Music Perception Systems

A proposal for a Ph.D. dissertation

by

Eric David Scheirer

22 October 1998

Thesis supervisor: ____

Prof. Barry L. Vercoe
Professor of Media Arts and Sciences
Massachusetts Institute of Technology

Thesis reader: ____

Prof. Rosalind W. Picard
Associate Professor of Media Technology
Massachusetts Institute of Technology

Thesis reader: ____

Prof. Perry R. Cook
Assistant Professor of Music
Assistant Professor of Computer Science
Princeton University

Thesis reader: ____

Dr. Malcolm Slaney
Member of the Research Staff
Interval Research Corporation

Music Perception Systems

Proposal for Ph.D. dissertation

Eric D. Scheirer
MIT Media Laboratory

22 October 1998

Abstract

I propose a dissertation that examines the psychoacoustic bases of the perception of music by human listeners. As music is fundamentally an acoustic phenomenon, a theory of music psychology is not a theory of music *perception* unless it can explain how a listener maps from the acoustic signal to an intermediate structural representation which allows cognition to occur. I propose to build computer models which can mimic basic musical perceptions such as parsing music into sections, tapping along with the beat, classifying music by genre and recognizing composers of pieces, upon taking complex acoustic signals as input. The dissertation that results will be both a scientific inquiry into the nature of the music-listening process, and a practical solution to certain difficult problems in computer-based multimedia.

My proposed research is supported by three bodies of background literature: psychoacoustics, music psychology, and musical signal processing. I review selected topics from these literatures pertinent to my research and contextualize my own proposals in light of this previous work. The predominant theoretical stance I propose is one in which the reliance of music-analysis systems on “notes” is set aside in favor of a more directly acoustic and perceptual model of the listening process. In particular, I wish to focus on acoustic signal-processing techniques that can model the perceptual formation of fused musical objects in complex sound scenes, and pattern-classification techniques that model the subsumption of musical objects into overall musical percepts.

The scope of my proposed research is three-fold: the discovery of musical-object-formation heuristics, the analysis of musical-object similarity, and the construction of musical-pattern-recognition systems. I will validate the models constructed by demonstrating performance on complex real-world musical stimuli taken directly from compact disc recordings, and by conducting supporting psychoacoustic experiments if time permits. In this proposal, I present a more detailed description of this scope, an outline of the schedule I plan to follow, and the resources required to undertake this project.

Thesis supervisor: Prof. Barry L. Vercoe

Contents

I. Introduction

This proposal describes a dissertation which examines the psychoacoustic bases for the perception of music by human listeners. The research I propose will explore the relationship between psychoacoustics and music perception from two perspectives: as a scientific inquiry into the nature of the music-listening process, and as a practical solution to difficult problems in computer-based multimedia such as automatic content indexing and the construction of systems for musical interaction.

Music is an acoustic medium—or as Bregman (1990, p. 455) writes, "Music is built out of sound." Every aspect of musical structure which allows compositions to differ from one another is conveyed in the sound-pressure waves which make up a musical performance. However, music-*listening* is a perceptual process. The manner in which listeners map from the acoustic realizations impinging upon their eardrums to a cognitive model of musical structure is subtly and crucially dependent upon fundamental principles of auditory organization. It is my goal to explain the abilities of listeners to dynamically organize musical sound signals using new models based in the study of psychoacoustics and auditory scene analysis, and to demonstrate how such a theory can be used to construct powerful and compelling new-media systems that use musical input data in novel ways.

In my proposal, I suggest a direction of research which differs from the bulk of existing research in several ways. It differs from the mainstream music-psychology literature by putting the emphasis on psychoacoustic explanation, rather than cognitive-structural explanation, and by primarily considering the music-perception abilities of non-musicians. It differs from the psychoacoustic literature by focusing on real-world musical sounds rather than specially-constructed test cases. And it differs from the music-signal-processing and multimedia-systems literature by treating models of perception as the best solution to most musical problems in these domains.

A first approximation to my scope can be expressed as follows. Consider a musically-unskilled listener turning on a radio and hearing five seconds of sound from the middle of a previously unheard piece of music. I want to build a computer system that can make the same judgments about this piece of music as the human listener can.

While such scenarios are not typically considered in music-psychology studies or experiments, it is clear that the listener can say many interesting things about the music that are beyond our current ability to model. The listener immediately orients himself or herself in the performance by categorizing the type of music and forming a primitive sense of structure. The listener will perhaps be able to identify the genre of the music, discuss what other pieces or kinds of music it bears similarity to, have an emotional reaction to the music, identify the instruments in the music, sing back a certain voice in the music, verbalize a "sketch" of the music, identify the composer or performer, clap along with rhythms in the music, classify the music as "simple" or "complicated," identify social scenarios in which the music is appropriate or inappropriate, make claims about the emotive intent of the composer or performer, think of other people who might like the music, and so on.

These capabilities fall along a continuum of mental abilities, from the very simple (tapping along) to highly conceptual, cognitive, and complex (identifying appropriate social scenarios). However, it is clear—and this is a distinction rarely made in music psychology—that very little musical input, long-term stimulus structure, or musical skill is needed to use these abilities. I wish to examine those aspects of the music-listening process responsible for organizing the five-second musical excerpt into a mental structure which allows for other cognitive abilities to operate. I will argue in this proposal, and pursue research attempting to demonstrate, that such judgments involve only everyday hearing ability, not specialized music-hearing capability, and that this "surface structure" of music is ripe for modeling with the right computational tools.

The title of my proposed dissertation touches on both aspects of this approach. First, I want to build *computer systems* that can perceive music, not just to articulate a perceptual theory of music listening. Implementation studies are crucial in the modern study of perceptual models. But to accomplish this, first I want to understand more about what makes it possible to *perceive* music, by study (theoretical, if not experimental) of humans, the only music perception systems which currently exist.

This document is structured as follows. Section II will discuss background for my proposed research, highlighting other studies that I consider most relevant to my goals within the psychoacoustic, music-psychological, and music-pattern-recognition literature. Section III outlines my view of the approach to take regarding these problems, describing the model of music perception I feel to be the most promising. Section IV provides a focused scope for my dissertation, explaining the aspects I will pursue first and those with which I will continue if time permits. Section V details practical considerations of my Ph.D. direction, including schedule and materials required. A brief conclusion and an extensive reference list conclude the proposal.

II. Background

Three areas of previous research bear the most direct relationship to my proposal. The literature on *psychoacoustics* describes the relationship between acoustic sound signals, the physiology of the hearing system, and the perception of sound. The literature on *music psychology* explores the processes which govern how composers and performers turn intentions into music, how listeners turn musical data into models of musical structure, and how cognitive music structures give rise to affective response. The literature on *musical signal processing* reports previous attempts to build computer systems that can process music, extract features from acoustic signals, and use them in practical applications.

Each of these three areas will be discussed in depth in subsequent subsections. I will not attempt to discuss all research in these disciplines, but only those which I see as most current and most directly connected to the main direction of my proposed research. At the end of Section II, I will discuss projects that cross this (somewhat arbitrary) division of boundaries. The projects in that final subsection are the ones closest to my own approach.

A. Psychoacoustics

Psychoacoustics as a field of inquiry has been in existence for many years; some of the earliest studies recognizable as psychological science in the 19th century were concerned with the perception of loudness and pitch of sound. Psychoacoustical thinking dates all the way back to the ancient Greeks; Pythagoras is credited with recognizing that strings whose lengths are related as the ratio of small integers sound good when plucked together.

Modern psychoacoustics, since the work of Stevens, Wegel, Fletcher and others in the early 20th century (Fletcher's research is reviewed in Allen 1996), has evolved sophisticated understanding of the psychophysics of audition. Robust and well-tested models have been developed, especially of single perceptual features (such as pitch and loudness) of simple stimuli, and the way in which one simple sound *masks* (hides) another depending on the time-frequency relationship between the two sounds. There is also a large body of research treating the perception of "roughness" in sounds, which relates to the study of musical consonance; I will discuss these results in Section II.B in the context of other studies on musical harmony. More recently, a research focus has developed to study the perceptual grouping and segregation of sounds under the broad heading of "auditory scene analysis" (Bregman 1990). I provide a brief review of theories of pitch and of auditory scene analysis in this subsection; I do not view loudness and masking models as directly relevant to my proposal.

1. Pitch theory

"Pitch" is the perceptual correlate of the frequency of a simple tone. It is the feature of a sound by which listeners can arrange sounds on a scale from "lowest" to "highest." The early days of pitch research dealt primarily with understanding the exact capabilities and psychophysical discrimination accuracy for pitch; more recently, research has focused on the construction of computational (or at least functional) models that mimic the human ability to determine pitch from acoustic signals.

In the process of building and testing such models, researchers can draw on the wealth of existing knowledge about the types of signals that generate a pitch sensation. Licklider (1951a) provided an early review of this research; Zwicker and Fastl (1990) and Hartmann (1996) provided more recent reviews of the data. The former gave a list of no less than eleven different types of pitched sounds—even more have been discovered since. Thus, the primary task in building functional models of pitch processing is to *explain* this data. There are two types of pitch models: *place* models of pitch, and *temporal* models of pitch.

In a place model of pitch, pitch is explained as the result of pattern-recognition analysis of a sound spectrum. The cochlea acts as a spectrum analyzer, and passes a set of "spectral peaks" to a central processor, which determines the pitch of the sound from the relationships of the peak positions.

A fine example of this traditional view of pitch perception was the Goldstein "optimum processor" model (Goldstein 1973). In this model, the instantaneous spectral peaks of a signal were extracted, and a maximum-likelihood processor (Therrien 1989) was used to decide which pitch best explains the spectrum under analysis. The model could explain a wide range of phenomena, including the pitch of sounds with missing fundamental, the percept of dichotic pitch (where only some harmonics are presented to each ear), and the "musical intelligibility" (capacity for describing musical intervals) of various pitch percepts. It was also presented in a rigorous mathematical formalism, which was viewed as more of a positive feature at the time than today.

Similar mechanisms have been proposed by Wightman (1973), Terhardt (1974), and Hermes (1988), among others. There are several disadvantages of such models. First, they require more spectral resolution in the front end of analysis than is known to exist in the ear. Signals with closely spaced spectral peaks can still give rise to a pitch sensation even though they are not resolved separately by the cochlea. Second, they do not easily explain certain pitch phenomena, such as iterated noise signals, which are spectrally flat. Additionally, they make scant predictions about the nature of the "central processor" that actually performs the analysis, and thus it is very difficult to evaluate this part of these models.

In a temporal model of pitch, pitch is explained as the result of temporal processing and periodicity detection on each cochlear channel. The cochlear acts as a spectrum analyzer, but rather than extracting the peaks from this representation, the band-passed signals output from the cochlea are inspected in the time domain. Pitched signals have periodic fluctuations in the envelopes of the subband signals, and a variety of methods have been proposed for measuring this periodicity.

The original development of this technique was due to Licklider (1951b), who presented a model based on a network of delay lines and coincidence detectors oriented in a two-dimensional representation. The first dimension corresponded to the spectral height of the signal, as analyzed by the cochlea, and the second to the *autocorrelation delay*, over the range of periods that evoke a pitch sensation. This construction calculated a running autocorrelation function in each channel; the peak of the function within a channel indicated the primary pitch to which that channel responded. Presumably, the information from multiple channels would then be integrated to give rise to a single sensation of pitch, but Licklider did not provide explicit details or predictions. Licklider termed this the *duplex* model of pitch.

Since Licklider’s formulation, this technique has been “rediscovered” several times, first by van Noorden (1983), who cast it in terms of the calculation of histograms of neural interspike intervals in the cochlear nerve. In the last decade, the model was reintroduced by Slaney and Lyon (1990), Meddis and Hewitt (1991b; 1991a), and others; it has since come to be called the *autocorrelogram* method of pitch analysis and is today the preferred model. The autocorrelogram is the 3-D volumetric function mapping a cochlear channel, temporal time delay, and time to the amount of periodic energy in that band at that delay period and time.

Meddis and Hewitt (1991a) specifically proposed that the cross-band integration of periodicity took the form of a simple summation across channels. They presented a number of analytic and experimental results showing that this model can quantitatively explain a great deal of the psychoacoustic data regarding pitch perception.

Patterson and his collaborators have spent several years recently developing a so-called Auditory Image Model that can achieve the same results with a somewhat different processing structure (Patterson, Allerhand and Giguere 1995). In this model, after cochlear filtering and inner-hair-cell transduction, a set of threshold detectors strobe and trigger integrators in each channel. Patterson has shown that such triggering rules, although simpler to compute than autocorrelation, can still be highly effective at “stabilizing” the acoustic data and explaining psychophysical pitch experiments. He claims as a major advantage that the model is asymmetric with regard to time, and has presented some experimental evidence (Irino and Patterson 1996) that seems to show that humans may indeed be sensitive to such temporal asymmetries.

Slaney (1997) presented a review of psychoacoustic and neurophysiological evidence for and against various types of correlogram processing: the modulation spectrogram (in which a short-time Fourier transform is calculated within each cochlear channel to analyze periodicity), the “true” autocorrelogram, and Patterson’s model. He concluded that there was little direct neurophysiological evidence for any of these methods, but that the explanatory power of the models with respect to the available psychoacoustic evidence (especially that taken from the literature on pitch phenomena) was such that it seemed likely that some temporal integration method similar to these was indeed used in the brain. In contrasting the models, he concluded that autocorrelation is less neurophysiologically plausible than a method like Patterson’s, and that the modulation spectrogram accounted less well for the psychoacoustic data than the other two.

The autocorrelogram and related representations are important to my research because their use is not restricted to the analysis of pitch in acoustic signals. As we have reasonable, albeit circumstantial, evidence that an important part of early auditory processing uses a periodicity-detection representation, examining such a framework to see “what else it’s useful for” in other auditory models is appropriate. For example, several recent computational auditory scene analysis systems (see below) have used the autocorrelogram as the front end.

Recently, de Cheveigné (1993; 1998) proposed a “cancellation” model of pitch perception, in which the multiplication operators used in calculating the autocorrelogram are replaced with half-wave rectified subtractions. He showed that this model is nearly equivalent to the autocorrelogram, except that it preserves certain temporal asymmetries; he also speculated that the cancellation model might be more useful for explaining multiple-pitch stimuli.

An important thread of research with which I am less familiar is the study of neural mechanisms for auditory processing. Recent reports suggest evidence of the information needed to extract pitch (Cariani 1996; Cariani and Delgutte 1996), temporal envelope (Delgutte *et al.* 1997), and other musically-relevant information in present in the neural anatomy (although this is different that demonstrating that this information is actually *used* the way we think it is). However, the neurological study of stimuli with any complexity is still in its infancy, and I do not view connections to neurophysiology as an important goal of my research.

2. Auditory scene analysis

Since the 1970s, the work of Bregman, his collaborators, and others has resulted in a new body of psychoacoustic knowledge collectively known as *auditory scene analysis* (ASA). The goal of this field is to understand the way the auditory system and brain process complex sound environments, where multiple sources which change independently over time are present. Two subfields are dominant: auditory *grouping* theory, which attempts to explain how multiple simultaneous sounds are partitioned to form multiple "auditory images"; and auditory *streaming* theory, which attempts to explain how multiple sequential sounds are associated over time into individual cohering entities.

The pioneering work of Bregman in this field was summarized in his classic text which named the field (Bregman 1990). He and his colleagues and students conducted dozens of experiments in which different *grouping cues* were put into conflict. In the generally-accepted Bregman model, the sound organization system groups primitive components of sound into sources and sources into streams. These grouping processes utilize rules such as "good continuation," "common fate," and "old + new" to decide what components belong together in time and frequency. In many ways, Bregman's articulation of perceptual grouping cues can be seen as a formalization and principled evaluation of quasi-scientific ideas proposed by the school of Gestalt philosophy/psychology in the early part of the century.

Bregman and his colleagues typically proposed grouping models in which the mid-level representation of sound—that is, the theorized representation lying between the signal and the final percept—was a collection of sinusoidal tones. This is a somewhat problematic view for constructing a proper functional model. On the front end, it seems that the cochlea does not provide sufficient time-frequency resolution to produce such a representation; and in pattern recognition, it has proven difficult to solve the *correspondence problem*, in which we must discover which sinusoids from a complex acoustic source "belong" together. The most current research uses the autocorrelogram and similar representations to examine the source-grouping problem, although the correspondence problem does not vanish in this domain.

