if mledebug then begin "print"
        print(crlf,"Maximum of MLE function = ",MLEval);
        print(crlf,"Choice of MLE period = ",pestlo);
        print(crlf,tab,"==============");
    end "print";

    if abs(perr←pestlo-pestin)≤gamma*pestin then begin
        if mledebug then print("$$$$$$$$ converging $$$$$$$$");
        pdevu←(pdevu min perr) max (pestlo*gamma);
        pdevl←pdevu;
    end else print("$$$$$$$$ NOT converging $$$$$$$$");

    if mledebug then begin
        print(crlf,"pestin",tab,"pestlo",tab,"perr",tab,"gamma",tab,"gamma*pestin");
        print(crlf,pestin,tab,pestlo,tab,perr,tab,gamma,tab,gamma*pestin);
    end;

    return(true);
```

```
end "MLE";
```

```
comment trackcycles(first,last,endseq);

procedure trackcycles(integer first,last;
                      real array endseq );

begin "trackcycles"

    integer oldpest,index,pest,pdevu,pdevl,window,wpratio,winopt,count,diffest,k,l,
            l1,k1,zxloc,window0;
    real pctpdev,delta,gamma,alpha,beta,tmp,zxval,zxslope,cye,cye0,phashi;
    boolean zxtrue,groping,notshiftyet,doneyet;

    begin "get par";
                                        @       winopt from mle;
                                        @       ;
                                        @       [0] first ← ref - Window div 2;
                                        @       [1] first ← ref;
                                        @       [2] first ← ref - Window;
        winopt←2;
        ttystr←null;
        tmp←query("window option 0/1/2 for estm/intrp/pred ("&cvs(winopt)&"): ",ttystr);
        if ttystr then winopt←tmp;
        window0←500;
        ttystr←null;
        tmp←query("maximum window length ("&cvs(window0)&"): ",ttystr);
        if ttystr then window0←tmp;
        cye0←.01;
        ttystr←null;
        tmp←query("cycle energy(11) threshold("&cvs(cye0)&"): ",ttystr);
        if ttystr then cye0←tmp;
        alpha←.9;
        ttystr←null;
        tmp←query("shoulder height cutoff of MLE distribution("&cvs(alpha)&"): ",ttystr);
        if ttystr then alpha←tmp;
        beta←.5;
        ttystr←null;
        tmp←query("shoulder width cutoff of MLE distribution("&cvs(beta)&"): ",ttystr);
        if ttystr then beta←tmp;
        gamma←.05;
        ttystr←null;
        tmp←query("fractional deviation of MLE from guess("&cvs(gamma)&"): ",ttystr);
        if ttystr then gamma←tmp;
        delta←.2;
        ttystr←null;
        tmp←query("MLE octave doubling error criterion("&cvs(delta)&"): ",ttystr);
        if ttystr then delta←tmp;
        pest←query("supply your best guestimate of frame length: ",ttystr);
        pctpdev←query("enter the % length deviation from your guess: ",ttystr);
        pdevu←(pctpdev/100)*pest;
        pdevl←pdevu;
        wpratio←query("enter window to period ratio: ",ttystr);
        window←pest*wpratio;

    end "get par";

    bosloc←begsamp(first,endseq);
    print(crlf,"begin @",bosloc);
    index←bosloc;
    cystarts[0]←index;
@   eosloc←index;
    eosloc←last;
@   winopt←2;
    count←1;
    groping←true;
    notshiftyet←true;
```

```
    zxtrue←false;

∂   while not endofsnd do begin "not eos"
!
    while not (index>eosloc) do begin "not eos"

∂       if index>eosloc then endofsnd←testeos(eosloc,sndseq);

        ∂   eosloc is that sample location before snd at which sample
            magnitudes are still significant, therefore there would be
            no need to check end of snd if k≤eosloc;

        oldpest←pest;

        begin "find time to shift"

            cye←llsum(index,pest,sndseq);
            print(crlf,"ll cycle-energy: ",cye);
            cye←llsum(index,pest/2,sndseq);
            print(crlf,"ll cycle-energy: ",cye);

            if cye>cye0 and notshiftyet then begin "shift cystart"

                k←index;
                while k<index+pest do begin
                    if zxtrue←zx(k,zxloc,zxval,zxslope,sndseq) then done;
                    k←k+1;
                end;
                if zxtrue then begin
                    phashi←zxloc-index;
                    print(crlf,"~~~~~~~~~~~~~~~ phase shift ~~~~~~~~~~~~~~~~~ ");
                    print(crlf,"~~~~~~~~~~~~~~~ by ",phashi," ~~~~~~~~~~~~~~~~~ ");
                    print(crlf,"~~~~~~~~~~~~~~~ phase shift ~~~~~~~~~~~~~~~~~ ");
                    for k←0 thru count-1 do cystarts[k]←cystarts[k]+phashi;
                    index←zxloc;
                    notshiftyet←false;
                end;

            end "shift cystart";

        end "find time to shift";

        if ~ale(groping,
                index,window,winopt,
                alpha,beta,delta,gamma,
                pest,pdevu,pdevl,
                sndseq,
                ttystr)
        then begin "window correction"
            groping←true;
            window←pest*upratio;
            print(crlf,"groping: new windowlength: ",window,tab,"pest: ",pest);
        end "window correction" else begin "update?"
            if (sndseq[index+1]-sndseq[index])>0 and
               (sndseq[index+pest]-sndseq[index+pest-1])>0 then begin "update ok"
                index←index+pest;
            end "update ok" else begin "look further"
                k←index+pest;
                ll←gamma*pest;
                l←1;
                doneyet←false;
                while not doneyet do begin "while"
                    kl←k-l;
                    if (sndseq[kl]-sndseq[kl-1])>0 and
                       (sndseq[index+1]-sndseq[index])>0 then doneyet←true
```

```
                else begin "not down"
                    kl←k+1;
                    if (sndseq[kl]-sndseq[kl-1])>0 and
                       (sndseq[index+1]-sndseq[index])>0 then doneyet←true
                    else begin "up !"
                        l←l+1;
                        if l>11 then doneyet←true;
                    end "up !"
                end "not down";
            end "while";
            pest←kl-index;
            print(crlf,"result of look-further: cystart=",kl,tab,"pest=",pest);
            print(crlf,sndseq[index],tab,sndseq[index+1]);
            print(crlf,sndseq[kl-1],tab,sndseq[kl],tab,sndseq[kl+1]);
            index←kl;
            groping←true;
        end "look further";
        cystarts[count]←index;
        print(crlf,count,tab,cystarts[count],tab,pest);
        if not groping then window←window+(pest/oldpest);
        if window>window0 then window←window0;
        print(crlf,"** update ** new windowlength: ",window,tab,"pest: ",pest);
        diffest←abs(pest-oldpest);
        if diffest>pdevu then begin
            print(crlf,"MLE period differs from guessed period by ",diffest,"samps",
                  crlf,"while prescribed margin was ",pdevu,"samps");
            if not groping then window←window-(diffest/pdevu);
            if window>window0 then window←window0;
        end else groping←false;
        if (oldpest-pdevl)<pest≤oldpest then begin
            pdevu←(oldpest-pest) max 5;
            pdevl←(pest-(oldpest-pdevl)) max 5;
        end;
        if oldpest≤pest<(oldpest+pdevu) then begin
            pdevl←(pest-oldpest) max 5;
            pdevu←((oldpest+pdevu)-pest) max 5;
        end;
        if diedebug then print(crlf,ttystr,tab,"new windowlength: ",window,tab,"pdevu: ",pdevu);
        count←count+1;

    end "update?";

end "not eos";

for k←0 thru count-2 do cylens[k]←cystarts[k+1]-cystarts[k];

nframes←count;

print(crlf,"cycle starts");
print(crlf,cystarts[0],tab);
for count←1 thru nframes-1 do begin
    if (count mod 2)=0 then print(crlf);
    print(cylens[count-1],tab,cystarts[count],tab);
end;

if savdat then begin "savdat"
    string prfilnam;
    print(crlf,"saving amplitude envelope in a text file ");
    errclr(outbuf);
    print(crlf,"number of maxima: ",count);
    for k←0 thru count-1 do begin
        outbuf[k]←cystarts[k];
        if k≥count-1 then auxseqnl[k]←0
        else auxseqnl[k]←cylens[k];
    end;
```