The approaches reported over the last 15 years in the ASA literature have been strongly functionalist and computational in nature. Brown and Cooke (1994) termed the discipline of constructing computer models to perform auditory source segregation *computational* auditory scene analysis (CASA). In the summary to follow, I combine results from perceptual studies and computational modeling, since the two projects support each other in a complex way.

In his dissertation and follow-on work, McAdams (1983; 1984; 1989) showed the important role of temporal modulation in separating sounds. He showed through psychoacoustic experiment that frequency modulation applied to one source from a mixture of synthetic vowels makes it "pop out" perceptually. Also, complex tones in which all partials are modulated coherently are perceived as more "fused" (see Section II.B) than tones in which partials are modulated incoherently. McAdams subsumes these results into a general interpretation and model of the formation of "auditory images."

Unlike most other ASA researchers, McAdams had explicitly musical motivations for his research. He presents his interests not only in terms of the scientific problems, but also in terms of providing advice to composers. For example, he identifies the following as a primary question:

> What cues would a composer or performer need to be aware of to effect the grouping of many physical objects into a single musical image, or, in the case of music synthesis by computer, to effect the parsing of a single musical image into many? (McAdams 1984, p. 3)

Weintraub (1985) used a dynamic programming framework around Licklider's autocorrelation model to separate the voices of two speakers whose voices interfere in a single recording. His goal originated in speech recognition and enhancement—he wanted to "clean up" speech signals to achieve better speech recognition and performance. The goal of enabling more robust speech recognition and speaker identification continues to be the strongest motivation for conducting research into CASA systems.

Summerfield, Lea, and Marshall (1990) presented a convolution-based strategy for separating multiple static vowels in the correlogram. By convolving a two-dimensional wavelet kernel possessing the approximate shape of the "spine and arches" of the pitch structure in a correlogram frame with an image representing the whole frame, they showed that multiple pitches with differing F_0 could be recognized. The stimuli used were simple synthesized vowels with F_0 not harmonically related.

Summerfield *et al.* drew an explicit contrast between "conjoint" grouping strategies, in which energy from each correlation channel is split up and assigned to several sources, and "disjoint" strategies, in which the channels themselves are partitioned between channels. Their method was a disjoint method; they do not provide psychoacoustic evidence for this decision, but base it on the grounds of physical acoustics ("when sounds with peaked spectra are mixed, energy from one or other source generally dominates each channel.") Bregman (1990) argued for a disjoint model, which he called the principle of *exclusive allocation.*

One of the key directions for my research is the finding of Duda *et al.* (1990) that *coherent motion* can be seen and easily used to visually segregate sources in an animation of an autocorrelogram moving over time. The excellent examples on Slaney and Lyon's video "Hearing Demo Reel" (Slaney and Lyon 1991) make this point very clear—anytime we perceive sounds as segregated in the ear, the moving correlogram shows them separated with coherent motion. There has been relatively little work on operationalizing this principle in a computer system, however; this is one of the starting points of my approach as discussed in Section III.5.

Mellinger's (1991) thesis (discussed in more detail later) contains a brief exploration of motion-based separation in the correlogram, but the techniques he developed for autocorrelogram analysis were never integrated into the main thrust of his system. McAdams also used this idea of coherent motion to drive his fusion research, but in the sinusoidal domain, not the correlogram domain. One might consider this a "place model" of grouping-from-motion, to contrast with the "timing model" suggested by Duda *et al.*

Over the last decade, a great number of psychophysical results for the so-called "double vowel" paradigm, in which two synthetic vowels are superposed and played to a listener, have accumulated. The listeners are required to report both vowels, properties of the vowels and the manner of the mixing are modified, and the effect on the accuracy with which the vowels are reported is tested. Certain of the pitch-processing algorithms above have been extended to allow the separation of simultaneous sounds as well as their analysis of their pitch (Meddis and Hewitt 1992; de Cheveigné 1993).

Once sounds are separated from mixtures with CASA systems, it is useful to be able to resynthesize the separands in order to compare them to the perceived sources in the mixture. This is easy for a sinusoidal representation, where additive synthesis regenerates the components after they have been isolated; however, it is more difficult for the autocorrelogram. Slaney and colleagues (1994) presented a method for accomplishing correlogram inversion, and reported that the sound quality of resynthesis is very good for the simple analysis-resynthesis loop, where no separation or modification occurred. It is difficult to modify and "play back" correlograms, because an arbitrary three-dimensional volumetric function is not necessarily the correlogram of any actual sound. Nonlinear internal-consistency properties of the correlogram must be used to create correct functions in order to allow resynthesis, and there is little research on this topic.

A system allowing guided (or constrained) analysis-modification-resynthesis from correlograms would be extremely valuable for examining the perceptual segregation and motion properties of the representation. For example, we could "by hand" eliminate certain channels or other structures, and listen to the perceptual effect on the reconstructed sound. On the other hand, the auditory system itself does not perform resynthesis; attempts to separate and resynthesize sound are not precisely research into perceptual modeling. A. Wang (1994) presents a discussion of these two tasks in the introduction to his thesis on signal-processing and source separation.

D. P. W. Ellis, who graduated in 1996 from the Media Laboratory, has been a leading proponent of the *prediction-driven* model of computational auditory scene analysis (Ellis 1996a; Ellis 1996b). In this framework (abbreviated PDCASA), *expectations* are developed using source models which "guess" what is about to happen in the signal. The bottom-up signal analysis is then used to confirm, deny, or augment the current set of expectations about the sound scene. The top-down expectations and bottom-up signal analysis continuously interact to give rise to a percept of sound. Such models are required to explain a number of known psychoacoustic effects, such as illusory continuity, phonemic restoration (Warren 1970; Warren, Obusek and Ackroff 1972), and other auditory illusions.

Ellis' dissertation (1996a) described a system that could analyze sound and segregate perceptual components from noisy sound mixtures such as a "city-street ambience." This is a difficult problem, since cues based on pitch or tracking of sinusoids are not always present. His system was the first to demonstrate an ability to be fooled by illusory-continuity stimuli. He conducted a psychoacoustic investigation to determine what humans hear in such noisy scenes, and concluded that his system showed equivalent performance.

Nawab, Lesser, and their collaborators developed a robust testbed for the construction of PDCASA systems, called *IPUS* for *Integrated Processing and Analysis of Sound* (Klassner, Lesser and Nawab 1998; Nawab *et al.* 1998). This system allows prototypes of rule structures and signal-processing methods to be quickly integrated and tested on sound databases. Klassner (1996) used this system in his dissertation to demonstrate robust segregation of known, fairly constrained environmental sounds (clocks, buzzers, doors, hair-dryers, and so forth). This system was not a perceptual model; Klassner used engineering techniques and constraints to perform segregation.

A second type of computational auditory scene analysis system of recent interest is the model of *oscillatory segregation*. In this model, input units (artificial neurons) respond with periodic behavior to an input signal with certain properties, such as possessing strong response at a certain point in the autocorrelogram. The input nodes influence, through a connectionist network, a highly cross-coupled set of *internal resonators.* The synchronization structures that result (in the internal resonators) can be highly predictive of certain perceptual source-grouping behavior. Models of this sort have been presented by D. Wang (1996) and by Brown and D. Wang (1997). A recent model by McCabe and Denham (1997) used a similar structure and also included elements meant to simulate attentional focus.

Recent work in psychological auditory scene analysis has led to a reexamination of the "classical" view of perceptual properties of sounds—such as loudness, pitch, and timbre—in an attempt to understand how such sound qualities are influenced by the perceptual grouping context. For example, McAdams *et al.* (1998) have elicited compelling evidence that loudness is not a property of an overall sound, but rather of each sound object in a perceived scene, and further, that fundamental auditory organization (the division of the scene into streams or objects) must precede the determination of loudness.

B. Music Psychology

There is a vast literature on music psychology, and many volumes have been written to survey it. In this section, I provide a highly selective overview of the work with which I am most familiar in five areas: pitch, tonality, and melody; tonal consonance and fusion; music and emotion; the perception of musical structure; and musical epistemology. A note on the use of musically experienced and inexperienced listeners in music-perception experiments concludes the section.

My proposed research will not fit directly into the mainstream of these sorts of endeavors; in fact, my own views are generally contrary to those most often articulated by music psychologists. However, it is my hope that my dissertation will be relevant to questions posed by psychologists as well as those posed by psychoacousticians, and so I am attempting to gain familiarity in the relevant subdisciplines of music psychology.

Many of these areas have more complete review articles in the literature. An extensive review of the research in the first two sections has been presented by Krumhansl (1991b). McAdams (1987) presented an excellent critical overview of connections between music psychology, music theory, and psychoacoustics. A more selective overview and critical essay regarding the relationship between music theory and music psychology has been presented by Rosner (1988).

1. Pitch, melody, and tonality

One of the central areas of research into the psychology of music during the last 25 years has been an exploration of the use of pitch in music. This includes the way multiple notes group horizontally into melodies, vertically into chords, and in both directions into larger-scale structures such as “harmonies” and “keys.” These latter concepts are somewhat theoretically nebulous (Thomson 1993), but crucial in the theory of Western music; the preponderance of formal music theory deals with the formation of notes into melodies and harmonic structures, and harmonic structures into areas of “key” or “tonality.”

Early work in this field explored hypotheses regarding pitch relationships drawn directly from music theory (discussed in Rosner 1988). This includes the now-classic presentation of the pitch helix Shepard (1964). The helical model of pitch provides a geometric representation of the two-dimensional nature of pitch similarity: tones are similar in pitch to other tones that are close in frequency (C is similar to C#)—the direction around the helix—and also similar to other tones whose frequency stands in an octave relationship (A440 is similar to A880)—the vertical direction on the helix. This paper by Shepard also developed the “Shepard tone” or “octave-complex” stimulus, which clearly shows that pitch chroma (the position of tones within an octave, which Western music denotes with letters A-G) is to some degree perceptually separable from pitch height. Various other geometric models of pitch similarity have been reviewed by Shepard (1982).

C. Krumhansl has done a great deal of work extending these ideas into a rich and robust framework. She developed, with Shepard, the crucial *probe-tone* technique for investigating the influence of what she terms tonal *context* (a music theorist might term this “local key sense”) on pitch relationships (Krumhansl 1979). In this method, a short context-determining stimulus (for example, a chord or scale) is played to a subject, and then a probe tone taken from the entire chromatic pitch set is played. The listener is asked to judge how well the probe tone “completes” or “continues” the stimulus; by testing each of the chromatic pitches in turn, a *hierarchy* of pitch-context relatedness can be measured.

The tonal hierarchy characterizing the relatedness of pitch and harmonic context turns out to be a very stable percept. Under a variety of context stimuli, including chords, scales, triads, harmonic sequences, and even individual notes, very similar response functions for the tonal hierarchy can be measured. Krumhansl modeled the dynamic motion of listeners through “key space” as the ebb and flow of predominating tonal hierarchies (Krumhansl and Kessler 1982). An excellent book by Krumhansl (1990) summarized her work on this topic. It included, interestingly, an algorithm for determining the dynamic key progression of a piece of music by correlating the pitch-chroma histogram with the tonal hierarchy of each of the 24 major and minor keys.

More recent work in these directions explores the importance of the rhythmic relations among notes in the formation of a sense of key (Schmuckler and Boltz 1994; Bigand, Parncutt and Lerdahl 1996), the relationships between melody, harmony, and key (Povel and van Egmond 1993; Thompson 1993; Holleran, Jones and Butler 1995) and the development of these processes in infants, a literature with which I am not familiar. Many researchers have also focused on the key-finding algorithm, attacking it as a practical problem somewhat distinct from its origins in perceptual science (Ng, Boyle and Cooper 1996; Vos and Van Geenen 1996; Temperley 1997).

Much of the work on tonality took as its goal an understanding of melodic continuation; the general conclusion is that melodies need to have a certain harmonic coherence if they are to be judged as pleasing. That is, there is an important interaction between the sequential nature of melodic continuation and the synchronic nature of harmonic sense. A different approach to melody perception has been presented by Narmour (1990), who developed what he terms an "implication-realization" model of melodic understanding.

Narmour's model drew heavily from Gestalt theories of psychology and from the work of L. B. Meyer (discussed below); he proposed several rules that describe what listeners prefer to hear in melodies, based on principles of good continuation, closure, and return-to-origin. He claimed that these rules represent *universal* guidelines for note-to-note transitions in melodic motion and that as such, they apply to atonal and non-Western musics as well as to Western tonal music. From the rules, he developed an extensive symbolic-analysis model of melody and proposed several experiments to analyze its predictions. An extension to this model, described in a second volume with which I am less familiar, presented a technique for the analysis of more complex melodic structures such as sequences (where the same figure is repeated in several registers, forming a hierarchically superior melody as in Schenkerian analysis).

Some psychological experiments to evaluate Narmour's model have recently been conducted along the lines he suggested (Schellenberg 1996; Krumhansl 1997). These experiments have found general support for his principles, but also have found that his model can be simplified significantly without affecting its ability to explain perceptual data. Thus, it seems at the moment that Narmour's model of melodic continuation is somewhat more elaborate than it needs to be.

The heavily structural nature of all of the work discussed in this section—building as it does from geometric and rule-based models—has made it attractive to computer scientists seeking to build "good-old-fashioned AI" models of music perception. For example, Longuet-Higgins (1994) developed several models of musical melody and rhythm around phrase-structure grammars (often used in linguistics and computer-language theory). Steedman (1994) discussed similar ideas, and also the use of various distance metrics in "tonality space" to describe the relationship between pitch, key, chord, and tonality.

Such models are typically very good at explaining a small subset of well-chosen examples, but make few testable predictions and are rarely considered in light of the full weight of experimental data available. They also typically make very strong claims about what constitutes musical competence[1]. Such properties are typical of the tradition in generative linguistics many of these theorists draw from. This is less of a concern in linguistics, since one of the fundamental tenets of modern linguistic theory is that judgments of grammaticality are shared by all native speakers of a language. An analogous argument does not hold for music listening.

[1] "A sure sign of musical competence is the ability to transcribe into stave notation a tune one hears played on the piano" (Longuet-Higgins 1994, p. 103) – a criterion which surely eliminates all but a vanishingly small proportion of the listening population from his consideration, not to mention the large number of musical cultures that have no canonical written form.

2. Perception of chords: tonal consonance and tonal fusion

There are relatively few reports in the literature examining the perception of vertical music concepts such as chord quality. This is somewhat surprising; given the music-theoretic importance of the roots of chords (discussed by Thomson 1993), and that of the invariance of chords under inversion, it would be revealing to have experimental data confirming or disputing these theories. However, the related concepts of *tonal consonance* and *tonal fusion* have received some attention.

The term *tonal consonance* is used to refer to the sense of "smoothness" or "pleasantness" that results when two sounds with certain properties are played together. Typically, this property results when the pitches of the sounds are in a simple integer ratio relationship such as 2:1 (an octave) or 3:2 (a fifth). The term *tonal fusion* refers to the sense of two sounds "merging" into a single sound in a musical context. This concept is very important in the theory of orchestration in music. The German word introduced by Stumpf and used in contexts where I say "tonal fusion" is *Verschmelzung*, literally "melting," a charming way of denoting the intermingled character of fused sounds. The early work by Stumpf on *Verschmelzung* is reviewed by Schneider (1997) in light of modern psychoacoustic theory.

Terhardt (1974) drew the important distinction between *musical consonance* and *psychoacoustic consonance*. That is, functionally (according to music theory) a major third is a consonant interval regardless of other concerns; this sort of consonance is termed *musical consonance*. However, as discussed below, depending on the height of the tones involved and their timbre, the major third may be relatively more or less *psychoacoustically consonant*. Terhardt (1982) viewed psychoacoustic consonance as a sensory aspect of sound, and musical consonance as a cultural aspect of sound.

The view of consonance developed in antiquity and carried forward by Galileo and Descartes into the Renaissance held that the concept of consonance was just as simple as the statement above: consonance is that percept which results from the playing of sounds with frequency ratios in small-integer relationship. However, this statement seems to miss certain important characteristics of tonal consonance, for example that low tones in a major-third interval sound less consonant than high tones in the same interval. It also ignores the relationship between harmony and timbre; spectrally rich sounds are often less consonant with other sounds than spectrally simple sounds.

Helmholtz, in his fundamental 19th century work on tone perception, recognized that the true cause of consonance and dissonance was conflict between *overtones*, not fundamentals; tones with fundamental frequencies in a simple harmonic ratio share many overtones. Helmholtz viewed the sharing of overtones as the origin of the pleasing sound of consonant intervals.