```
        infilename←file:name[in];
        prfilnam←infilename[1 for 2] & "cin" & ".dat";
        wrtdata(auxseqn1,outbuf,1,count,prfilnam);
    end "savdat";

end "trackcycles";
```

from MARIMS.SND/DEFAULT WITH MINVAR OPTION(3 JUL 1986 at 12:53 am)
length: 183

0	1.00000	54.0000
1	55.0000	54.0000
2	109.000	54.0000
3	163.000	54.0000
4	217.000	54.0000
5	271.000	54.0000
6	325.000	54.0000
7	379.000	54.0000
8	433.000	54.0000
9	487.000	54.0000
10	541.000	54.0000
11	595.000	54.0000
12	649.000	54.0000
13	703.000	54.0000
14	757.000	54.0000
15	811.000	54.0000
16	865.000	54.0000
17	919.000	54.0000
18	973.000	54.0000
19	1027.00	54.0000
20	1081.00	54.0000
21	1135.00	54.0000
22	1189.00	54.0000
23	1243.00	54.0000
24	1297.00	54.0000
25	1351.00	54.0000
26	1405.00	54.0000
27	1459.00	54.0000
28	1513.00	54.0000
29	1567.00	54.0000
30	1621.00	54.0000
31	1675.00	54.0000
32	1729.00	54.0000
33	1783.00	54.0000
34	1837.00	54.0000
35	1891.00	54.0000
36	1945.00	54.0000
37	1999.00	54.0000
38	2053.00	54.0000
39	2107.00	54.0000
40	2161.00	54.0000
41	2215.00	54.0000
42	2269.00	54.0000
43	2323.00	54.0000
44	2377.00	54.0000
45	2431.00	54.0000
46	2485.00	54.0000
47	2539.00	54.0000
48	2593.00	54.0000
49	2647.00	54.0000
50	2701.00	54.0000
51	2755.00	54.0000
52	2809.00	54.0000
53	2863.00	54.0000
54	2917.00	54.0000
55	2971.00	54.0000
56	3025.00	54.0000
57	3079.00	54.0000
58	3133.00	54.0000
59	3187.00	54.0000
60	3241.00	54.0000
61	3295.00	54.0000
62	3349.00	54.0000

63	3403.00	54.0000
64	3457.00	54.0000
65	3511.00	54.0000
66	3565.00	54.0000
67	3619.00	54.0000
68	3673.00	54.0000
69	3727.00	54.0000
70	3781.00	54.0000
71	3835.00	54.0000
72	3889.00	54.0000
73	3943.00	54.0000
74	3997.00	54.0000
75	4051.00	54.0000
76	4105.00	54.0000
77	4159.00	54.0000
78	4213.00	54.0000
79	4267.00	54.0000
80	4321.00	54.0000
81	4375.00	54.0000
82	4429.00	54.0000
83	4483.00	54.0000
84	4537.00	54.0000
85	4591.00	54.0000
86	4645.00	54.0000
87	4699.00	54.0000
88	4753.00	54.0000
89	4807.00	54.0000
90	4861.00	54.0000
91	4915.00	54.0000
92	4969.00	54.0000
93	5023.00	54.0000
94	5077.00	54.0000
95	5131.00	54.0000
96	5185.00	54.0000
97	5239.00	54.0000
98	5293.00	54.0000
99	5347.00	54.0000
100	5401.00	54.0000
101	5455.00	54.0000
102	5509.00	54.0000
103	5563.00	54.0000
104	5617.00	54.0000
105	5671.00	54.0000
106	5725.00	54.0000
107	5779.00	54.0000
108	5833.00	54.0000
109	5887.00	54.0000
110	5941.00	54.0000
111	5995.00	54.0000
112	6049.00	54.0000
113	6103.00	54.0000
114	6157.00	54.0000
115	6211.00	54.0000
116	6265.00	54.0000
117	6319.00	54.0000
118	6373.00	54.0000
119	6427.00	54.0000
120	6481.00	54.0000
121	6535.00	54.0000
122	6589.00	54.0000
123	6643.00	54.0000
124	6697.00	54.0000
125	6751.00	54.0000
126	6805.00	54.0000
127	6859.00	54.0000

128	6913.00	54.0000
129	6967.00	54.0000
130	7021.00	54.0000
131	7075.00	54.0000
132	7129.00	54.0000
133	7183.00	54.0000
134	7237.00	54.0000
135	7291.00	54.0000
136	7345.00	54.0000
137	7399.00	54.0000
138	7453.00	54.0000
139	7507.00	54.0000
140	7561.00	54.0000
141	7615.00	54.0000
142	7669.00	54.0000
143	7723.00	54.0000
144	7777.00	54.0000
145	7831.00	54.0000
146	7885.00	54.0000
147	7939.00	54.0000
148	7993.00	54.0000
149	8047.00	54.0000
150	8101.00	54.0000
151	8155.00	54.0000
152	8209.00	54.0000
153	8263.00	54.0000
154	8317.00	54.0000
155	8371.00	54.0000
156	8425.00	54.0000
157	8479.00	54.0000
158	8533.00	54.0000
159	8587.00	54.0000
160	8641.00	54.0000
161	8695.00	54.0000
162	8749.00	54.0000
163	8803.00	54.0000
164	8857.00	54.0000
165	8911.00	54.0000
166	8965.00	54.0000
167	9019.00	54.0000
168	9073.00	54.0000
169	9127.00	54.0000
170	9181.00	54.0000
171	9235.00	54.0000
172	9289.00	54.0000
173	9343.00	54.0000
174	9397.00	54.0000
175	9451.00	54.0000
176	9505.00	54.0000
177	9559.00	54.0000
178	9613.00	54.0000
179	9667.00	54.0000
180	9721.00	54.0000
181	9775.00	53.0000
182	9828.00	.000000

from AHS.SND/DEFAULT WITH MINVAR OPTION(3 JUL 1986 at 12:52 am)

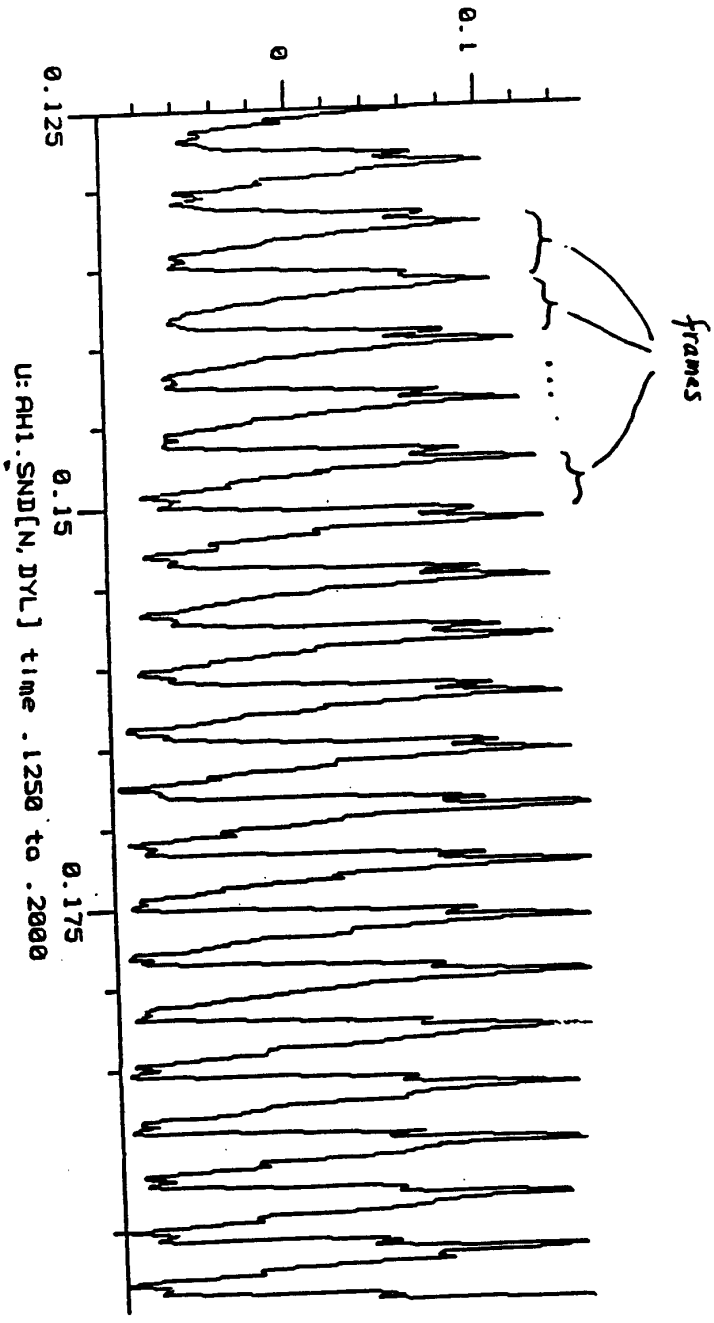
length:	131	
0	141.000	26.0000
1	167.000	50.0000
2	217.000	55.0000
3	272.000	55.0000
4	327.000	50.0000
5	377.000	51.0000
6	428.000	50.0000
7	478.000	50.0000
8	528.000	52.0000
9	580.000	54.0000
10	634.000	56.0000
11	690.000	56.0000
12	746.000	55.0000
13	801.000	54.0000
14	855.000	53.0000
15	908.000	52.0000
16	960.000	53.0000
17	1013.00	53.0000
18	1066.00	54.0000
19	1120.00	53.0000
20	1173.00	53.0000
21	1226.00	52.0000
22	1278.00	52.0000
23	1330.00	52.0000
24	1382.00	51.0002
25	1433.00	53.0006
26	1486.00	54.0000
27	1540.00	51.0000
28	1591.00	49.0000
29	1640.00	51.0000
30	1691.00	52.0000
31	1743.00	53.0000
32	1796.00	50.0000
33	1846.00	51.0000
34	1897.00	48.0000
35	1945.00	47.0000
36	1992.00	47.0000
37	2039.00	47.0000
38	2086.00	47.0000
39	2133.00	47.0000
40	2180.00	47.0000
41	2227.00	47.0000
42	2274.00	47.0000
43	2321.00	48.0000
44	2369.00	47.0000
45	2416.00	48.0000
46	2464.00	47.0000
47	2511.00	49.0000
48	2560.00	48.0000
49	2608.00	48.0000
50	2656.00	49.0000
51	2705.00	48.0000
52	2753.00	50.0000
53	2803.00	49.0000
54	2852.00	51.0000
55	2903.00	50.0000
56	2953.00	51.0000
57	3004.00	51.0000
58	3055.00	51.0000
59	3106.00	53.0000
60	3159.00	53.0000
61	3212.00	55.0000
62	3267.00	52.0000

63	3319.00	54.0000
64	3373.00	54.0000
65	3427.00	56.0000
66	3483.00	58.0000
67	3541.00	60.0000
68	3601.00	63.0000
69	3664.00	61.0000
70	3725.00	64.0000
71	3789.00	66.0000
72	3855.00	69.0000
73	3924.00	72.0000
74	3996.00	78.0000
75	4074.00	81.0000
76	4155.00	84.0000
77	4239.00	78.0000
78	4317.00	87.0000
79	4404.00	88.0000
80	4492.00	91.0000
81	4583.00	90.0000
82	4673.00	93.0000
83	4766.00	96.0000
84	4862.00	95.0000
85	4957.00	98.0000
86	5055.00	101.000
87	5156.00	101.000
88	5257.00	101.000
89	5358.00	95.0000
90	5453.00	99.0000
91	5552.00	103.000
92	5655.00	98.0000
93	5753.00	102.000
94	5855.00	99.0000
95	5954.00	101.000
96	6055.00	102.000
97	6157.00	102.000
98	6259.00	99.0000
99	6358.00	103.000
100	6461.00	102.000
101	6563.00	100.000
102	6663.00	105.000
103	6768.00	110.000
104	6878.00	102.000
105	6980.00	107.000
106	7087.00	111.000
107	7198.00	110.000
108	7308.00	105.000
109	7413.00	104.000
110	7517.00	108.000
111	7625.00	112.000
112	7737.00	108.000
113	7845.00	102.000
114	7947.00	106.000
115	8053.00	101.000
116	8154.00	105.000
117	8259.00	109.000
118	8368.00	108.000
119	8476.00	105.000
120	8581.00	109.000
121	8690.00	108.000
122	8798.00	105.000
123	8903.00	109.000
124	9012.00	114.000
125	9126.00	119.000
126	9245.00	124.000
127	9369.00	123.000

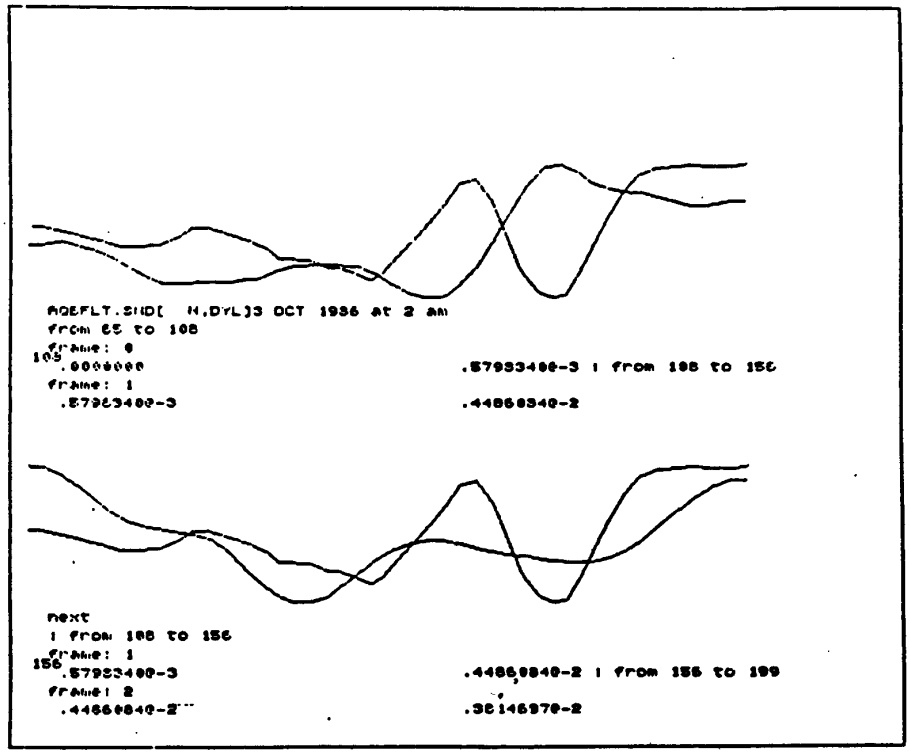
128	9492.00	129.000
129	9621.00	134.000
130	9755.00	.000000

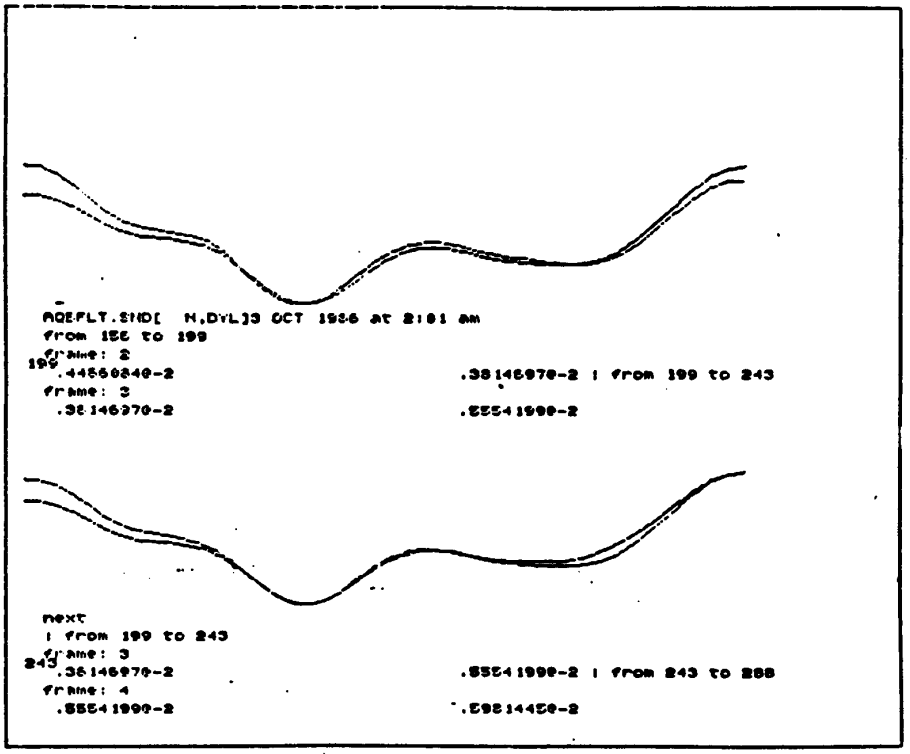
APPENDIX B

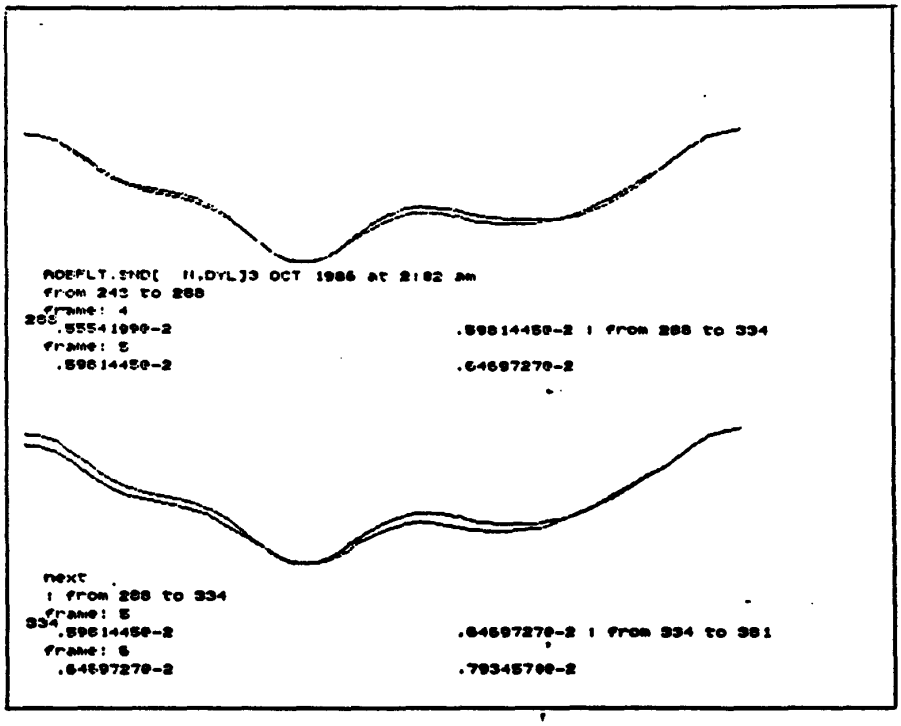
An illustration of the notion of frames as maximally similar neighboring local patterns of variation, using a segment of the /a/ waveform as an example, follows.

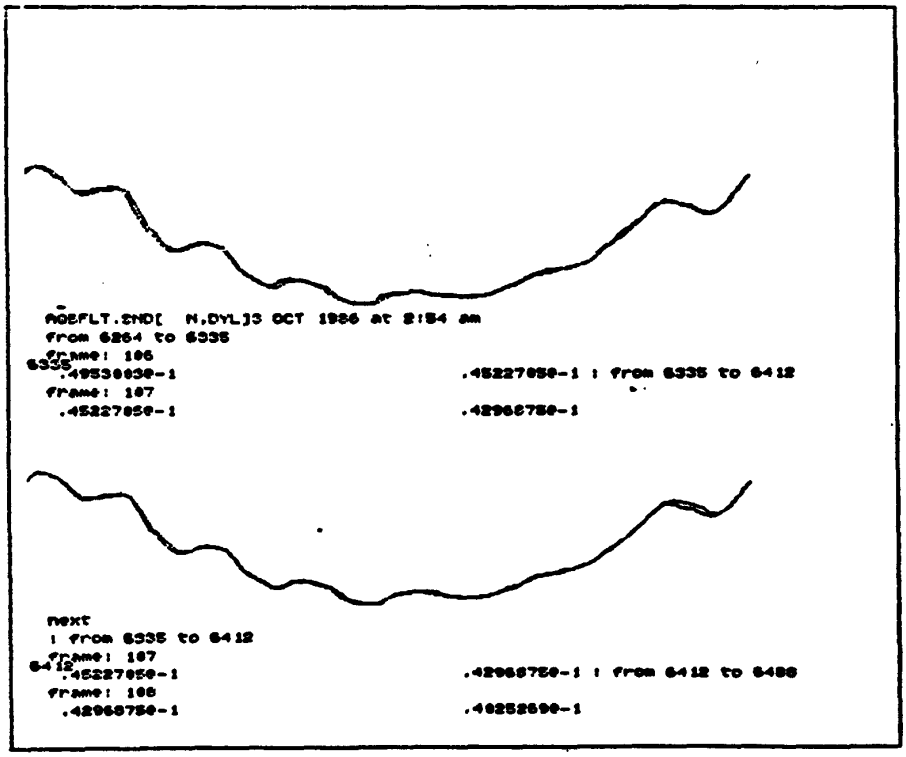


230a


AQEFLT.SND[N,DYL]3 OCT 1986 at 2 am
from 85 to 108
frame: 0
105 .0000000 .57933400-3 | from 108 to 156
frame: 1
.57933400-3 .44868840-2
next
| from 108 to 156
frame: 1
156 .57933400-3 .44868840-2 | from 156 to 199
frame: 2
.44868840-2 .38146970-2


AQEFLT.SND[N,DYL]3 OCT 1986 at 2:01 am
from 156 to 199
frame: 2