A reexamination of the relationship between "frequency ratios" and "consonance" was undertaken by Plomp and Levalt in a now-classic paper (Plomp and Levelt 1965). They developed a model in which not only the ratio of fundamentals, but the overall spectral composition, was taken into consideration in analyzing consonance. They related the quality of consonance to the critical band model of the cochlea, producing compelling evidence that consonant pairs of complex tones are ones in which there are no (or few) harmonics competing within the same critical band, and dissonant pairs ones in which harmonics compete and cause a sensation of roughness. Finally, they demonstrated through statistical analysis of a small set of musical scores (J.S. Bach *Trio Sonata for Organ no. 3* and the 3rd movement of A. Dvorák *String Quartet op. 51*) that the statistical frequency of occurrence of various vertical intervals in musical works can be explained by a model in which composers are attempting to reduce critical band "clashes" among overtones. This work was highly influential and still stands as one of the few successful unifications of theories of music and psychoacoustics.

David Huron has engaged in a number of studies of the musical repertoire (of the sort Plomp and Levalt did on a smaller scale in the study cited above) using computerized tools that examine the relationship between certain psychological hypotheses and the statistical patterning of occurrence of musical constructs. One paper (Huron and Sellmer 1992) criticized Plomp and Levalt (1965) on methodological grounds, but arrived at the same conclusions with a more robust methodology. Another paper (Huron 1991) considered tonal consonance and tonal fusion – Huron analyzed a sample of keyboard works by J. S. Bach and found that Bach's use of vertical intervals could be viewed as a tension between a principle of avoiding tonal fusion, and one of seeking tonal consonance. He distinguished, following Bregman and Terhardt, the cases of "sounding smooth" (i.e., consonant) and "sounding as one" (i.e., fused). A valuable extension to this sort of project would be to develop similar rules in acoustic analysis and recreate the experiment using acoustic signals.

In all of these cases, the view of the relationship between timbre and harmonic structure was somewhat simplistic. Plomp and Levalt used a timbre model in which all sounds are represented by nine even-strength sinusoids with no temporal evolution; Huron (1991) failed to consider an interaction between harmony and timbre at all in one study; and used a single "average" timbre computed from static spectra of a range of musical instruments in another (Huron *et al.* 1992).

Sandell (1995) conducted one of the few studies integrating timbre, vertical structure, and acoustic data. He analyzed several acoustic features, including onset asynchrony, spectral centroid, and envelope shape in an attempt to determine what makes some pairs of tones "blend" better than other pairs (for example, why clarinet and horn blend but oboe and trumpet clash). He concluded that there is an important role for the spectral centroid—in cases where the tones formed a minor-third interval, both the overall spectral height and the distance between the individual tone centroids correlated negatively with "blend" as judged directly by musicians. He did not evaluate time-domain or correlation models of the acoustic signal in relation to his findings; a natural extension would be to examine whether the centroid interactions he found in his data can be explained using more fundamental operations on correlograms. He also considered his results in comparison to experiments in the "double-vowel" paradigm.

The broader literature on timbre is extensive but out of the scope of this proposal. A recent overview, presented from a similar philosophical stance to mine, was presented by Martin and Kim (1998a).

Composers of the late 20th century have been frequently interested in the relationship between harmony and timbre. Some, for example, Erickson (1985) believe that this is a continuum, and organize their own music to explore its boundaries and the "grey area" in between. McAdams and Saariaho (1985) published a set of "criteria for investigating the form-bearing potential of a proposed musical dimension," which they used to explore the possibilities of timbrally-organized music and the relationship between harmony and timbre in an analytical essay. Many composers are beginning to take results from psychoacoustics and auditory grouping theory as creative impetus to explore new areas of sonic texture (Hartmann 1983; Gordon 1987; Belkin 1988; Chowning 1990).

On the other hand, there has been relatively little work relating traditional music-theoretical views of musical elements such as "melody" and "accompaniment" to the acoustic signal. A paper by Povel and van Egmond (1993) claimed to provide evidence that melodic perception and harmonic perception are processed separately and interact little (which is highly contrary to traditional musical thinking). However, their study suffered greatly from not considering at all possible interactions between timbre, tonal fusion, and harmonic/melodic processing.

3. Music and emotion

There is a long-standing thread in music psychology which tries to understand how it is that music communicates emotion to listeners. This is often viewed as the primary function of music; Dowling and Harwood write:

> When we discuss our work with nonpsychologists, the questions that most often arise concern music and emotion. Music arouses strong emotions, and they want to know why ... Not only do listeners have emotional reactions to music, but pieces of music also *represent* emotions in ways that can be recognized by listeners. (Dowling and Harwood 1986, p. 201)

Much of the early (pre-1950's) work in musical emotion as reviewed by Dowling and Harwood focused on theories of musical *semiotics* – the ways in which listeners receive signs, indexical associations, and icons when they listen to music, and the ways in which such entities become associated with significands containing emotional connotations. Such work, by necessity, is highly theoretical and generally disconnected from practical issues of performance and acoustic realization; nonetheless, it has been a fertile and productive area of study.

The major organizing force in modern thinking on emotion in music was the work of L. B. Meyer, particularly an extensive volume on the topic. Meyer made a number of key points that continue to influence thinking about the aesthetics and philosophy of music today. First, he explicitly denies that music gives rise to a consistent, *differentiated* affective behavior (such as a sadness response), focusing instead on level of arousal. He equates the affective response to music with arousal, using terms familiar to musicologists such as *tension* and *release*, and attempts to relate the musicological use of this terminology to more accurate emotional-psychology definitions.

Second, he denies that the affective response to music is based on designative semiotics in the sense described above. Rather, he claims that all emotion and meaning in music is intra-musical; music only references itself and thus the excitement and tension in music are present only insofar as a listener understands the references. He thus views *expectation* and fulfillment or denial of expectation in music as the singular carriers of emotion in music.

A series of papers by Crowder and various collaborators (Crowder 1985a; Crowder 1985b; Kastner and Crowder 1990) evaluated the most well-known (not to say well-understood) of emotional distinctions in music: the association of the major tonality with "happiness" and the minor tonality with "sadness." Crowder *et al.* explored this issue variously from historical, experimental, aesthetic, and developmental viewpoints.

More recently, experimentalists have attempted to quantify emotional *communication* in music. For example, Juslin conducted an experiment in which professional musicians (guitarists) were instructed to play a short piece of music so as to communicate one of four basic emotions to listeners (Juslin 1997). He then analyzed acoustic correlates of tempo, onset time, and sound level in the performances, and tried to correlate these physical variables to the emotional intent of the performers. While he found that listeners were reliably able to detect the emotion being communicated, it was hard to determine exactly the physical parameters which conveyed the emotional aspects of the performance. However, Juslin must be credited for acknowledging the importance of the physical performance to transport and mediate the emotional material, through his adaptation of the "lens model" from emotion psychology.

A potential application of my research would be an attempt to correlate perceptual models of music-listening with affective response as measured through physiological measurement. This would be a valuable contribution to the music-and-emotion literature, which is somewhat impoverished with regard to serious psychoacoustical approaches, as well as a good opportunity for inter-group collaboration at the Media Lab. It would also be a useful example of a study in individual differences in the response to music, a topic which is largely unaddressed in the music-psychology literature.

4. Perception of musical structure

One of the primary ways in which both musicians and non-musicians understand music is through the perception of musical *structure.* This term refers to the understanding received by a listener that a piece of music is not static, but evolves and changes over time. Perception of musical structure is deeply interwoven with memory for music and music understanding at the highest levels, yet it is not clear what features are used to convey structure in the acoustic signal or what representations are used to maintain it mentally. I am interested in exploring the extent to which "surface" cues such as texture, loudness, and rhythm can explain the available data on the structural perception of music. This is in contrast to many theories of musical segmentation that assume this process is a crucially cognitive one, making use of elaborate mental representations of musical organization.

Clarke and Krumhansl (Clarke and Krumhansl 1990) conducted an experiment that analyzed listeners' understanding of sectional changes and "segmentation" in two pieces of music. One of the pieces was atonal or "textural" music (Stockhausen's *Klavierstück XI*), and one employed traditional tonal material (Mozart's *Fantasie*, K. 475). All of their listeners were highly trained musicians and/or composers. They present a characterization of the types of musical changes which tend to promote temporal segregation of one section from another, which include heuristics such as "Return of first material," "Arrival of new material," and "Change of texture." They also find that similar principles govern segmentation in the atonal and tonal works, and interpret their results as general support for the Lerdahl and Jackendoff grouping rules (Lerdahl and Jackendoff 1983), which are discussed below.

Deliège and her collaborators have conducted extensive studies (Deliège *et al.* 1996 for one) on long-term cognitive schemata for music; these studies include components similar to the Clarke and Krumhansl work that analyze musical segmentation. This research is connected to my own interests, as results on intra-opus segmentation might be extended to analyze similarities and differences *between* musical pieces as well as *within* pieces.

Clarke and Krumhansl also presented their work in the context of theories of time perception and dynamic attending, and Deliège in terms of long-term memory and music understanding; these topics fall less within the proposed scope of my dissertation, as they deal essentially with the long-term development of a single piece over time. Deliège *et al.* (1996) were also concerned with the relationship between "surface structure" and "deep structure" of music, and on the similarities and differences in processing between musicians and nonmusicians, which are topics of great interest to me. In particular, they wrote:

> [R]esults reported here can be thought of as providing information about processes that might be implicated in everyday music listening ... [R]esults for nonmusician subjects ... are indicative of a reliance on elements of the musical surface in listening to and manipulating the materials of complete tonal pieces ... [T]he primacy afforded to harmonic structure in Lerdahl and Jackendoff's theory may only be operational for musicians. (Deliège *et al.* 1996, p.153)

A different study by Krumhansl (1991a) has revealed data that support the importance of "musical surface." Using an atonal composition with a highly idiosyncratic rule structure (Messiaen's *Mode de valeurs et d'intensités*), she examined the perceptions of skilled musical listeners regarding the "surface" (pitch-class distribution and rhythmic properties) and "underlying" (strict correlations between pitch, loudness, and duration) properties of the music. She found that listeners were able to reject modifications to the surface structure as possible continuations of this work, but unable to reject modifications to the underlying structure. Listeners rapidly learned the "rules" of the surface structure of the piece—they performed as well on the first trial as on repeated trials—and were never able to learn the underlying structure. The latter was true even for professional musicians who had a verbal description of the rule structure provided to them.

N. Cook (1990), in a fascinating book on the topic, explored the relationship between musical understanding, musical aesthetics, and music structure.

5. Epistemology/general perception of music

Several writers—largely music theorists—have brought forth proposals for the manner in which music is "understood," or in which a listener "makes sense of it all." Many of these writings come from an attempt by music theorists to take a more "perceptual" stance, and include aspects of real-time thinking, retrospection, or perceptual limitations in their models.

The music theorist F. Lerdahl, in collaboration with the linguist R. Jackendoff, developed an extensive theory of musical "grammar" (Lerdahl *et al.* 1983). This theory has as its goal the "formal description of the musical intuitions of a listener who is experienced in a musical idiom." Their study takes the view that "a piece of music is a mentally constructed entity, of which scores and performances are partial representations by which the piece is transmitted." That is, the *piece* of music has a status that is neither the music-on-the-page (which generally concerns theorists) nor the music-in-the-air (which is my main interest).

The roots of such a theory in Chomskian linguistics are clear: Lerdahl and Jackendoff were concerned not with *performance* (in the linguistic sense, which includes operational details of memory and attention), but with *competence*: how can we characterize the mental structures of an idealized listener *after* listening has been completed? Their answer was an elaborate theory involving the integration of rhythm, phrasing, and structural rules with roots in Schenkerian music analysis.

Lerdahl and Jackendoff's model was highly influential in the development of the music-perception community. As such, it has drawn both support and criticism from experimenters and from other theorists. Although Lerdahl and Jackendoff asserted that their theory was not a theory of written music, in point of fact all of their analyses use only the written score as the starting point. Smoliar (1995) made a critical point which I feel is essential: that this theory was really one of musical structure, not musical perception. Jackendoff's linguistic concerns made these notions central in his thinking; however, Smoliar questioned their relevance to the "real" perception of music, as contrasted to the analysis of music-on-the-page. N. Cook (1990, Ch. 1) explicates the aesthetic stance represented by this idea when he constrasts the basic perception (hearing a work that is a sonata) with a more imagined or idealized perception (hearing a work *as* a sonata). Any listener (according to Cook) may easily *hear* a sonata, but more expertise is required to hear it *as* a sonata. The theory of Lerdahl and Jackendoff seems to mainly consider the latter case.

Lewin (1986) developed a sophisticated model of the phenomenology of music—that is, what goes on in the conscious mind during attentive music listening—and the relation between structural perception, memory, and expectation. This model was quasi-formalized using a structure akin to "frames" as used in artificial intelligence research. He used it to develop a theory which connected musical structure, abstracted away from any particular theoretical framework, to real-time listening and perception. He was particularly concerned with the relationship between expectation and realization and the "strange loops" (in the phrase of Hofstadter) this relationship creates in the perceptual structuring of music.

Lewin's theory came out of a larger literature on the phenomenology of time and music with which I am not generally familiar. He was especially critical of the conflation of structural theory with aesthetic theory and with perception theory – he clearly drew distinctions between music-as-cultural-artifact, music-as-acoustic-signal, and music-as-mental-object. He also drew a strong connection between music perception and music-as-behavior. He apparently believed that music is not "perceived" unless it leads to a creative music act by the listener, such as a composition, an article of criticism, or a performance. I disagree with this stance, excepting the circular case where any listening is a "creative act." However, his crucial point that listening is a *behavior* (and thus can only be properly addressed with progressive models) is also a central part of my approach.

Minsky (1989) presented a discussion of how music-listening could be viewed in the context of his well-known and influential "society of mind" theory of intelligence (Minksy 1985), although this book did not itself treat music. In Minsky's view, music-listening is best understood as the interaction of *musical agents*, each with a particular focus. For example, *feature-finders* "listen for simple time-events, like notes, or peaks, or pulses;" and *difference-finders* "observe that the figure *here* is same as that one *there*, except a perfect fifth above." Minsky's primary concern was with the "highest" levels of musical perception and cognition, in which the full panoply of human intellectual abilities is used. He also presented interesting thoughts on music appreciation—where does "liking" music come from—and argued that it is a long-term similarity matching process, that we only like music that is similar to other music we like in some (unspecified) structural ways (it is not clear how he believes this process is bootstrapped).

Minsky also argued strongly for linkages between music and emotion as the primary motivating factor behind music-making; indeed, as the primary reason for the existence of music. He argued that music *encapsulates* emotional experiences and allows us to examine them more closely. His arguments were largely drawn from his own intuitions about music-making (he improvises on piano in the style of Bach) and conversations with music theorists and others, not from the music literature or experimental evidence.

M. Clynes has spent many years developing an elaborate theory of "composers' pulses," which he claims are important cues to the correct performance of Western classical music in various styles. He has recently (Clynes 1995) presented evidence that listeners as well may be sensitive to these systematic performance-time deviations; if true, such a feature might be used to distinguish composers working within this tradition from one another.

It seems difficult for many theorists involved in this sort of work to avoid making prescriptive judgments; that is, to use their theoretical stance as an argument for the "goodness" or "badness" of types of music or types of listeners. For example, Minsky argued against the repetitiveness he perceives in popular music:

> [W]e see grown people playing and working in the context of popular music that often repeats a single sentence or melodic phrase over and over and over again, and instills in the mind some harmonic trick that sets at least part of one's brain in a loop. Is it o.k. that we, with our hard-earned brains, should welcome and accept this indignity—or should we resent it as an assault on an evident vulnerability? (Minsky and Laske 1992, p. xiv)

Lewin complained that not enough music-listeners are music-makers, and alluded to this fact as emblematic of larger cultural problems:

> In other times and places, a region was considered "musical" if its inhabitants habitually made music, one way or another, to the best of their various abilities; nowadays and here, regional music "lovers" boast of their "world-class" orchestras (whose members probably commute), their concert series of prestigious recitalists, their improved attendance at concerts (especially expensive fund-raising concerts), their superb hi-fis, their state-of-the-art compact disc players, and so on. (Lewin 1986, p. 380)

As a closing remark, the composer and computer scientist D. Cope has built several computer systems that can mimic the style of various composers (Cope 1991; Cope 1992). These systems work from a stochastic analysis-synthesis basis: they take several pieces by a composer as input, analyze the style using statistical techniques, and "recombine" the pieces into new pieces with similar statistical properties. This work is a fascinating exploration into the mechanisms of musical style; however, it has been popularly cited (and argued by Cope himself) as an example of "musical artificial intelligence". To me, it is more a demonstration of how easy it is to create new works in the style of old composers—most music students can write simple works in the style of Bach, Mozart, Beethoven, and so forth—and how simple is the perceptual apparatus that is used to do such categorization. A truer sort of "intelligent" musical system would itself be able to evaluate the quality of its compositions, identify composers, or perhaps innovate *new* styles rather than only replicating what is given.