199 .44868840-2 .38146970-2 | from 199 to 243
frame: 3
.38146970-2 .8554 1990-2
next
| from 199 to 243
frame: 3
243 .38146970-2 .8554 1990-2 | from 243 to 288
frame: 4
.8554 1990-2 .698144E0-2


ADEFLT.END[11,DYL]3 OCT 1986 at 2:02 AM
from 242 to 288
frame: 4
288 .5554199D-2 .89814458-2 ! from 288 to 334
frame: 5
.89814458-2 .64697278-2
next
! from 288 to 334
frame: 5
334 .89814458-2 .64697278-2 ! from 334 to 381
frame: 6
.64697278-2 .79345708-2

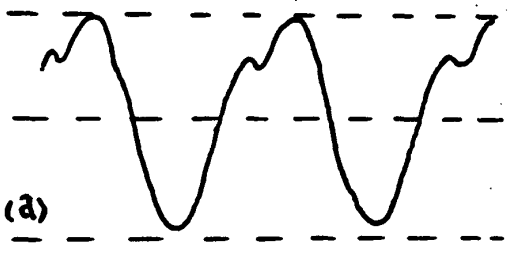

ADEFLT.END[N,DYL]3 OCT 1986 at 2:54 AM
from 6264 to 6335
frame: 186
6335 .49538038-1 .45227858-1 ! from 6335 to 6412
frame: 187
.45227858-1 .42968758-1
next
! from 6335 to 6412
frame: 187
6412 .45227858-1 .42968758-1 ! from 6412 to 6488
frame: 188
.42968758-1 .40252698-1

APPENDIX C

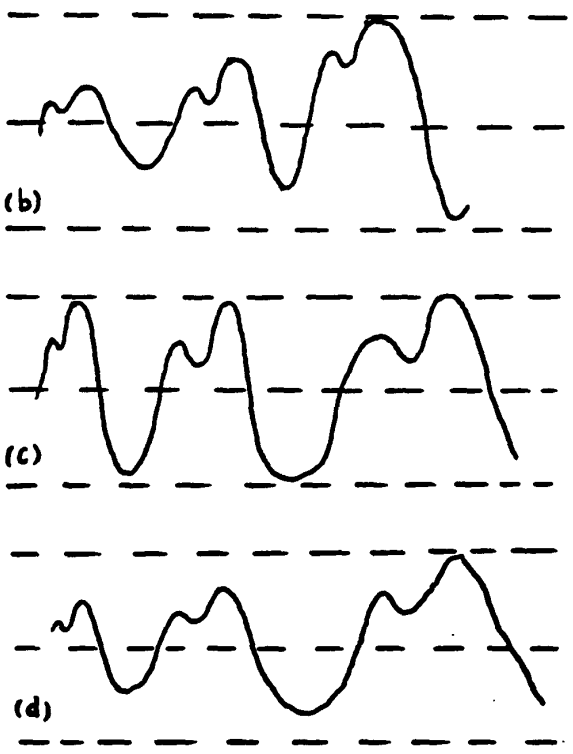
Definition of Temporal Acoustic Features in Timbre and their Perceptual Origin

The idea of acoustic features important to perception of timbre is necessarily a pattern-recognition concept. The following is merely a guide.

A temporal acoustic feature, for short, a temporal feature, is a composition of locally coherent amplitude- or time-scale changes with respect to elemental acoustic events in the acoustic waveform as a function of time. These elemental acoustic events are auditorily identifiable acoustic wave patterns such as peaks, valleys, or maximum slope excursions. Throughout, we assume the existence of an acoustic observer which actively seeks to adapt its own algorithms and contextual data so as to organize the incoming acoustic waveform into maximally structured information.



Scale change refers to change in the grid or coordinate system within which the elemental acoustic events of a waveform is represented. For example, graph (a) at left indicates no amplitude- or time-scale changes.



Graph (b) indicates an amplitude scale change only. Graph (c) indicates time-scale change only while graph (d) indicates both time- and amplitude-scale changes. Coherence is indicated by similar scale changes over the same duration across the spectral or vibrating components of the signal. That is, a scale change must be indicated by monotonic time segments in the same direction of variation among the vibrating components. And the variations must be interpolated from one end to the other over the frequency range.

The extent of locality varies from signal to signal, but it must be at least as long as a couple of periods of the fastest vibrating components of the waveform. Coherent scale–changes give rise to temporal features. Features are scale-changes as a function of time. An elemental temporal feature is a monotonic segment. A feature is a composition of monotonic segments. The most elementary composition is straightforward concatenation.

A concatenation can be a simple extension of a monotonic segment or a merger of two or more. More complex composition involves concatenation of transformed features. Transformation includes scaling, shifting, reflection in time or amplitude, etc. A meaningful composition may mean formation of a recognizable pattern. An amplitude envelope is a superfeature of the amplitude-scale changes in the waveform. A period trajectory is a superfeature of the time-scale changes in the waveform.

If the acoustic signal exhibits coherent amplitude- or time-scale changes, the basilar membrane response will mirror the behavior because of built-in redundant characteristics of the fibers in the basilar membrane. Specifically, "slowly"-varying temporal characteristics are broadly reflected in the responses of the fibers across the