6. Musical experts and novices

As a final note, it is important to realize that the vast majority of findings I have summarized in this subsection are taken from research on musical *experts*; that is, composers, music graduate students, and analysts with many years of formal training and scholarship. It is not at all clear that these results extend to the perceptions of non-musician "naive" listeners. In fact, the evidence is quite the opposite. An excellent article by Smith (1997) reviewed the literature on the music perceptions of non-musicians, then made this point incisively:

> The situation is that novices do not resonate to octave similarity; they often cannot identify intervals as members of overlearned categories; they seem not to know on-line what chromas they are hearing; in many situations, they may even lack abstract chroma and pitch classes; they seem not to appreciate that the different notes of their scale have different functional and closural properties; they have little desire for syntactic deviation or atypicality; they dislike the formal composition elegance that characterizes much of Western music during its common practice period; indeed, they reject music by many of the composers that experts value most. (Smith 1997, pp. 251-252)

A paper by Robinson (1993) went even further, to suggest that for non-musical listeners, even pitch is a cue of low salience compared to "surface" aspects such as broad spectral shape!

This is not to dismiss non-expert listeners as musically "worthless"; rather, it is to say that unless we want music systems and theories of music perception to have relevance only to the skills and abilities of listeners who are graduate-level musicians, we must be cautious about the assumptions we follow. Naïve listeners can extract a great deal of information from musical sounds—for example, Levitin and colleagues have shown (Levitin 1994; Levitin and Cook 1996)that non-musician listeners can not only extract, but preserve in long-term memory, the absolute tempo and absolute pitch of popular songs that they like—and they are likely to be the primary users of many of the applications we would like to build. At present, we have no theoretical models of music-listening from acoustic data that explain the behavior even of the least musically-sophisticated human listener. I will return to this point in Section III.9.

An alternative thread of research which relates to the relationship between novices and experts with which I am less familiar is the study of the development of musical thinking (Serafine 1988; Kastner *et al.* 1990; Bamberger 1991).

C. Musical signal processing

The third area of research I review reports on the construction of *musical-signal-processing systems*. Researchers have attempted to build systems which could analyze and perform music since the dawn of the computer age; the motivation of making computers able to participate musically with humans at some level appears to be a strong one.

By musical signal processing, I mean the study of techniques to apply in the analysis (and synthesis) of musical signals. There are overlaps between general audio signal processing and musical signal processing; for example, the FFT is useful in many ways to analyze musical signals as well as others. However, I am especially interested in signal-processing techniques that are specifically targeted to musical signals.

I review four main areas of research. *Pitch-tracking* systems are computer systems which attempt to extract pitch from sound stimuli. They differ from the pitch models in Section II.A.1 primarily in focus—the systems described here are meant for practical use, not scientific explication. A related area is *automatic music transcription* or *polyphonic pitch tracking*; these systems attempt to extract multiple notes and onset times from acoustic signals and produce a musical score or other symbolic representation as output. A short digression on representations for musical signal processing follows. I also summarize recent research on tempo and beat analysis of acoustic musical signals; such systems attempt to "tap along" with the beat in a piece of music. Finally, I describe a few recent systems which demonstrate classification of whole audio signals.

While most of these systems are motivated from an engineering viewpoint, not a scientific one, it is important to appreciate the connection between these goals. Systems with scientific goals may have great practical utility in the construction of multimedia systems (Martin, Scheirer and Vercoe 1998b); conversely, the activities involved in prototyping working systems may lead to better scientific hypotheses regarding music perception and psychoacoustics. P. Desain and H. Honing have been among the strongest proponents of the view that research into modeling *per se*, or even engineering-oriented "computer music" research, can lead us to better insights about the music perception process in humans. They articulated this view in a paper (Desain and Honing 1994) which argued on the basis of research into "foot tapping" systems that pursuing well-organized engineering solutions to certain problems can result in more robust formal models of music-listening.

This summary covers only half of the world of musical signal processing; the other half is that devoted to sound systems and sound effects algorithms. There is an extensive literature on these topics that has recently been exhaustively reviewed in tutorial form by Roads (1996). Recent papers of a more technical bent are collected in a volume edited by Roads and three colleagues (1997).

1. Pitch-tracking

Perhaps the most-studied engineering problem in musical signal processing is *pitch-tracking* – extracting pitch from acoustic signals[2]. This task has wide practical application, as it is used both in speech coding as an input to linear-prediction models of speech (Makhoul 1975), and in music systems, where it is used to "follow" or "ride" acoustic performances and control musical input devices.

An early review of pitch detection algorithms (Rabiner *et al.* 1976) was presented in the context of speech-processing systems. Rabiner and his colleagues gave references and comparatively tested seven methods from the literature on a speech database using a variety of error measurements. They concluded that no method was overall superior, but that different techniques were useful for different applications. Note that there was no direct comparison with perceptual models of pitch such as the template-matching or autocorrelogram models described above.

More recently, studies have focused on developing models of pitch that respect the known psychoacoustic evidence as discussed above, but are still efficient and practical for computerization. Van Immerseel and Martens (1992) constructed an "auditory model-based pitch extractor" which performs a temporal analysis of the outputs emerging from an auditory filterbank, haircell model,

[2] Most systems of this sort are more properly described as *fundamental frequency* tracking systems. Engineering approaches are typically not concerned with the wide variety of signals that give rise to pitch sensation as discussed in Section II.A.1, and the engineering applications cited actually make better use of fundamental frequency information that a true "pitch" analysis. I use the incorrect ("pitch") terminology in this section for consistency with the literature.

and envelope follower. They reported real-time performance and robust results on clean, band-pass-filtered, and noisy speech.

In commercial music systems, pitch-tracking is often used to convert a monophonic performance on an acoustic instrument or voice to symbolic representation such as MIDI. Once the signal has been converted into the symbolic representation, it can be used to drive real-time synthesis, interact with “synthetic performer” systems (Vercoe 1984; Vercoe and Puckette 1985) or create musical notation. Especially in the first two applications, systems must be extremely low-latency, in order to minimize the delay between acoustic onset and computer response. Vercoe has developed real-time pitch-tracking systems for many years for use in interactive performance systems (Vercoe 1984); a paper by Kuhn (1990) described a robust voice pitch-tracker for use in singing-analysis systems.

2. Automatic music transcription

In a broader musical context, pitch-tracking has been approached as a method of performing *automatic music transcription*. This task is to take acoustic data, recorded from a multiple-instrument musical performance, and convert it into a “human-readable” or “human-usable” format like traditional music notation. As best I can determine, Piszczalski and Galler (1977) and Moorer (1977) coined the term contemporaneously; however, I prefer not to use it, because it conflates the issue of performing the musical analysis with that of printing the score (Carter, Bacon and Messenger 1988). It is the former problem that is more related to this proposal, and I will term it *polyphonic pitch tracking*.

There are many connections between the polyphonic pitch-tracking problem, the engineering approaches to auditory source segregation discussed in Section II.A.2, and research into sound representations, which will be discussed in Section II.C.3. Many researchers in musical-signal-processing equate the problems of music source segregation and polyphonic pitch-tracking. I do not; I will discuss this viewpoint in Section II.3.

The work by Piszczalski and Galler (1977) focused only on single instrument analysis, and only “those instruments with a relatively strong fundamental frequency.” Thus, their work was not that different from pitch-tracking, except that the system tried to “clean up” results for presentation since it was not operating in real-time. Their system operated on an FFT front-end, and tried to measure the fundamental directly from the spectrogram.

Moorer’s system (Moorer 1977) was the first in the literature to attempt separation of simultaneous musical sounds. His system could pitch-track two voices at the same time, given that the instruments were harmonic, the pitches of the tones were piece-wise constant (i.e., no vibrato or jitter), the voices did not cross, and the fundamental frequencies of the tones were not in an 1:N relationship (unison, octave, twelfth, etc). He demonstrated accurate analysis for a synthesized violin duet and a real guitar duet obeying these constraints. His system, like Piszczalski and Galler’s, worked directly from a short-time spectral analysis.

The research group at Stanford’s CCRMA did extensive research in polyphonic pitch-analysis during the early 1980s. They began with work on monophonic analysis, beat-tracking and notation systems (Foster, Schloss and Rockmore 1982); these systems were envisioned to support an “intelligent editor of digital audio” (Chafe, Mont-Reynaud and Rush 1982). This early work soon gave way to a concerted attempt to build polyphonic pitch-tracking systems, largely for acoustic piano. Their reports (Chafe *et al.* 1985; Chafe and Jaffe 1986) were heavy on details of their analysis techniques, which were based on grouping partials in sinusoidal analysis, but very sketchy on results. It is unclear from their publications whether their systems were particularly successful, or whether they attempted validation on music with more than two voices.

Vercoe (1988) and Cumming (1988) presented sketches of a system which would use a massively parallel implementation on the Connection Machine to perform polyphonic transcription using modulation detectors over the output of a frequency-decomposition filterbank, but this system was never fully realized.

R. C. Maher's work (Maher 1990) resulted in the first system well-described in the literature which could perform relatively unconstrained polyphonic pitch-tracking of natural musical signals. He developed new digital-signal-processing techniques which could track duets from real recordings, so long as the voices did not cross.

The front-end of his system used McAuley-Quateiri (MQ) analysis (1986), which represents the signal as the sum of several quasi-sinusoids which vary over time in frequency and amplitude. He extended the MQ analysis to include heuristics for the analysis of "beating," which occurs when the frequency separation of two partials becomes smaller than the resolution of the short-time Fourier analysis component of the MQ analysis. He also developed a "collision repair" technique which could reconstruct, through interpolation, the damage that results when multiple sinusoids come too close in frequency and cause phase interactions. He considered but abandoned the use of spectral templates to analyze timbre. Finally, Maher performed a meaningful evaluation, demonstrating the performance of the system on two synthesized examples and two natural signals, a clarinet-bassoon duet and a trumpet-tuba duet.

Kashino and his colleagues have used a variety of formalizations to attempt to segregate musical sounds. An early system (Kashino and Tanaka 1992) performed a sinusoidal analysis and then grouped partials together using synchrony of onset, fundamental frequency, and common modulation as cues. The grouping was performed using a strict probabilistic framework. A second system (Kashino and Tanaka 1993) used dynamic timbre models in an attempt to build probabilistic expectations. Tested on synthesized random two- and three-tone chords built from flute and/or piano tones, this system recognized 90% and 55% of the stimuli correctly, respectively.

More recently, Kashino's efforts have been focused on the "Bayesian net" formalism. A system built by Kashino *et al.* (1995) used both an analysis of musical context and low-level signal processing to determine musical chords and notes from acoustic signals. Operating on sampled flute, trumpet, piano, clarinet, and violin sounds, their system identified between 50% and 70% of chords correctly for two-voice chords, and between 35% and 60% (depending on the type of test) for three-voice chords. They found that the use of musical context improved recognition accuracy between 5% and 10% in most cases. An extended version of this system (Kashino and Murase 1997) recognized most of the notes in a three-voice acoustic performance involving violin, flute, and piano.

Hawley's dissertation at the Media Lab (Hawley 1993) discussed piano transcription in the context of developing a large set of simple tools for quick-and-dirty sound analysis. He used a short-time spectral analysis and spectral comb filtering to extract note spectra, and looked for note onsets in the high-frequency energy and with bilinear time-domain filtering. He only evaluated the system on one example (a two-voice Bach piano excerpt without octave overlaps); however, Hawley was more interested in describing the *applications* of such a system within a broad context of multimedia systems than in presenting a detailed functional system, so evaluation was not crucial.

My own master's thesis (Scheirer 1995; Scheirer 1998c) extended the idea of incorporating musical knowledge into a polyphonic transcription system. By using the score of the music as a guide, I demonstrated reasonably accurate transcription of overlapping four- and six-voice piano music. The system used both frequency-domain and time-domain methods to track partials and detect onsets. I termed this process "expressive performance analysis," since the goal was to recover performance parameters with accurate-enough time resolution to allow high-quality resynthesis and comparison of timing details between performances.

This system was validated by capturing both MIDI and audio data from the same performance on an acoustic-MIDI piano (Yamaha Disklavier). My algorithms were applied to the acoustic performance; the symbolic data thus recovered was compared to the ground-truth MIDI data. I tested the system on scales, on a polyphonic piece with many overlaps (a fugue from the *Well-Tempered Clavier* of Bach), and on a polyphonic work with pedaling and many simultaneous onsets (Schubert *Kinderszenen*). The system proved accurate enough to be used for tempo analysis and some study of expressive timing. It was not generally good enough to allow high-quality resynthesis. It stands as a proof-of-concept regarding expectation and prediction; that is, if the expectations/predictions of a polyphonic pitch-tracking system can be made good enough, the signal-processing components can succeed. Of course, building a system that can generate good expectations is no easy task.

Most recently, Martin (1996b) has demonstrated use of a blackboard architecture (which was a technique also suggested at CCRMA) to transcribe four-voice polyphonic piano music without using high-level musical knowledge. His system was particularly notable for using the autocorrelogram as the front end (Martin 1996a) rather than sinusoidal analysis. Rossi *et al.* (Rossi, Girolami and Leca 1997) built a system for polyphonic pitch identification of piano music around an automatically-collected database of example piano tones. The spectra of the tones were analyzed and used as matched filters in the spectral domain. This system also transcribed four-voice music correctly. Although neither of these systems has been extensively evaluated (for example, they were not tested with music containing overlapping notes, only with simultaneous onsets), they currently stand as the state of the art in polyphonic pitch-tracking.

The recent development of *structured audio* methods for audio coding and transmission (Vercoe, Gardner and Scheirer 1998) ties research in musical signal processing (especially polyphonic pitch-tracking and music synthesis) to new methods for low-bitrate storage and transmission. In a structured audio system, a representation of a musical signal is stored and transmitted using an algorithmic language for synthesis such as Csound (Vercoe 1996) or SAOL (Scheirer 1998a) and then synthesized into sound when it is received. High-quality polyphonic pitch tracking, timbre analysis, and parameter estimation are necessary in this framework if the structured descriptions are to be automatically created from sound. Currently, structured audio descriptions must be authored largely by hand, using techniques similar to MIDI composition and multitrack recording. Vercoe *et al.* (1998) reviewed much of the same literature I have examined here, in the context of articulating the goals of structured audio transmission and representation.

3. Representations and connections to perception

Every analysis system depends upon representation and transformation. The representation used informs and constrains the possible analysis in subtle and potentially important ways. I provide a short digression here to discuss the most common signal representations used in music-pattern-recognition systems. A recent collection (De Poli, Piccialli and Roads 1991) extensively described many different representations suitable both for music analysis and for music synthesis.

The earliest discussion of signal processing in the ear was by Ohm and Helmholtz, who observed that the ear is like a Fourier analyzer; it divides the sound into spectral components. But Gabor (1947) objected that Fourier theory is an abstract, infinite-time basis of analysis, while humans understand sound as evolving over time. He developed a theory which has now come to be called " wavelet analysis" around the problem of simultaneous analysis of time and frequency.

Many engineering approaches to signal analysis have utilized one or another time-frequency transformation with mathematically simple properties, including the Discrete Fourier Transform (Oppenheim and Schafer 1989), the "constant-*Q*" transform (Brown 1991; Brown and Puckette 1992) or various "wavelet transforms" (Kronland-Martinet and Grossman 1991). These models have varying degrees of affinity with the current understanding of perceptual frequency analysis by the cochlea; the *gammatone filterbank* is a closer approximation used in perceptual studies.

These representations also have various properties of time/frequency resolution and efficiency which make them more or less suitable for use in various applications. For example, the DFT can be computed using the Fast Fourier Transform which is extremely efficient to calculate. A recent review article (Pielemeier, Wakefield and Simoni 1996) compared the properties of various time/frequency representations.

Musical-signal-processing systems have often worked by transforming a signal into a time-frequency distribution, and then analyzing the resulting spectral peaks to arrive at a representation which tracks sinusoids through time. This may be termed an *additive synthesis* model of musical sound, and techniques such as the phase vocoder (Flanagan and Golden 1966) and McAuley-Quatieri analysis (McAulay *et al.* 1986) can be used to extract the sinusoids. These techniques have also been used in a more perceptually-motived framework; Ellis (1994) used a McAuley-Quatieri front end to extract sinusoids from musical sound and then developed grouping heuristics based on those of Bregman to regroup them as sources.