membrane. This redundance is a consequence of the attenuated low-pass frequency response characteristic of all fibers (see figure 2.2.1). A Fourier interpretation of place redundancy (where place refers to location on the basilar membrane) exhibited by the fibers is shown in figure A3.a. Place redundancy is the basis for detection of coherent scale changes and hence the basis for temporal feature extraction. Similarly, the all familiar temporal redundancy in the temporal response patterns which naturally originate in the periodic or quasi-periodic behavior of the signal provides the basis for extraction of place features (i.e., spectral features such as formants and spectral shape). Figure A3.b shows a schematic for feature detection in place and time after Minsky's societal model (see text for references and details).

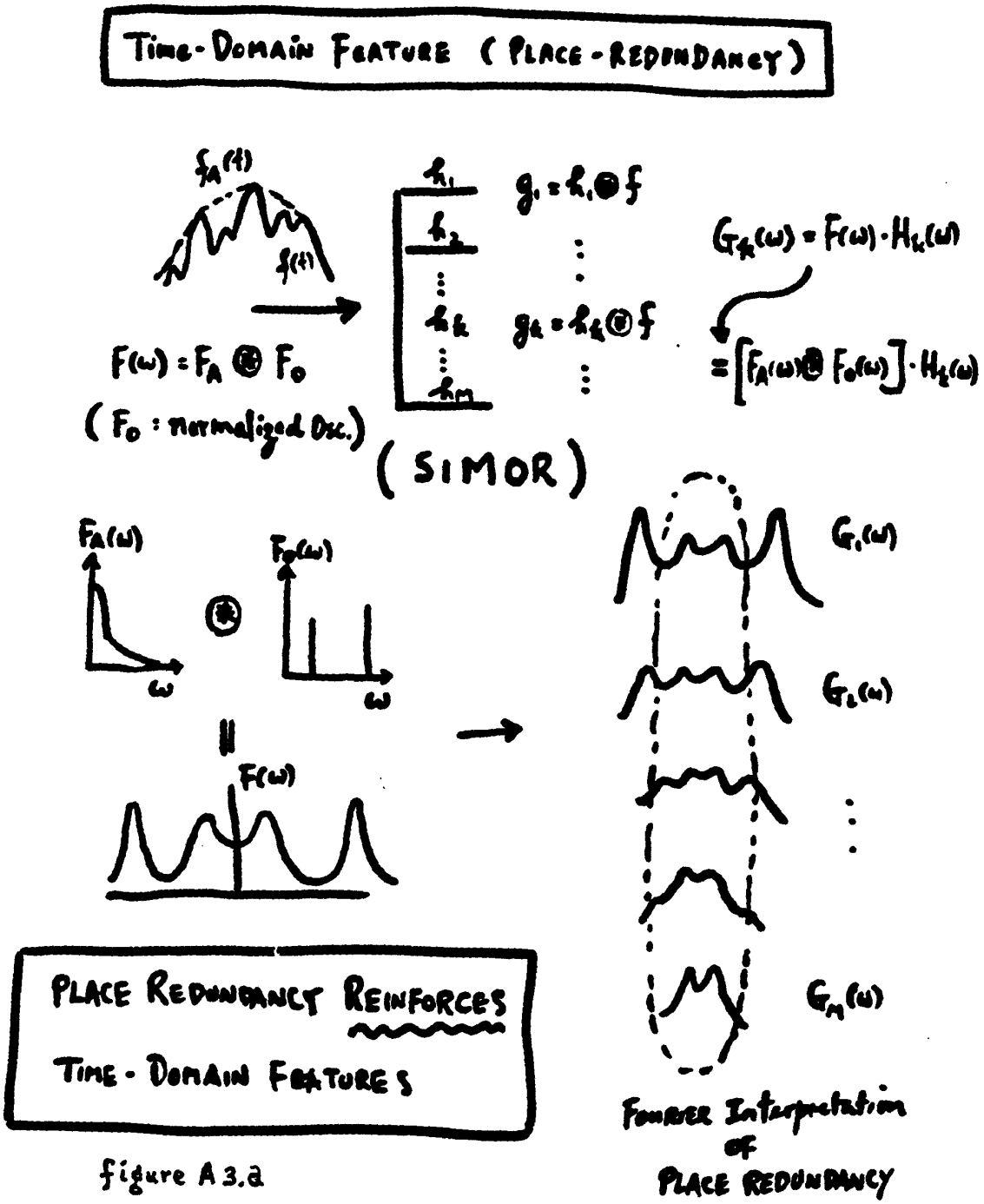

TIME-DOMAIN FEATURE (PLACE-REDUNDANCY)
$f_A(t)$
$f(t)$
$F(\omega) = F_A \circledast F_0$
(F_0 : normalized Osc.)
h_1
h_2
h_k
h_m
$g_1 = h_1 \circledast f$
$g_k = h_k \circledast f$
$G_k(\omega) = F(\omega) \cdot H_k(\omega)$
$= [F_A(\omega) \circledast F_0(\omega)] \cdot H_k(\omega)$
(SIMOR)
$F_A(\omega)$
$F_0(\omega)$
ω
$F(\omega)$
$G_1(\omega)$
$G_2(\omega)$
$G_m(\omega)$
PLACE REDUNDANCY REINFORCES
TIME-DOMAIN FEATURES
FOURIER Interpretation
OF
PLACE REDUNDANCY

figure A3.a

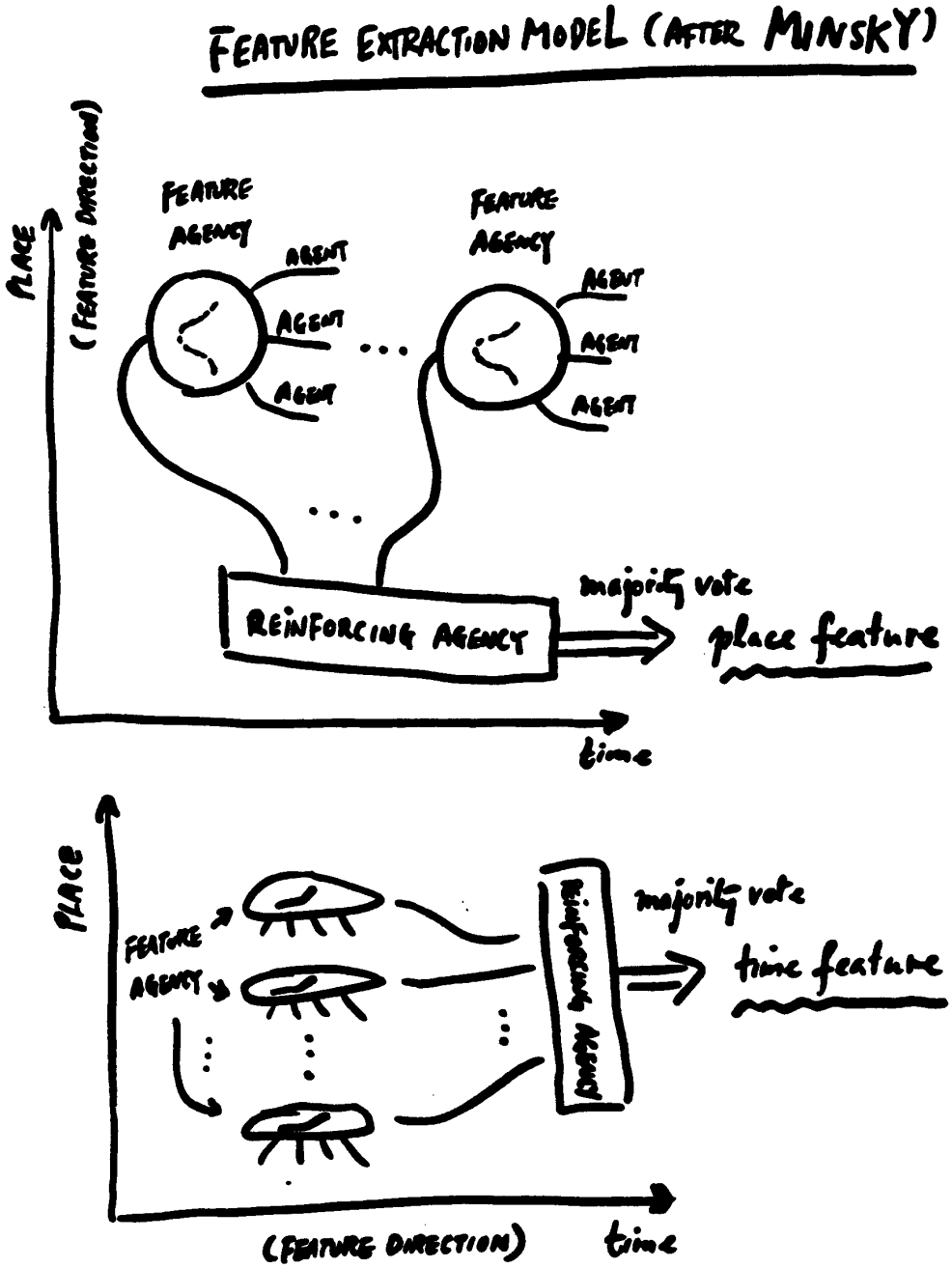

FEATURE EXTRACTION MODEL (AFTER MINSKY)
PLACE
(FEATURE DIRECTION)
FEATURE AGENCY
AGENT
AGENT
AGENT
FEATURE AGENCY
AGENT
AGENT
AGENT
REINFORCING AGENCY
majority vote
place feature
time
PLACE
FEATURE AGENCY
Reinforcing Agency
majority vote
time feature
(FEATURE DIRECTION)
time

Figure A3.b

REFERENCES

Alles, H. G., 1975. "A Hardware Digital Music Synthesizer," *EASCON*: 217A–217E.

Altes, R. A., 1978. "The Fourier-Mellon Transform and Mammalian Hearing," *J. Acoust. Soc. Am.* **63**(1): 174–183.

American Standard Association, 1960. "American Standard Acoustical Terminology."

Backhaus, H., 1932. "On the Importance of Transients in Acoustics," Strawn, J. (translator). Originally appearing in *Zeitschrift für Technische Physik,* **13**(1):31–46.

Baastiaans, M.J., 1980. "Gabor's Expansion of a Signal into Gaussian Elementary Signals," *Proceedings of the IEEE,* **68**(4):538–539.

Beauchamp, J.W., 1975. "Analysis and Synthesis of Cornet Tones Using Non-Linear Interharmonic Relationships," *Journal of the Audio Engineering Society,* **23**:778–795.

Békésy, G., 1960. *Experiments in Hearing,* (translated and edited by E.G. Wever) McGraw-Hill.

Bell, C.G., H. Fujisaki, J.M. Heinz, K.N. Stevens, and A.S. House, 1961. "Reduction of Speech Spectra by Analysis-by-Synthesis Techniques," *J. Acoust. Soc. Am.,* **33**:1725–1736.

Benade, A.H., 1976. *Fundamentals of Musical Acoustics,* London and New York: Oxford U. Press.

Berger, T., 1971. *Rate Distortion Theory,* Prentice-Hall.

Blumstein, S., and K. Stevens, 1979. "Acoustic Invariance in Speech Production: Evidence from Measurements of the Spectral Characteristics of Stop Consonants," *J. Acoust. Soc. Am.,* **66**(4):1001–1017.

Borish, J., 1984. *Electronic Simulation of Auditorium Acoustics,* Ph.D. dissertation, Stanford University.

Bracewell, R., 1978. *The Fourier Transform and its Applications,* McGraw-Hill.

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

Chowning, J.M., 1973. "The Synthesis of Complex Audio Spectra by Means of Frequency Modulation," *Journal of the Audio Engineering Society,* **21**(7):526–534.