The most sophisticated model to use this approach was the work of A. Wang in his dissertation (Wang 1994). He developed a set of signal-processing techniques for working with a number of novel types of phase-locked loops (PLLs), including notch-filter PLLs, comb-filter PLLs, frequency-tracking PLLs, and harmonic-set PLLs. He applied these techniques to the separation of voice from musical signals, and evaluated the resulting system on several real-world musical examples, with quite satisfactory results. He also discussed the relationship between this sort of heavy-duty signal processing and models auditory perception, although to some degree the question is left hanging in his presentation. He did provide straightforward discussion of the practical utility of source separation systems for applications other than perceptual modeling.

Among the contributions made by Ellis in his dissertation (1996a) was a novel intermediate representation he termed the *weft* (Ellis 1997; Ellis and Rosenthal 1998). The weft allows simultaneous representation of pitch and spectral shape for multiple harmonic sounds in a complex sound scene. He provided algorithms for extracting wefts from autocorrelograms as well as details on their use in sound-scene analysis. Ellis (1996a) also discussed the general problem of trying to discover the true perceptual representations of sound.

4. Tempo and beat-tracking models

A final commonly-approached problem in musical signal processing is *beat-tracking* or *foot-tapping*; that is, the construction of systems which can “find the beat” in a piece of music. There is a fair amount of experimental data on special, non-ecological examples collected in studies that try to explain the behavior of humans, but I am not aware of any perceptual data for “real music” except that contained in the short validation experiment in my own work (Scheirer 1998b) on this topic. Beat perception models are important because the beat of a piece of music is a universal, immediately perceived feature of that piece, and it is thus crucial in developing an understanding of how listeners orient themselves to a new composition.

Many early systems, as reviewed by Desain (1995) and Scheirer (1998b), attempted to explain the perception of beat (and rhythm in some cases) starting from a series of inter-onset-interval times. This model assumed that prelimary processing has produced an onset stream. Of more interest in the present context are the few systems which can operate directly on the acoustic signal; I review three here.

Vercoe (1997) reported a system for beat-tracking acoustic music which used a constant-*Q* filterbank front-end and a simple inner-hair-cell rectification model. The rectified filter channels were summed, and a “phase-preserving narrowed autocorrelation” analyzed the periodicity of the signal. The output of the model was primarily visual; Vercoe provided anecdotal verification of its performance on a piece of simple piano music.

M. Goto and his collaborators have done extensive research into real-time beat-tracking systems. Two systems that they built could perform real-time acoustic beat-tracking on real musical signals, and even make simple rhythmic judgments. In the first (Goto and Muraoka 1998), a template-matching model was used, in which signal processing models attempted to locate bass drum and snare drum sounds in a complex sound source; the extracted onset times were then matched against a database of known patterns. The system could accurately track tempo, beat, half-note, and measure from rock-and-roll songs which use these drum patterns.

A further extension to this system used more sophisticated signal-processing (Goto and Muraoka 1997) to extract onset times from signals without drums and make similar judgments to the earlier system. It used a robust onset-detector and chord-change locator as the fundamental features which drive the system. Goto and Muraoka reported excellent results for both of these systems on rock-and-roll music, although it is not clear whether their methods were applicable to other genres or to signals with changing tempo. They also used their system to drive interactive computer graphics.

I have also done work in music beat-tracking (Scheirer 1998b). My system was somewhat similar to Vercoe's, but was evaluated more extensively and demonstrated to give good results on a wide range of music. It could also track tempo changes (if not too unusual) and was analytically somewhat simpler than Goto's. This system operated by decomposing the signal into six bands with sharply-tuned bandpass filters, and then analyzing the periodicity of each band's envelope independently using envelope detection and banks of parallel comb filters. The periodicity was detected within each channel by searching through the lag space of the multiple comb filters; the feedback delay with maximum response was selected. The estimates from the multiple subbands were combined to give an overall estimate, and then the beat phase of the signal was estimated using simple heuristics.

The algorithm performed very well on any sort of music in which the beat was conveyed clearly without a lot of syncopation. It had a difficult time analyzing signals where the beat was very slow, conveyed by instruments with slow onsets, or highly syncopated. The system was validated by testing it on 60 musical excerpts of various styles, including rock, jazz, classical, new-age, non-Western music, and salsa, and comparing the performance to human performance in a psychoacoustic experiment. The behavior of the system was demonstrated to be quantitatively similar to that of human listeners engaged in a "tapping task" where they tap along with the musical signals.

This system and Vercoe's are different than other rhythm-analysis techniques that have been presented in the literature in two ways relevant to the present discussion. Most importantly, no explicit segmentation, note analysis, or source-separation was undertaken to enable the analysis. Other systems explicitly or implicitly assumed that the signal had already been, or needed to be, transcribed into sequences of note onsets before rhythmic analysis could occur. These systems were much "closer to the signal." Rather than representing tempo analysis as the high-level juxtaposition of multiple notes output by a separation system, this model of tempo viewed tempo itself as a fundamental low-level feature of the acoustic signal.

Additionally, there are strong connections between the analysis model developed for my system and components of modern hearing theory (Scheirer 1997). Both my rhythm-analysis model and the temporal model of pitch hearing combine a front-end filterbank and rectification process with a subband periodicity analysis and cross-band integration stage. In fact, I demonstrated (Scheirer 1997) that existing models of pitch could be used directly to analyze tempo from acoustic musical signals.

Beat-tracking is of potentially great interest in the construction of general music perception systems; it has been implicated, especially in the work of M. R. Jones (Jones and Boltz 1989) as a strong cue to attentional set. That is, Jones argued, the rhythmic aspect of music provides a framework for alternately focusing and relaxing attention on the other features of the signal. If possible, I will try to integrate my beat-perception model into whatever music analysis systems I build.

5. Audio classification

As well as the specifically music-oriented systems described above (and the great wealth of speech-recognition and speech-analysis systems not discussed here), there have been a few efforts to conduct a more general form of sound analysis using techniques from the pattern-recognition literature.

I have previously conducted one well-organized study in sound-signal classification; M. Slaney and I built a system which could robustly distinguish speech from music by inspection of the acoustic signals (Scheirer and Slaney 1997). This system was an exemplar of the classical approach to pattern recognition; we chose 13 features by developing heuristics we thought promising, and then combined them in several trained-classifier paradigms. The system performed well (about 4% error rate, counting on a frame-by-frame basis), but not nearly as well as a human listener. This was not a "machine listening" system—there was little attempt to consider the perceptual process when building it.

There has been some recent work to attempt classification of sounds, musical excerpts, or even whole soundtracks by type. E. Wold *et al.* (1996) described a system which analyzed sound for pitch, loudness, brightness, and bandwidth over time, and tracked the mean, variance, and autocorrelation functions of these properties to create a feature vector. They reported anecdotally on the use of a simple spatial partition in the resulting feature space to classify sounds by type and of the use of simple distance metrics to do perceptual similarity matching. They included speech, sound effects, and individual musical notes, but did not report results on musical style classification or speech-vs.-music performance. They attempted to justify their feature set on perceptual grounds, but not the combination of features nor the way they were used in applications. They also provided a useful list of potential applications for this sort of system.

Smith *et al.* (Smith, Murase and Kashino 1998) discussed the time-domain use of zero-crossing statistics to retrieve known sounds very quickly from a large database. Any noise in their system had to be minimal and obey simple statistical properties; their method was not robust under signal transformation or in the presence of interfering sounds.

Dannenberg *et al.* (1997) reported on a pattern recognition system which classified solo improvised trumpet performances into one of four styles: "lyrical," "frantic," "syncopated," or "pointillistic" (such classes are useful for real-time collaboration in a modern jazz idiom). They used a 13-dimensional feature set taken from MIDI data acquired through real-time pitch-tracking with a commercial device. These features included average and variance of pitch height, note density, and volume measures, among others. Using a neural network classifier, they reported 1.5% error rate on ten-second segments compared to human classification.

A recent article (Foote in press) reviews other systems for automatic analysis of audio databases, especially focusing on research from the audio-for-multimedia community. This review is particularly focused on systems allowing automatic retrieval from speech or partly-speech databases, and thus makes a good complement to my music-focused approach here.

A fascinating system constructed by Katayose and Inokuchi (1989) attempted to model emotional reactions as a direct pattern-extraction system. Their system first transcribed a musical signal, or scanned a musical score using optical music recognition (Carter *et al.* 1988), and then "extract[ed] sentiments using music analysis rules that extract musical primitives from transcribed notes and rules that describe the relation between musical primitive and sentiments." The report of their system was still highly speculative. As music seems to be an important vehicle of emotional communication between humans, the development of truly “affective” computer systems (Picard 1997) requires more research of this sort.

D. Recent cross-disciplinary approaches

In this section, I examine some recent studies which report attempts similar to the project I propose; that is, to construct computational systems with a basis in the study of human perception that can analyze acoustic musical signals. These studies are the most direct background for my research.

E. Terhardt was the first to make a concerted effort at the computational construction of perceptual models of musical phenomena. His research is broad and yet disciplined, with a constant focus on presenting correct and consistent definitions of terms and maintaining a separation between structural, physical, and perceptual aspects of music. He also has a great knowledge of the broad directions of the field of computer music, as evidenced by an essay on that topic (Terhardt 1982). He is best known today for his “virtual pitch” model, which is a template-matching model of the pitch of complex tones (Terhardt 1974).

He extended this model to analyze the roots of musical chords (Terhardt 1978). He observed that the relationship between chord tones and chord roots is very similar to the relationship between overtones and fundamentals in complex tones. Thus, his model predicted the root of a chord as the subharmonic which best explains the overtones as “harmonics.” While this model is not without its problems, it has been the most influential model of acoustic correlates of musical harmony to this point.

R. Parncutt, a student of Terhardt, extended and formalized this model in a number of important directions (Parncutt 1989). His model, based on four free parameters that control how analytic the modeled listener is, predicts pitch, tonality (how audible the partials of a complex tone are), and multiplicity (the number of tones noticed in a chord) of individual sounds and the similarity of sequential sounds. The model incorporated masking and loudness effects and a template-matching model.

Parncutt (1989) conducted extensive psychoacoustic tests to determine listeners’ settings for the model’s free parameters and to psychophysically validate its principles. He shows that the model’s predictions are generally accurate and that the model can also be used to determine musical key and the roots of chords. The model takes as input a fully-resolved spectrum; that is, a list of all the harmonic components present in a chord. Thus, all of his testing was conducted on synthetic test sounds rather than ecological music examples; a logical next step would be to attempt the analysis of real music for spectral components and use this as input to his model.

Recent psychoacoustic experiments have provided further evidence for Parncutt’s model. Thompson and Parncutt (1997) found that the model alone could explain 65% of the variance in a chord-to-tone matching task, and 49% of the variance in a chord-to-chord matching task. This indicates that the model is a good predictor of whether subjects (expert musicians in their case) find two chords (or a chord and a complex tone) “similar” or “different.”

Parncutt has continued to conduct work on modeling the perception of chords. A recent chapter (Parncutt 1997) took a more directly music-theoretical stance, attempting to explain the root of a chord using a sophisticated model incorporating the pitches in the chord, its voice, the local key sense (“prevailing tonality”), and voice-leading considerations. As he admitted, this model has not been systematically tested.

D. K. Mellinger did extensive work in his thesis (1991) attempting to unify contemporaneous results from music transcription and auditory scene analysis in a robust polyphonic music-analysis system. He developed a set of two-dimensional filters which operated on a time-frequency "cochleagram" image. By using this technique, the sound-understanding problem can make use of techniques developed in the image processing world.

Mellinger's technique was explicitly both perceptual and musical – he cited Bregman (1990) and Marr (1982) as influences in the perceptual theory, and considers source grouping only in music, not in other sounds. The system was similar to the sinusoidal analysis systems cited in Section II.C.2 in that it operated in time-frequency space; however, it used a perceptually motivated front-end (Slaney 1994) rather than a spectral analysis, and analyzed large stretches of time at once through the use of two-dimensional filtering in the time-frequency place. Also notable was the use of modulation-detection kernels that were convolved with the image to detect frequency variation of partials.

Mellinger analyzed his system's performance on a few ecological music examples. The system performed well on a piano performance with two-voice polyphony, and could separate simultaneous synthesized complex tones with vibrato, by making use of common-modulation cues between partials. It had a much harder time with real instruments with slow onsets – it could not accurately separate voices from a Beethoven violin concerto or from sections of a Beethoven octet with two to four woodwinds playing at once.

The approach most similar to mine is that of Leman (1994; 1995). Leman presented methods for analyzing acoustical signals and, using the Kohonen-map framework, allowing a self-organizing architecture to represent tonality in music. His system consists of two parts with interchangeable mechanisms:

1. An auditory model, which incorporated either the Terhardt model for pitch analysis or a cochlear filterback process based on the model of Van Immerseel and Martens (1992) and a short-term within-channel autocorrelation (which he calls the "tone completion image" although it was essentially equivalent to the Licklider (1951b) model for pitch).

2. A self-organizing tone-center cognitive model based on the Kohonen-map formalism. In this technique, the output of the auditory model was presented as vectors of data to a two-dimensional grid of computational neural-network elements. As successive stimuli were presented, the map self-organized, and topological patterns of regions of grid neurons responding to certain sorts of inputs began to appear. At this point, when new stimuli were presented, they activated certain regions of the self-organizing map strongly; this regional allocation was a form of classification.

There are interesting comparisons between this sort of connectionism and the sort represented in the oscillatory grouping system of Wang (1996). While the latter views the dynamic evolution of the oscillatory state as critical—since source grouping is represented through coherent oscillation—the former hopes that the oscillations will die away and the map will reach stability.

Leman trained the network with a set of synthesized cadence sequences, moving through all keys. He viewed this set of training data as representative of the important relationships in tonal music. He then analyzed the performance of his methods in two ways. First, he examined the internal structure of the Kohonen map after learning had occured. Certain structural equivalencies to theories of tonal music perception as described in Section II.B.1 were present; the Kohonen map was an accurate reflection of "tonality space" as predicted by such theories. Second, he played musical examples to the system after it had been trained, and let it produce output judgments of tonality moving through listening time. He included three examples of real, ecological music (Debussy *Arabesque No. 1* and Bartók *Through the Keys* for solo piano, and Brahms *Sextet No. 2* for string sextet), and found that the judgments of the model corresponded reasonably well with the analysis of music theorists.

My approach will differ from Leman's in a number of ways. Most importantly, Leman had as a goal the *analysis* of music – the subtitle of his book is "cognitive foundations of systematic musicology." As such, he is interested in *musicology* and hopes to provide tools of interest to music theorists in understanding the nature of music. He is less interested in understanding the perceptions of everyday listeners than in showing correspondence with the predictions of music theorists. In contrast, I am much more interested in the perceptions of non-musically skilled listeners; there are serious questions about whether the kinds of judgments his system can make are perceptually relevant to non-musicians.

Second, he did not try to incorporate models of source grouping and segregation in his model. The perception of music with multiple sources (even if some sources are "virtual"; see Section III.4) surely depends on our ability to segrate the sound scene into multiple sources and attend to them holistically or selectively. One of the first components of my approach will be a re-examination of musical source grouping using the autocorrelogram framework. Leman and I share the goal of building *subsymbolic* models of music processing (Leman 1989) – models that operate directly on acoustic signals and do not involve an explicit symbolization step. Leman's work involves self-organizing structures as subsymbolic components, whereas mine will be based on signal-processing and pattern-recognition techniques. Finally, I hope to demonstrate the performance of my models on a much broader range of stimuli than Leman did.

It is important to understand that I am not criticizing the *physiological* plausibility of Leman's model (or any of the other models I discuss). It is unlikely that the human perceptual system analyzes sound in exactly the *manner* proposed by any existing or near-future model. My criticism is only lodged on *behavioral* grounds; I don't believe that Leman's model mimics the behavior of human listeners accurately, and I hope to do better at this.

Drawing from the approach of Leman, Izmirli and Bilgen (1996) presented a model for dynamically examining the "tonal context" key of music from acoustical signals. The method used was straightforward: they tracked the acoustic strength of fundamental pitches from audio processed using a constant-*Q* transform, and measured the note onsets and offsets. The pitch results were collapsed across octaves to form a 12-element vector, which was averaged over time using leaky integration. This vector was correlated with the Krumhansl (1990) tonal-context profiles for various keys, and the low-passed result was taken as the dynamic key measurement. This method can easily produce graphs of dynamic key strengths for various pieces of music taken as acoustic signals; however, its relation to perception is unclear. Izmirli and Bilgen made no attempt to validate the results produced by their system against perceptual data, only against music-theoretical analysis.

III. Approach

My approach to the research I propose centers upon the analysis of several hypotheses regarding the perception of music and the construction of music-analysis systems. Thus, I will organize this section as a series of assertions which are not yet generally accepted in the literature. By articulating these claims and the way I hope to address them, my general thinking and approach will become clear. I provide general summary and thoughts on the application of my research at the end of this section.

1. Much of music perception is pre-structural

This claim, on face value, is uncontroversial. Everyone understands that even if most of the interesting aspects of music perception are due to structural relationships, pre-structured "auditory features" of the sound signal must at some point be used. For example, musical pitch is typically considered as a fundamental feature of a sound; however, template models of pitch reduce this even further, to say that pitch arises from the interrelationship of more fundamental features, such as the frequencies of single sinusoidal partials.

I wish to argue the other direction: not only is pitch a fundamental feature, but other aspects of music typically considered as structurally based may be equally well-explained with a subsymbolic process model. For example, I have presented a processing model for musical tempo perception (Scheirer 1997; Scheirer 1998b) which is similar to theories of pitch-processing and does not require any musical knowledge. Rather than viewing tempo as a high-level attribute of music which results from the structural interaction of primitive events, this model views tempo itself as a primitive, holistic, feature of sound.

I hypothesize that similar analyses are possible for vertical harmony (chord quality), and aspects of tonal fusion and the perception of the resulting sound. If these hypotheses are correct, then listeners have a great deal of information available for classifying musical sounds at a very early stage of auditory cognition. Addressing this claim will be approached through direct demonstration: the construction of music-analysis systems which can analyze music in interesting, perceptually-relevant terms directly from the acoustic signal, without building elaborate grammatical or relational structures. I hope to build systems as successful at these tasks as the one I built for tempo and beat analysis.

This assertion is similar to complaints which have been lodged regarding the focus on symbolization in the artificial intelligence community at large. In particular, the writing of H. L. Dreyfus makes well-articulated arguments against over-symbolization and over-simplification in "microworlds" approaches to building AI systems (Dreyfus 1981 for example). In recent years, more research has taken place on building non-symbolic AI systems or hybrid symbol/signal systems (Brooks 1991); however, much of this work has not yet moved into the music-cognition and "music AI" literature. In my view, symbolization is a necessary step to understanding music cognition, just as it is in understanding human language, for language and music are both symbolic systems. However, it is important to use the *proper* symbolization, and to understand at a deep level what is done pre-symbolically and what is done post-symbolically.

Minsky has addressed my concern in a footnote to an article on music and meaning, in which he argues for a highly-constructed view of music understanding:

> What makes those thinkers who think that music does not make them do so much construction so sure that they know their minds so surely? It is ingenuous to think you "just react" to anything a culture works a thousand years to develop. A mind that thinks it works so simply must have more in its unconscious than it has in its philosophy.
> (Minsky 1989, p. 655 footnote 1)

This viewpoint argues against the empirical results on "musical surface listening" of Deliège *et al.* (1996); my response to it will be simply to *build* "cognitive artifacts" which demonstrate interesting musical functionality without sophisticated internal structuring. Such a construction will serve as a demonstration that at least internal construction is not *necessary* to perform certain musical tasks.

Note that in the writing of Leman and his colleagues (Leman 1995) the term "pre-symbolic" is used to refer to neural-network processing with perceptrons and Kohonen nets. When I use the term, though, I mean something more like digital-signal-processing, where the percept is represented by the continuous output of a (nonlinear) digital-filtering system. These uses are not mutually exclusive, of course. An article by Desain (1993) contrasts connectionist and "traditional AI" approaches to a single musical problem (beat-tracking) and develops a behavioral approach for comparing these sorts of models. Nearly any other sort of model could be subsumed by their approach as well.

This component of my research is the one which has the most long-term promise as a theoretical development. As described in Section 2, the vast majority of studies in music psychology have taken a structuralist viewpoint; that is, they view the perception of music as arising from the interrelation and juxtaposition of fundamental objects. While this model is undoubtedly a useful metaphor for thinking about certain phenomena (especially those of most interest to music theorists), it leaves many interesting questions about the perception of music unaddressed. For example, how does the mapping from acoustic signal to fundamental musical object happen? What are the fundamental perceptual objects in music-listening? How much preattentive and precognitive organization of musical input occurs? What aspects of music-listening must be learned specifically, and what aspects emerge from more basic organizational principles of the auditory system? I hope to begin addressing these important questions in my research.

2. "Notes" are not the fundamental perceptual entities of music

I have argued this claim in great detail elsewhere (Scheirer 1996), but I will recapitulate the main points here. Most models of music perception use what I call a "transcriptive" metaphor for music processing. That is, there is a clear hierarchy of data representations: the first stage is the sound signal, followed by some early organization such as harmonic partials or correlation structure, the next stage are the notes in the composition, and final cognition uses the notes to make structural judgments. Some authors make this assumption explicit (Piszczalski and Galler 1983; Longuet-Higgins 1994), while others (and I feel this is more dangerous) implicitly assume it, for example by describing "perceptual analysis" of written music notation (Lerdahl *et al.* 1983; Narmour 1990).

It is a trap set by centuries of musicology and music theory that we believe that symbolic models which use the score of the music as the starting point have a strong connection with perceptual music psychology. This is not to say that we can't learn anything from scores, but that theorists are making very strong assumptions about the early stages of perceptual organization when they assume the mid-level representation of music is "like" a score. These assumptions are as yet largely unjustified.

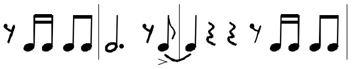
I think that a reconsideration of the fundamental organization of musical sound will lead us to find that so-called "tonal fusion" – that is, the perception of multiple physical sound sources as a single "auditory object" – plays a much larger role in the perception of music than it is normally assigned. Bregman has written about this hypothesis, using the term *chimera* to describe a tonally-fused object:

> Natural hearing tries to avoid chimeric percepts, but music often tries to create them. It may want the listener to accept the simultaneous roll of the drum, clash of the cymbal, and brief pulse of noise from the woodwinds as a single coherent event with its own striking emergent properties. The sound is chimeric in the sense that it does not belong to any single environmental object. (Bregman 1990, p. 459-460)

I hypothesize that these chimerae are, in fact, the most typical mid-level representation of musical sound, and that single-pitched auditory objects ("notes") are in fact only a special case of this more general object. I view a deeper understanding of the formation and perceptual properties of fused sounds, in opposition to a score-based viewpoint based on notes, as the strongest goal of my research.

This holds not only for vertical (harmonic/timbral) organization of music, but for horizontal (rhythmic) organization as well. To take one case for concreteness, in a paper on rhythm perception (Johnson-Laird 1991b) the following example was presented:

> The opening phrase of "Walkin'," a composition by Miles Davis, has the following rhythm:



> The first phrase ends, not with the first note in the second measure, but with the accented syncopation at the end of that measure.

There are many important assumptions buried in this innocuous-looking example. First, Davis' music comes from an oral tradition, not a written tradition. It is likely that this phrase was never written down until years after it had been invented. In addition, there is no "canonical" performance of this composition; in ten different recordings, the tempo and rhythm of this segment of the piece will be performed in ten different ways. Thus, to say that *this* is the rhythm of the piece is to make a choice about which version is the "most important."

Next, the Western-style notation obscures the fact that the actual durations and attack points are not spaced as notated here in an actual performance. Even for a performance where this notation would be used by an after-the-fact transcription to *represent* the rhythm to a jazz musician, the *actual* timing – due to swing phrasing and expressive playing – is potentially much different. Johnson-Laird, of course, is himself well aware of these distinctions, as evidenced by his other writing on jazz (Johnson-Laird 1991a). Using the same symbol (the "eighth note") to represent the third, fourth, ninth, and tenth notes of the phrase is an indication that, according to the symbolic theory in use, these notes are somehow "the same" or that something important is shared by all of them. This is not a conclusion, but an assumption, and one which has great ramifications for the analytic results which follow. Especially for music like jazz, rock, and the various other oral music traditions of the world, the acoustic signal is a much better starting point than a notation invented to serve an entirely different mode of thought.

I am not the first to present this argument against note-based music perception theories. Smoliar made it eloquently in a review of a book on "syntactic" music processing by Narmour (1990):

> The problem with a system like music notation is that it provides an *a priori* ontology of categories – along with labels for those categories – that does not necessarily pertain to categories that are actually formed as part of listening behavior. If we wish to consider listening to music as a cognitive behavior, we must begin by studying how categories are *formed* in the course of perception rather than trying to invent explanations to justify the recognition of categories we wish to assume are already present. (Smoliar 1991, p. 50)

Serafine made it as well in the introduction to her book on music cognition:

> Traditionally, the elements of music are assumed to be tones and assemblages of tones called chords. Such a view critically determines how we conceive composition and perception. For example, tones may be considered the material with which a composer works, by arranging and conglomerating them, or tones may be considered the basic units processed by the listener. The present view, however, holds that tones and chords cannot in any meaningful and especially psychological way be considered the elements of music. Rather, tones and chords are viewed as the inevitable by-product of musical writing and analysis, and as such are useful, even necessary analytic tools with minimal cognitive reality (Serafine 1988, p. 7)

The study of the true perceptual elements of music, as opposed to the study of the historical remnants of music theory, is the fundamental organizing principle of the research I propose. I hold firm to the position that the early stages of music perception have little to do with notes. My key goal is to move beyond theoretical argument for this position to practical demonstration that it allows a coherent and predictive theory of music perception to be constructed around it.

3. Polyphonic pitch-tracking is an engineering problem, not a scientific one

Much of the literature on polyphonic pitch-tracking and "automatic transcription" as described in Section II.C.2 conflates and confuses the goals of building systems to perform useful tasks with that of understanding the processes which underlie human sound perception. While they are both useful goals, and somewhat related, they are not the same: the first is an *engineering* goal, and the second a *scientific* goal. Several authors, in writing about engineering systems, allude to the goal of understanding human perception. For example, Piszczalski and Galler wrote:

> Exploring automatic music recognition may also help us understand how humans recognize sound patterns in general. The music notation associated with the played music represents an idealized, prototypical sound pattern that the performer tries to produce when playing a literal interpretation of the printed music.... The musically-sophisticated listener can then map the audible performance back into the idealized CMN [common music notation] form. In the recognition process, the listener somehow "internalizes" the perceived patten, considering it as an acceptable match of idealized patterns internally stored in the brain. (Piszczalski *et al.* 1983, p.405)

Surely, it "may" be the case that such an exploration will help understand human audition, but their sketch of the relationship between the score, the performer, the acoustics, the listener, and the mental representation is highly idiosyncratic and largely at odds both with intuition and with evidence from music psychology.

In my own research, I reject the usefulness of polyphonic pitch-tracking as a metaphor for human sound-understanding. Pre-symbolic computation, and processes which don't rely on a full transcriptional analysis of the music, are a much better model for the hearing processing. Smoliar made a similar point:

> When one considers a representation based directly on auditory stimuli, one is inevitably confronted with the question of *segmentation*—the assumption that, given any sort of audio input, one is obliged to parse it into notes before doing anything further with it. However, excessive attention to segmentation may ultimately be a distraction from more fundamental issues of behavior. There is no reason to assume that any fundamental ingredients based on immediate stimuli must be isomorphic to some [written] passage which can be clipped out of a score, so to speak. (Smoliar 1992, p. 591)

Acker and Pastore (1996) present experimental evidence in line with this view. In a test of musically-trained subjects using the "Garner paradigm" for testing integrality or separability of covariant discrimination, they found that subjects could not independently judge the changes in the pitch of the third of a major triad and changes in the pitch of the fifth of a major triad. That is, when the pitch of the fifth was altered, it affected the perception of the third, and vice versa. This indicates that these components are perceptually integral, not separately perceived and then combined to create a higher-level percept related to the chord quality. An experiment on musical duplex perception (Hall and Pastore 1992) showed that even subliminal—individually imperceptible—tones can be used to distinguish one chord from another.

Related to this point is the question of resynthesis. Many of the auditory scene analysis systems described in Section II.A.2 have been validated by resynthesizing extracted components and comparing them to the original sound or sounds. In my view, this is a false goal if we are actually studying human perception rather than simply attempting to build practical sound-separation systems. The human sound-understanding apparatus surely does not separate the sounds and *then* listen to each of them in term; rather, the perceptual segregation is illusory, highly dependent on expectation and attention. In many cases, we perceive sound which is not physically present as demonstrated by Warren's experiments on phonemic restoration (Warren 1970; Warren *et al.* 1972).

It would be silly to try to "resynthesize" a restored phoneme into a sound source, since the conditions leading to the perception are such complex cognitive ones. While it may be a useful test of a restoration model, this is *synthesis*, not *resynthesis*, since the "actual" phoneme was never present in the acoustic source (or, perhaps more accurately, there is no "actual" phoneme). The appropriate validatation for perceptual models is behavioral, not practical.

4. Auditory stream organization is crucial to music perception

When stated this way, this is not a new claim – Bregman (1990) devotes a chapter to exploring the role of auditory streaming in music perception and the theses of McAdams (1984) and Mellinger (1991) center on it. However, I view it in a slightly different way. In most discussions of auditory streaming in music, two aspects are discussed: the formation of harmonic overtones into notes, and the grouping of notes into streams. Too often, the notion of *auditory stream*, that of *musical voice*, and that of *homophonic line* are conflated; the examples of fugues and the way in which Bach used principles of auditory streaming to keep the voices separate are frequently used.

While it is true that auditory streaming applies to fugues, it's a very special case, in which stream, voice, and line are all the same thing. In the broader musical world, there are many cases in which multiple voices fuse to create a single line of polyphony, and the notion of auditory stream in such cases is much more subtle. If my hypothesis on tonally-fused musical objects proves a useful model, then the result of sequentially grouping these fused objects is more complex than what is typically considered. As multiple notes produced at the same time are perceived as a single object, so are multiple voices playing similar phrases perceived as a single homophonic line.

McAdams and Saariaho (1985) provided a similar analysis of perceptual entities in music:

> An entity, for us, is an auditory image whose constituent elements behave coherently as a group. This holds as much for the frequency components of a complex tone as for the different voices in homophonic orchestral writing. (McAdams *et al.* 1985, p. 367)

It is not at all clear what ability listeners have to track voices through polyphony. Huron (1989) found that even experienced listeners had a difficult time simply counting—a task that should be easier than following—more than three simultaneous voices concurrently. Tasks analyzing the perception of polyphony clearly relate to attentional control over listening; as such, I am not likely to explore them deeply in my research. A goal that I do have, though, is to attempt to articulate a theory of polyphony versus homophony in multiple-instrument textures; that is, to explain on a psychoacoustic basis why certain instruments are perceived "together" as a blended unit, and certain instruments, and groups of instruments, perceived as "separate."

There are no computational systems which have been presented in the literature that take this stance. I hope to describe source-grouping algorithms that lead to this sort of grouping naturally.

5. Music analysis systems must be perceptually motivated

The field of music signal processing has traditionally, as discussed in Section II, been one dominated by practical concerns: how do we find the pitch of a signal fast enough to use the results to control a synthesizer in real-time? How do we time-scale a signal without causing audible artifacts? How can we turn acoustic signals into symbolic representations? Many of these problems have proven amenable to approximate or engineering solution using more direct methods than full-scale perceptual modeling—Rabiner *et al.* (1976) gave an overview, for example, of different methods of approximating pitch.

However, music signal-processing systems which attempt to make broader judgments about sound will require a deeper grounding in the methods used by humans understanding sound. The reason is a fundamental one regarding the epistemology of music: the structural meaning of music is phenomenologically defined. To put this another way, there is no "ground truth" for pitch, or loudness, timbre, or rhythm, or harmonic implication, or other musical qualities of a sound signal, except what a listener hears in the signal. Pitch-tracking methods which are not based on models of pitch perception still work in many cases because there are other ways to extract representations similar to what people hear; however, when there is a conflict between the perception and a particular model, the model must be fixed to reflect the perception. This idea was explored in more detail by Martin, Scheirer, and Vercoe (Martin *et al.* 1998b).