Chowning, J.M., 1980. "Computer Synthesis of the Singing Voice," *Sound Generation in Winds, Strings, Computers,* Royal Swedish Academy of Music, Publication no. **29**, Kungl. Musikaliska Akademien, Stockholm.

Clarke, A., 1982. *2010: Odyssey Two,* Random House.

Claasen, T.A.C.M., and W.F.G. Mecklenbräuker, 1980. "The Wigner Distribution—A Tool for Time-Frequency Analysis," *Philips Journal of Research,* **35**, Part I:217–250, Part II:276–300, Part III: 372–389.

Cohen, A., and J. t'Hart, 1962. "Speech Synthesis of Steady-State Segments," *Proc. Stockholm Speech Comm. Seminar,* R.I.T., Stockholm, Sweden.

Crawford, F.S., 1980. *Waves* (Berkeley Physics Course—volume 3), McGraw-Hill.

Dürer, A. See Thompson, D'Arcy, *On Growth and Form.*

Ehresman, D., and D.L. Wessel, 1978. *Perception of Timbral Analogies,* Technical report **13**, IRCAM, Paris. (See also Wessel, D.L., 1979. "Timbre Space as Musical Control Structure," *Computer Music Journal,* **3**(2):45–52.)

Erickson, R., 1975. *Sound Structure in Music,* Berkeley, California: University of California Press.

Fant, G., 1960. *Acoustic Theory of Speech Production,* Gravenhage: Mouton.

Feynman, R.P., R.B. Leighton, and M. Sands, 1963. *The Feynman Lectures on Physics,* Addison-Wesley.

Flanagan, J.L., 1972. *Speech Analysis, Synthesis, and Perception,* second edition, Berlin: Springer-Verlag.

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

Gabor, D., 1947. "Acoustical Quanta and the Theory of Hearing," *Nature,* **159**:591–594.

Gambardella, G., 1971. "A Contribution to the Theory of Short-Time Spectral Analysis with Non-Uniform Bandwidth Filtering," *IEEE Trans. on Circuit Theory,* **CT-18**(4):455–460.

Gordon, J.W., 1984. *Perception of Attack Transients in Musical Tones,* Ph.D. dissertation, Stanford University.

Green, D., 1983. "Profile Analysis," *American Psychologist* **38**:133–142.

Grey, J.M., 1975. *An Exploration of Musical Timbre,* Ph.D. dissertation, Stanford University.

Halle, M., and K.N. Stevens, 1959. "Analysis by Synthesis," *Proc. Sem. Speech Compression and Processing,* W. Wathen-Dunn and L.E. Woods (editors), AFCRC-TR-59-198, **2**, paper D7.

Helmholtz, H., 1877. *Sensations of Tone,* English translation with notes and appendices by E.J.Ellis, 1954. New York: Dover.

Heinz, J.M., and K.N. Stevens, 1961. "On the Properties of Voiceless Fricative Consonants," *J. Acoust. Soc. Am.* **33**:589–596.

Hitt, D. Private communication.

Hoel, P., S. Port, and C. Stone, 1971. *Introduction to Probability Theory,* Houghton Mifflin Company.

Huggins, W.H., 1952. "A Phase Principle for Complex-Frequency Analysis," *J. Acoust. Soc. Am.* **24**:582–589.

Jeans, J., 1937. *Science and Music,* reprinted by Dover.

Kruskal, J.B., 1964. "Multidimensional Scaling by Optimizing Boodness of Fit to a Non-Metric Hypothesis," *Psychometrika* **29**:1–27.

Kajiya, J.T., 1979. *Toward a Mathematical Theory of Perception,* Ph.D. dissertation, University of Utah.

Kashima, K., 1985. "The Bounded-Q Frequency Transform," Dept. of Music Tech. Rep. STAN-M-28.

Keeler, J.S., 1972. "Piecewise-Periodic Analysis of Almost-Periodic Sounds and Musical Transients," *IEEE Transactions on Audio and Electroacoustics, AU-ZO*:338-344.

Knudsen, V.O., 1963. "Architectural Acoustics," *Scientific American,* reprinted in *The Physics of Music* (Readings from *Scientific American*), 1978. San Francisco: Freeman and Company.

Lakatos, I., 1978. *The Methodology of Scientific Research Programmes, Philosophical Papers Volume* **1**, Cambridge U. Press.

Lerdahl, F., 1986. "Timbral Hierarchies," *Contemporary Music Review* **1.3**.