The search for a model that could explain the extremely broad range of signals which give rise to pitch perception has been the primary impetus for many of the modern hearing models discussed in Section 2. Unfortunately, for most other perceptual qualities, we don't have nearly as many phenomena to explain, and this lack means that it is hard to construct falsifiable models. Bregman (1990, p. 122) discusses "metameric" timbres (hypothetical sounds with differing spectra that give rise to the same percept) and how useful it would be to discover and study them; but in the absence of such data, only hypotheses are possible.

I submit that it is possible and useful, however, to explore the application of existing psychoacoustic models in domains outside of their original conception. For example, I have discussed (Scheirer 1997) the application of a pitch-tracking model to the study of musical beat perception. In this study, I found that the auditory-computation structure currently hypothesized for pitch perception was also a good model for tempo perception, insofar as we can collect data explaining how tempo is perceived. The *result* of this research, though, was the construction of a robust, real-time beat tracking system (Scheirer 1998b) that operated in a wider musical domain than other such methods in the literature (Goto *et al.* 1998). Thus, by application of a hearing model to a different feature (a pitch model applied to the percept of tempo), I derived a pattern-recognition method for tempo analysis that seems to be superior to other solutions that originated from a strictly pattern-recognition approach.

To pursue this idea in more generality, I propose to look in detail at the application of temporal-correlation-based analysis models to complex musical sources. As discussed in Section 2, various methods of temporal correlation (the autocorrelogram, the weft, the strobed auditory image, the modulation spectrogram) are the predominant mid-level representations examined in hearing models today. Although it seems likely that some kind of periodicity detection is performed in the auditory system, it is still an open question which model of these (or others) is the best representation of the actual processing performed. It is also unknown what the "next" stages of processing are; the typical model is that some kind of auditory stream formation is performed, and "properties" or "features" such as pitch or loudness are associated with auditory objects.

The hypothesis I will pursue is that the proper method of extracting and analyzing periodic representations will lead to a feature set which allows for the natural comparison and classification of musical sounds. In particular, I want to investigate the representation of fused musical sounds and of perceptually segregated musical sounds in the correlogram. I further hypothesize that in order to pursue this, I will be able to make significant progress without conducting extensive psychoacoustic tests investigating human perception of fused and segregated sounds—that certain features will "fall out" naturally from the correct processing methods, and that they will allow interesting musical classification to be performed.

6. Music perception is not a logical process

When I say this, I don't mean that listening to music is "irrational" in the colloquial sense, or that some aspect of music perception is "noncomputational" (Kugel (1990) has presented an exposition of this viewpoint, with which I heartily disagree). Rather, I mean that the goal of a theory of music perception is not to arrive at a set of "fundamental propositions" which can then be interrelated using the machinery of formal logic.

Many theories of music-listening use formalisms drawn from mathematical logic to represent music perception and/or cognition. For example, Kunst (1978) uses the modal logic of Putnam and Kripke to represent "possible" and "necessary" conclusions about a piece during listening evolution. Deutsch (1981) uses a string-manipulation grammar to describe transformations of sequences of tones. The perceptual connections of these sorts of approaches are hazy at best.

I assert that there is an important distinction between the kind of mathematical-logical thought which is used in proving theorems or by some students in harmonizing four-part exercises, and the kind of “fluid” perceptual thinking which occurs in most real-world situations. Minsky writes:

> [T]he term [“logic”] originally referred to matters involving understanding and knowledge. But in current computer science “logical” now means mainly to try to express everything in terms of two quantifiers, viz., “for all X,” and “for some X,”—and two values, viz., “true” and “false.” But those menus seem just too small to me! We need much richer ideas than that ... I want AI researchers to appreciate that there is no one “best” way to represent knowledge. Each kind of problem requires appropriate types of thinking and reasoning—and appropriate kinds of representation. (Minsky *et al.* 1992, p. xi)

In places where formal methods are required (such as the construction of computer programs), the tools and techniques of pattern recognition are much more applicable to all kinds of perceptual systems, including music listening. In their ability to represent subtle aspects of multidimensional similarity and difference, nonlinear distances, and scale to a wide range of simple and complex problems, pattern analysis and classification techniques are the right ones to use in investigations of music.

7. The use of “real” musical stimuli is essential in music-analysis studies

Psychoacoustic studies have traditionally dealt with very simple stimuli. The “classical” method of inquiry into psychoacoustics involves extremely constrained sounds such as white-noise bursts and sine tones, and the analysis of perceptual qualities which result from listening to them, typically within the context of a narrowly-defined “decision criterion” paradigm (Durlach 1968). As discussed by Hartmann (1983), this methodology is essential if we are to properly understand the percept of fundamental sounds. However, the recent research on “top-down” processing, and binaural, cross-frequency, and cross-modal auditory phenomena calls into question the degree to which the perception of interesting ecological stimuli in auditory perception may be succinctly explained as a simple juxtaposition of underlying auditory objects.

Several of the projects described in Section II.D attempted to solve one specific problem in music theory: that of the determination of the root of a musical chord, or the tonality of a segment of music. This narrowness goes even as far as to develop whole papers around the examination of one *specific* chord “known” by music theories to be problematic (the so-called “Tristan” chord from Wagner). Such detailed analysis seems like an attempt at justifying the use of psychoacoustic models to music theorists; to show that a scientific approach can elucidate a problem which interests critics.

In contrast, many studies in music psychology (particularly “high-level” study of phenomena such as affective response to music) have taken a broader view of the experimental stimulus. Even musical research projects using simple stimuli often develop a broad stance within which the resulting theories can be scaled up to treat musical stimuli as whole

Very recent research in computational psychoacoustics (Leman 1994) has begun to examine processing models for more complex stimuli. However, there as yet have not been attempts to examine recent psychoacoustic models using a wide range of musical stimuli. I believe that to do so at this time is a possible and fruitful direction. I plan to conduct all validation experiments for the models I build using real musical data collected directly from commercial compact disc recordings.

The lack of ground truth for such musical examples (that is, for many samples we don’t have the score readily available) is not a problem in my research approach, since I am not trying to recover this information. Rather, I am trying to match human perception of the excerpts, and so validation data comes either from intuition (in straightforward judgments like genre) or psychoacoustic experiment (for similarity matching and other complex judgments).

Most research projects in studying music perception from a scientific view have focused on the classical tradition of Western music. This is natural, since most of the researchers are themselves most skilled in this genre. This cultural bias is sometimes defended as a beginning simplification; Smoliar has criticized this argument:

> One cannot simply say, "First we shall explain the cognition of Western tonal music, and then we shall explain everything else." All that "everything else" is going to require division, too; and unless we have some basic understanding of the discriminatory criteria we use to divide up into different subdomains, each of which may be "conquered" in isolation, we risk creating a situation in which those conquered pieces will not fit together for want of a clear understanding of how they were taken apart in the first place. (Smoliar 1995, p. 25)

If we want to build music systems of use to everyday listeners (not just music theorists), it is crucial not to take the view that other forms of music are a curiosity or a "simplified" case. Jazz, rock, blues, folk, and what are today called "world" musics should be considered as primary cases alongside common-practice Western music[3]. As my project attempts only to explain the auditory process as applied to music, any findings should be pre-cultural to begin with; my goal is to explain the *early* stages of perception and bring the model to a point where acculturation can begin. Unless the role of expectation is very strong and important at this early stage, the musical background of the listener should matter little.

8. A standard music database is a necessary addition to the field

For previous research projects in musical-signal-processing (Scheirer *et al.* 1997; Scheirer 1998b), I collected a set of 120 short musical excerpts from a radio tuner. Having this data has proven very valuable for a wide range of uses, from validating performance to testing hypotheses. In addition, using the same data over a period of more than a year has allowed me to become very familiar with the properties of the various excerpts and utilize them effectively. However, this database is not completely adequate for a research project larger in scope. It is impoverished with regard to certain musical genres, there is not enough data overall, and (especially) I do not know the exact composers, performers, or even title of many of the excerpts. I propose to collect a larger database, have it better documented, and (if I can get copyright clearance through fair use guidelines, or through donation from Media Lab sponsors) release it to the larger music-processing world as a standard database of acoustic musical signals for computational music psychology.

The existence of standard databases of images, textures, and video sequences in the image-processing and speech-processing fields has proven a great help to researchers trying to compare results and cross-validate their ideas. To do the same for music signal processing will similarly be a boon to researchers in musical acoustics.

Some of the desirable properties of such a database are: it should be extensively documented both with copyright-style information, such as composer, title, and catalog number, and as much humanistic information, such as genre, style, and quality, as possible (this is a labor-intensive process). It should be very broad with regard to musical style, including non-Western music, popular music such as rock, music with and without vocals, music which is only vocal, music with elec-

[3] It is of course a misnomer to categorize all such music "together." The musics of different cultures of the world differ as much from each other as each does from Western music. The point is only to consider such diversity when conducting music research.

tronic instruments, music with heavy drums, and music with a poor signal-to-noise quality (for example, with heavy audience noise or reverberation). It should contain music that is well-known and music that is not well-known. It should be publically distributable.

If data is recorded monophonically, at 44 KHz sampling rate about 2 hours of sound fits on a CD-ROM. This corresponds to 240 examples of 30 seconds each; 1000 examples will easily fit on a set of four CDs. This should be enough data for many researchers to make good use.

9. The perceptions of "naive" listeners are important

Many studies in the psychology of music use only expert musicians as subjects. The rationale for such a restriction is that only expert musicians have the skills required (analytic hearing) to decode the intentions of the composer, and thus to "understand" the music. However, as noted by Smith (Smith 1997), this has the danger of causing music psychology to become entirely self-referential, in which experiments conducted by experts on experts serve to confirm the expectations of experts. Too often, Smith writes, "These [expert] subjects can predict the character of the answers wished and can provide those answers if they choose." (p. 238)

As discussed in Section II.B.6, non-musicians seem not to share the structural facilities of music perception with experts. However, when music perception is considered in a broader sense, it is clear that non-musician "naive" listeners also participate a great deal. As discussed in the Introduction, many of the instantaneous judgments of music, such as beat and genre, are shared by musicians and by non-musicians. If we wish to study the fundamental organization of music, it may be more appropriate to also consider listeners who do not apply cognitive analysis to the listening process, as the fundamental aspects of organization may then become clearer.

It is possible that musical novices have a stronger connection between "musical surface" cues such as tempo, texture, loudness, and height, and emotional reactions. Smith (1987) compared the musical preferences of experts and novices, and found that novices value composers judged (by experts) as "sensual" and "emotional," while experts overwhelmingly value composers judged as "syntactic."

In my modeling approach, and in any psychoacoustic experiments I conduct, I will be primarily interested in understanding the perceptions of untrained listeners.

10. General Discussion and Applications

As discussed in the preceding sections, the approach underlying my research is one of *unification*. I want to examine signal-processing implementations of techniques arising from the psychoacoustic literature, and demonstrate how they can be used to build better music analysis systems. The resulting systems will be embodiments of a scientific theory of music perception which will be articulated in my dissertation; they also have practical application for use in multimedia systems.

The applications for high-quality music perception systems are numerous; several are articulated by Martin *et al.* (1998b) and Vercoe *et al.* (1998). These include:

- *Access to musical databases.* Using robust automated music perception, systems could be built to augment Internet-based music recommendation services with the ability to make musical judgments. Currently, such systems work through collaborative filtering, which is a powerful but limited technique – for example, it is difficult to make recommendations about music that few people have listened to. This is the one application I would like to demonstrate and hope to build.

- *Synthetic performance systems.* With the ability to hear music, systems for enabling musical collaboration between humans and machines could become more expressive and capable.

- *Algorithmic composition.* A current limitation on computer-composition artificial intelligence systems is that their output is evaluated by humans. By enabling machines to evaluate and critique their own compositions, a more natural (and more truly automated) automatic composition process is enabled.

- *Structured audio encoding.* A theory of musical listening could shed light on structural regularities in music that could be used in structured audio encoding. This is not only in the simple case of music-transcription for generating note lists, but in more subtle ways such as identifying perceptually equivalent sounds for coding. As loudness and masking models led to the development of low-bit rate perceptual coders (Jayant, Johnston and Safranek 1993), so can psychoacoustic music processing shed light on new techniques for audio coding and compression.

- *Visual music displays.* Real-time computer graphics routines (or physical devices such as those employed by TechnoFrolics systems) can generate compelling multimedia experiences when synchronized with music. By enabling real-time automated music listening as a component of such a system, the display can be made to work immediately with any music desired.

IV. Scope

This section describes and outlines the particular activities I plan to pursue in the course of my dissertation research. Of necessity, it is somewhat speculative; the particular activities to be pursued and their order will depend on the results of the early experiments. A short description of deliverables concludes the section.

A. Modelling musical object formation

In this component, I will construct a model of the separation and formation of fundamental objects in music perception. This component differs from other projects intending to accomplish the same thing (Mellinger 1991, for example) in its broader conception of “perceptual object.” In particular, I will not try to transcribe and separate tonally-fused musical objects (notes in a chord), but rather, I will treat fused timbres as singular objects.

The goal and idea is to build this system on top of a correlogram representation, or another one (Patterson’s SAI model, or the modulation spectrogram) that provides equivalent information. There have been few results reported on pattern-recognition-based modelling of correlogram “motion.” I hope to arrive at a set of heuristics and procedures which can track multiple moving sources through the correlogram representation, and thereby understand which components of the correlogram must be grouped together. The output of this grouping process will be taken as a model of fundamental music object formation.

B. Examining musical object properties

Since my view of musical object is more complex than that used in a transcriptive model, a more elaborate model of musical object properties is required. In a transcription model, objects (notes) have properties such as pitch, loudness, and timbre; if multiple sources with different pitches fuse into a single object, the properties of the object must represent more difficult-to-define considerations such as harmonic implication and overall “texture.”

This work will begin with an investigation of chord quality within a correlogram representation. As discussed above, there is a clear notion in music psychology and music theory of the perceptual equivalence of various inversions and orchestrations of the same chord. I would like to examine potential perceptual analogues to this equivalence within the correlogram, by trying to build a small system which can classify individual chords by quality without regard to timbre or inversion. Ideally, if this model is successfully constructed, then it will also apply largely without modification to the objects which were formed by the model discussed above.

A second direction requiring study is the investigation of remaining perceptual properties of fused objects. I use the term "texture" to represent the analogue in fused objects to what the term "timbre" represents for single-source notes. However, I do not yet have a clear idea of how to study this, and it may have to remain outside the scope of the thesis. A recent idea that has appealed to me is the decomposition of sounds into source-filter models. This was the approach taken by both Ellis (1996a) and Casey (1998) in their recent dissertations at the Media Lab, and was also pursued by Martin in his timbre work (Martin 1998) and Slaney (1996) in his work on "audio morphing." It is plausible, but requires more evaluation, that the "chord quality" component of a harmonic object might correspond to certain properties of the sound source in a source-filter model, and that the "texture" component to properties of the filter function.

C. Automatic classification of music

As discussed above, the first step in experimenting with the classification of music using a model of object separation and object properties is collection of a sufficiently thorough and documented database of musical excerpts for use in the experiments. This activity can proceed in parallel with the technical work described above, and will likely make use of several undergraduate assistants hired for the purpose of helping to select, document, and organize these musical examples.

The actual experiments in music classification and understanding will make use of the models developed in the above activities, and the database of test excerpts, to investigate whether these models are adequate for performing genre categorization, similarity measurement, and other solutions to practical problems in multimedia content retrieval. Tools from the pattern classification literature will be applied in these experiments; I do not currently foresee the development of new classification methods.

D. Psychoacoustic justification

If time permits, it would be a valuable step to provide experimental evidence in defense of the psychoacoustic claims made by the models I develop. However, robust and careful experiments are a difficult and long-term proposition, and I do not foresee that it will be possible to do as much as might be desirable. I can identify a few experiments which would shed light both on the technical direction of the object formation and feature analysis process, and on validation methodology for the model as a whole.

1. Responses of non-musicians to short musical excerpts

As I described in the Introduction, a "toy problem" which represents my proposed work is the modelling of the percepual response of non-musicians to short excerpts. A qualitative experiment on this problem would be useful as a possible ground truth. Broadly described, the design for such an experiment might simply ask subjects to "describe what they hear" in such excerpts. With appropriate guidance, it might be possible to learn from this the kinds of representations which non-musicians maintain during music listening. Other possible tasks include music similarity judgment (which could lead to a multidimensional-scaling analysis of music types) or others of the "surface music perception" tasks described in the Introduction.