Lo, Y., 1986. "Techniques of Timbral Interpolation," *Proceedings of the ICMC '86.*

Luce, D.A., 1963. *Physical Correlates of Non-Percussive Musical Instrument Tones,* Ph.D. dissertation, M.I.T.

McAdams, S., 1984. *Spectral Fusion, Spectral Parsing, and the Formation of Auditory Images,* Ph.D. dissertation, Stanford University.

McAdams, S., and K. Saariaho, 1985. "Qualities and Functions of Musical Timbre," *Proceedings of ICMC '85.*

Matthews, M.V., 1970. *The Technology of Computer Music,* M.I.T. Press.

Meyer, E., and Buchmann, G., 1931. *Die Klangspektren der Musikinstrumente,* Berlin: Akedemie der Wissenschaften.

Minsky, M., 1981. "Music, Mind, and Meaning," *Computer Music Journal,* 5(3):28–44.

Minsky, M., 1986. *The Society of Mind,* New York: Simon and Schuster.

Moorer, J.A., 1974. "The Optimum Comb Method of Pitch Period Analysis of Continuous Digital Speech," *IEEE Trans. on Acoustics, Speech, and Signal Processing,* **22**:330–338.

Moorer, J.A., 1975. *On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer,* Ph.D. dissertation, Stanford University.

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

Morrill, D., 1977. "Trumpet Algorithms for Computer Composition," *Computer Music Journal,* 1(1):46–52.

Noll, A.M., 1969. "Pitch Determination of Human Speech by Harmonic Product Spectrum, the Harmonic Sum Spectrum, and a Maximum LIkelihood Estimate," *Proceedings of the Symposium on Computer Processing in Communications,* **19**:779–797.

Parsons, T., 1987. *Voice and Speech Processing,* McGraw-Hill.

Petersen, T.L., and S.F. Boll, 1983. "Critical Band Analysis-Synthesis," *IEEE Transactions on Acoustics, Speech, and Signal Procesing, ASSP*-**31**(3):656-663.

Peterson, G.E., and H.L. Barney, 1952. "Control Methods used in a Study of the Vowels," *J. Acoust. Soc. Am.* **24**:175–184.

Pippard, A.B., 1978. *The Physics of Vibration, volume I* (The Simple Classical Vibrator), Cambridge U. Press.

Pippard, A.B., 1985. *Response and Stability,* Cambridge U. Press.

Plomp, R., 1964. "The Ear as a Frequency Analyzer," *J. Acoust. Soc. Am.* **36**(9):1628–1636.

Plomp, R., 1970. "Timbre as a Multi-Dimensional Attribute of Complex Tones," *Frequency Analysis and Periodicity Detection in Hearing,* R. Plomp and G.S. Smoorenburg, editors. Leiden: Sijthoff.

Rabiner, L.R., and R.W. Schafer, 1978. *Digital Processing of Speech Signals,* Prentice-Hall.

Rayleigh, J.W.S., 1877. *The Theory of Sound,* Dover, 1945.

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

Risset, J.-C., 1969. "An Introductory Catalogue of Computer-Synthesized Sounds," Bell Laboratories, Murray Hill, New Jersey.

Risset, J.-C., and M.V. Matthews, 1969. "Analysis of Musical Instrument Tones," *Physics Today* **22**(2):23–40.

Risset, J.-C., and D.L. Wessel, 1982. "Exploration of Timbre by Analysis and Synthesis," D. Deutsch (editor), *Psychology of Music*: 25–58, Academic Press.

Rodet, X., 1979. "Time-Domain Formant-Wave-Function Synthesis," *Computer Music Journal,* 8(3):9–14.

Rudin, W., 1964. *Principles of Mathematical Analysis,* Second Edition, McGraw-Hill.

Rumelhart, D.E., and A.A. Abramson, 1973. "Toward a Theory of Analogical Reasoning," *Cognitive Psych.* **5**:1–28.

Sachs, M.B., and E.D. Young, 1979. "Representation of Steady-State Vowels in the Temporal Aspects of the Discharge Patterns of Populations of Auditory-Nerve Fibers," *J. Acoust. Soc. Am.,* **66**(5):1381–1407.

Schaeffer, P., 1966. *Traité des Objets Musicaux,* Paris: Ed. du Seuil.

Schoenberg, A., 1922. *Harmonic Lehre,* Vienna: Universal. Translated (1978) as *Theory of Harmony* by R.E. Carter, University of California Press.

Schoenberg, A., 1969. *Structural Functions of Harmony,* New York: Norton and Company.

Schouten, J.F., 1940. "The Residue, a New Component in Subjective Sound Analysis," *Proc. Kon. Ned. Akad. v. Wetensch.* **43**:356–365, Amsterdam.

Schouten, J.F., 1968. "The Perception of Timbre," *Report of the Sixth International Congress on Acoustics, Tokyo,* **GP-6-2**:35–44.

Schroeder, M.R., 1984. *Number Theory in Science and Communication,* Berlin: Springer-Verlag.

Schroeder, M.R., and S. Mehrgardt, 1982. "Auditory Masking Phenomena in the Perception of Speech," *The Representation of Speech in the Peripheral Auditory System,* R. Carlson and B. Granström (editors), Elsevier Biomedical Press.

Schloss, W.A., 1985. *On the Automatic Transcription of Percussive Music*, Ph. D. dissertation, Stanford University.

Schubert, E.D., 1980. *Hearing: Its Function and Dysfunction,* Berlin: Springer-Verlag.

Schubert, E.D. Private communication.

Schwede, G., 1983. "An Algorithm and Architecture for Constant-Q Speech Analysis", *ICASSP 83*:1384–1387

Serra, X., 1984. "Simulating Marimba Tones," Poster, *ICMC'84.*

Serra, X., 1986. "A Computer Model for Bar Percussion Instruments," *Proc. ICMC'86*:257-262.

Shannon, C., 1949. "Communication in the Presence of Noise," *Proc. IRE,* **37**:10–21.

Shannon, C., 1959. "Coding Theories for a Discrete Source with a Fidelity Criterion," *IRE Nat. Corr. Rec., pt.* 4:142–163. *Transactions in Information Theory.*

Shannon, R. Private communication.

Shepard, R.N., 1966. "Metric Structures in Ordinal Data," *Journal of Math. Psychology* **3**:287–315.

Shepard, R.N., 1972. "Psychological Representatives of Speech Sounds," *Human Communication,* E.E. Davis and P.B. Denes (editors), McGraw-Hill.

Slawson, W., 1985. *Sound Color,* University of California Press.

Sommerfeld, A., 1947. *Mechanics of Deformable Bodies,* translated by G. Kuerti, 1964, Academic Press.

Strawn, J.M., 1981. "Approximation and Syntactic Analysis of Amplitude and Frequency Functions for Digital Sound Synthesis," *Computer Music Journal* 4(3):3–24.

Strawn, J.M., 1982. "Research on Timbre and Musical Context at CCRMA," *Proc. ICMC '82.*

Stumpf, C., 1926. *Die Sprachlaute,* Springer-Verlag.

Sundberg, J., 1977. "The Acoustics of the Singing Voice," *Scientific American* **236**:82–91.

Teaney, D.T., V.L. Moruzzi, and F.C. Mintzer, 1980. "The Tempered Fourier Transform," *J. Acoust. Soc. Am.* **67**(6):2063–2067.

Tenney, J.C., 1965. "The Physical Correlates of Timbre," *Gravesaner Blätter* **7**(26):106–109.

Thom, R., 1975. *Structural Stability and Morphogenesis,* Benjamin/Cummings.

Thompson, D'Arcy, 1961. *On Growth and Form* (abridged and edited by J.T. Bonner), Cambridge U. Press.

Walker, J., 1977. *The Flying Circus of Physics,* Wiley.

Wedin, L. and G. Goude, 1972. "Dimension Analysis of the Perception of Instrumental Timbre," *Scandinavian Journal of Psychology* **13**:228–240.

Weinreich, G., 1979. "The Coupled Motions of Piano Strings," *Scientific American* **240**(1):118–127.

Wessel, D.L., 1979. "Timbre Space as a Musical Control Structure," *Computer Music Journal* 3(2):45–52.

Winckel, F., 1967. *Music, Sound, and Sensation,* Dover.

Yougrau, W., and S. Mandelstam, 1968. *Variational Principles in Dynamics and Quantum Theory,* Dover.

Youngberg, J.E., 1979. *A Constant Percentage Bandwidth Transform for Acoustic Signal Processing,* Ph.D. dissertation, University of Utah.