2. Fundamental object formation

In a previous paper (Scheirer 1996), I discussed a few sound examples which I claim as evidence for the notion that chimeric objects, or tonally-fused music objects, are the fundamental units of music perception. To fully develop this argument requires conducting psychoacoustic experiments with these sorts of stimuli. The general paradigm I propose is to place harmonic perception in conflict with auditory stream formation in a standard “stream capture” experiment. I hypothesize that immediate harmonic perception is only possible when sources giving rise to such a percept are grouped into the same auditory stream. Thus, when key sound sources disambiguating harmonic sounds are captured by competing streams, the harmonic implications of the now-ambiguous sounds will not be able to be perceived.

Such a result would provide evidence that harmonic perception occurs only within streams, not across streams, and thus cannot be explained as the symbolic interaction of inter-stream objects.

3. Similarity of tonally-fused objects

There are few reports in the literature on the perception of musical objects in tonal fusion. Most such studies have been principally concerned with *when* tonal fusion occurs in musical contexts, not the result of such fusion. In order to validate the model of object properties, an experiment investigating the nature of harmonic quality of different objects, and perhaps another investigating textural properties, would be very useful. I don’t know whether similarity-matching in a multidimensional scaling framework, or some other methodology, would be most appropriate here.

E. Deliverables

The primary deliverables for this research will be the code that implements the models I develop of music perception, and the dissertation that articulates the science, philosophy, and operation of the models. If time permits, I also plan to develop a prototype music-searching interface that allows a database of music to be searched using perceptual categories such as genre, complexity, tempo, and artist, and allows for music similarity matching demonstrations.

Such an interface will make a valuable demonstration to Media Lab sponsors as well as an example of applying the research.

V. Practicalia

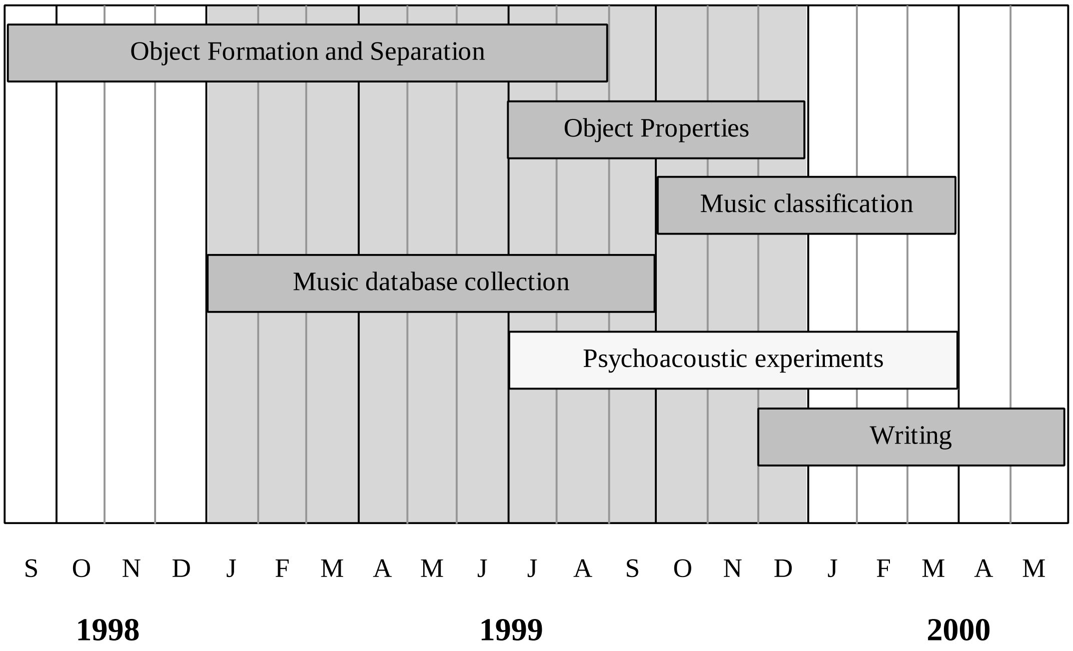
This section discusses the schedule and resources for the research proposed in previous sections.

A. Schedule

I propose to complete the research described in Section IV in the next 18 months, and present and defend my dissertation by May 2000. The following chart shows a tentative schedule for work progress on the dissertation. The activities listed are those discussed in the previous scope. Many of the activities may be pursued in parallel or with some overlap.

In general, study of the two major technical contributions (**Object Formation and Separation** and **Musical Object Properties**) precedes the other activities. I have allocated one year for the former and six months for the latter, with a three-month overlap. After significant headway has made made in these tasks, I will begin to examine music-classification systems using these results with known methods of discrimination, categorization, and classification. The collection of a database for this task will be performed in parallel with the earlier scientific work.

If time permits, I will also attempt to conduct psychoacoustic experiments in parallel with the music classification studies in order to evaluate key results from the modelling research. I have allocated ten months for this task, partially overlapping with the six months allocated for writing the dissertation.



B. Resources

Here is a list of resources required to carry out this work:

1. **Computer facilities**

   My current machine, an SGI Octane, should be quite adequate for this purpose.

2. **Disk space allocation**

   Research on correlogram structures, and the collection of a large music database, are very storage-intensive activities. At least 5 GB of disk space, shared between local and networked resources, should be made available.

3. **Undergraduate assistance**

   Both the collection of a music database and the execution of any psychoacoustic experiments could make use of undergraduate assistants hired through the UROP program. I request funding for hiring one UROP, on average, over the whole research period. I am not presently sure what the allocation of this funding will be (I might hire two for the database collection activity), but not more than two UROP-years in total will be required.

4. **Audio equipment**

   Headphones, loudspeakers, amplification, etc are required. Current and future Machine Listening Group resources should be adequate for this requirement.

VI. Conclusion

In this proposal, I have reviewed literature pertaining to the fields of music psychology, psychoacoustics, and music signal processing. I have discussed the relation of each of these fields to my proposed dissertation research, which examines the psychoacoustic bases for the perception of music. I have articulated my own approach to these problems, the scope of research I plan to undertake, and the resources this research will require.

A dissertation that solves the problems I propose to examine here will be a major contribution to the literature on music perception, psychoacoustics, and music signal processing. It will provide a point of reconciliation between the first two of this fields, showing how music perception depends on and is mediated by the psychoacoustic faculties of human listeners. It will also point a new direction in the construction of music-listening computer systems; such systems will be of great practical use for multimedia tasks such as annotating and searching musical databases.

Presenting a theory of perception as a computer-implemented set of algorithms, and evaluating it through the presentation of ecological stimuli to the algorithms, is a technique that has only recently become available to perception researchers. As other research at MIT, the Media Laboratory, and elsewhere has demonstrated, it is a powerful addition to the set of analytic tools at the disposal of a scientist. I am firmly committed to presenting my results as part of the scientific discourse on the relevant topics, not only to constructing demonstration and prototype systems.

My work fits well into the larger context of Machine Listening and Perceptual Computing research at the Media Lab. Like many other projects in these areas, it seeks to synthesize results from disparate disciplines in creating a unique theory of Media as distinct from the component parts. Following my dissertation, I hope to continue exploring the connections between musical systems, musical signal processing, musical multimedia, and musical perception.

Acknowledgements

The support and assistance of the following people is gratefully acknowledged: Prof. Barry Vercoe, for supervising the research; Prof. Rosalind Picard, Prof. Perry Cook, and Dr. Malcolm Slaney, for agreeing to serve as readers; the current and past members of the Machine Listening Group, especially Keith Martin, Dr. Dan Ellis, and Dr. Bill Gardner, for continuous discussion, argument, support, and criticism; Prof. Carol Krumhansl, for her support and supervision when I was her undergraduate assistant; and especially my wife Jocelyn, for her love and unconditional emotional and intellectual support.

References

Acker, B. E. and R. E. Pastore (1996). “Perceptual integrality of major chord components.” *Perception & Psychophysics* **58**(5): 748-761.

Allen, J. B. (1996). “Harvey Fletcher's role in the creation of communication acoustics.” *Journal of the Acoustical Society of America* **99**(4): 1825-1839.

Bamberger, J. (1991). *The Mind Behind the Musical Ear: How Children Develop Musical Intelligence*. Cambridge, MA: Harvard University Press.

Belkin, A. (1988). “Orchestration, perception, and musical time: A composer's view.” *Computer Music Journal* **12**(2): 47-53.

Bigand, E., R. Parncutt and F. Lerdahl (1996). “Perception of musical tension in short chord sequences: The influence of harmonic function, sensory dissonance, horizonal motion, and musical training.” *Perception & Psychophysics* **58**(1): 125-141.

Bregman, A. (1990). *Auditory Scene Analysis*. Cambridge MA: MIT Press.

Brooks, R. A. (1991). "Intelligence without reason." MIT Artificial Intelligence Lab Technical Report #1293, Cambridge MA.

Brown, G. J. and M. Cooke (1994). “Computational auditory scene analysis.” *Computer Speech and Language* **8**(2): 297-336.

Brown, G. J. and D. Wang (1997). “Modelling the perceptual segregation of double vowels with a network of neural oscillators.” *Neural Networks* **10**(9): 1547-1558.

Brown, J. C. (1991). “Calculation of a constant *Q* spectral transform.” *Journal of the Acoustical Society of America* **89**(1): 425-434.

Brown, J. C. and M. S. Puckette (1992). “An efficient algorithm for the calculation of a constant *Q* transform.” *Journal of the Acoustical Society of America* **92**(5): 2698-2701.

Cariani, P. A. (1996). "Temporal coding of musical form." In *Proc. International Conference on Music Perception and Cognition*, Montreal.

Cariani, P. A. and B. Delgutte (1996). “Neural correlates of the pitch of complex tones. I. Pitch and pitch salience.” *Journal of Neurophysiology* **76**(3): 1698-1734.

Carter, N. P., R. A. Bacon and T. Messenger (1988). “The acquisition, representation and reconstruction of printed music by computer: A review.” *Computers and the Humanities* **22**(2): 117-136.

Casey, M. A. (1998). *Auditory Group Theory with Applications to Statistical Basis Methods for Structured Audio.* Ph.D. thesis, MIT Media Laboratory, Cambridge MA.

Chafe, C. and D. Jaffe (1986). "Source separation and note identification in polyphonic music." Stanford University Technical Report #STAN-M-34, Palo Alto, CA.

Chafe, C., D. Jaffe, K. Kashima, B. Mont-Reynaud and J. Smith (1985). "Techniques for note identification in polyphonic music." In *Proc. ICMC*, Tokyo.

Chafe, C., B. Mont-Reynaud and L. Rush (1982). “Toward an intelligent editor of digital audio: recognition of musical constructs.” *Computer Music Journal* **6**(1): 30-41.

Chowning, J. M. (1990). "Music from machines: Perceptual fusion and auditory perspective - for Ligeti." Stanford University Technical Report #STAN-M-64, Palo Alto, CA.

Clarke, E. F. and C. L. Krumhansl (1990). “Perceiving musical time.” *Music Perception* **7**(3): 213-252.

Clynes, M. (1995). “Microstructural musical linguistics: composers' pulses are liked most by the best musicians.” *Cognition* **55**: 269-310.

Cook, N. (1990). *Music, Imagination, and Culture*. Oxford: Clarendon Press.

Cope, D. (1991). *Computers and Musical Style*. Madison, WI: A-R Editions.

Cope, D. (1992). “Computer modeling of musical intelligence in EMI.” *Computer Music Journal* **16**(2): 69-83.

Crowder, R. G. (1985a). “Perception of the major/minor distinction: I. Historical and theoretical foundations.” *Psychomusicology* **4**(1): 3-12.

Crowder, R. G. (1985b). “Perception of the major/minor distinction: II. Experimental investigations.” *Psychomusicology* **5**(1): 3-24.

Cumming, D. (1988). *Parallel algorithms for polyphonic pitch tracking*. M.S. thesis, MIT Media Laboratory and Dept. of Electrical Engineering, Cambridge MA.

Dannenberg, R. B., B. Thom and D. Watson (1997). "A machine learning approach to musical style recognition." In *Proc. ICMC*, Thessaloniki GR.

de Cheveigné, A. (1993). “Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing.” *Journal of the Acoustical Society of America* **93**(6): 3271-3290.

de Cheveigné, A. (1998). “Cancellation model of pitch perception.” *Journal of the Acoustical Society of America* **103**(3): 1261-1271.

De Poli, G., A. Piccialli and C. Roads, Eds. (1991). *Representations of Musical Signals*. Cambridge MA, MIT Press.

Delgutte, B., B. M. Hammond, S. Kalluri, L. M. Litvak and P. A. Cariani (1997). "Neural encoding of temporal envelope and temporal interactions in speech." In *Proc. XIth International Conference on Hearing*, Grantham UK.

Deliège, I., M. Melen, D. Stammers and I. Cross (1996). “Musical schemata in real-time listening to a piece of music.” *Music Perception* **14**(2): 117-160.

Desain, P. (1993). “A connectionist and a traditional AI quantizer: symbolic versus sub-symbolic models of rhythm perception.” *Contemporary Music Review* **9**(1&2): 239-254.

Desain, P. and H. Honing (1994). "Can music cognition benefit from computer music research? From foot-tapper systems to beat induction models." In *Proc. ICMPC*, Liege BE.

Desain, P. and H. Honing (1995). "Computational models of beat induction: The rule-based approach." In *Proc. IJCAI-95*, Montreal.

Deutsch, D. (1981). “The internal representation of pitch sequences in tonal music.” *Psychological Review* **88**(6): 503-522.

Dowling, W. J. and D. L. Harwood (1986). *Music Cognition.* San Diego: Academic Press.

Dreyfus, H. L. (1981). "From micro-worlds to knowledge representation: AI at an impasse." In *Mind Design*, J. Haugeland, ed. Cambridge MA: MIT Press**:** 161-204.

Duda, R. O., R. F. Lyon and M. Slaney (1990). "Correlograms and the separation of sounds." In *Proc. IEEE Asilomar Workshop*, Asilomar CA.

Durlach, N. I. (1968). "A decision model for psychophysics." MIT Research Laboratory of Electronics Technical Report #1, Cambridge MA.

Ellis, D. P. W. (1994). "A computer implementation of psychoacoustic grouping rules." In *Proc. 12th ICPR*, Jerusalem.

Ellis, D. P. W. (1996a). *Prediction-Driven Computational Auditory Scene Analysis*. Ph.D. thesis, MIT Dept. of Electrical Engineering and Computer Science, Cambridge MA.

Ellis, D. P. W. (1996b). "Prediction-driven computational auditory scene analysis for dense sound mixtures." In *Proc. ESCA workshop on the Auditory Basis of Speech Perception*, Keele UK.

Ellis, D. P. W. (1997). "The weft: A representation for periodic sounds." In *Proc. ICASSP*, Munich.

Ellis, D. P. W. and D. F. Rosenthal (1998). "Mid-level representations for computational auditory scene analysis: The weft element." In *Readings in Computational Auditory Scene Analysis*, D. F. Rosenthal and H. G. Okuno, ed. Mahweh NJ: Lawrence Erlbaum**:** 257-272.

Erickson, R. (1985). *Sound Structure in Music*. Berkeley, CA: University of California Press.

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

Foote, J. (in press). "An overview of audio information retrieval." *ACM Multimedia Systems Journal*.

Foster, S., W. A. Schloss and A. J. Rockmore (1982). "Toward an intelligent editor of digital audio: Signal processing methods." *Computer Music Journal* **6**(1): 42-51.

Gabor, D. (1947). "Acoustical quanta and the theory of hearing." *Nature* **159**: 591-594.

Goldstein, J. L. (1973). "An optimum processor theory for the central formation of the pitch of complex tones." *Journal of the Acoustical Society of America* **54**(6): 1496-1516.

Gordon, J. W. (1987). "The role of psychoacoustics in computer music." Stanford University Technical Report #STAN-M-38, Palo Alto CA.

Goto, M. and Y. Muraoka (1997). "Real-time rhythm tracking for drumless audio signals -- chord change detection for musical decisions." In *Proc. IJCAI-97 Workshop on Computational Auditory Scene Analysis*, Tokyo.

Goto, M. and Y. Muraoka (1998). "Music understanding at the beat level: Real-time beat tracking for audio signals." In *Readings in Computational Auditory Scene Analysis*, D. F. Rosenthal and H. Okuno, ed. Mahweh, NJ: Lawrence Erlbaum.

Hall, M. D. and R. E. Pastore (1992). "Musical duplex perception: Perception of figurally good chords with subliminal distinguishing tones." *JEP:HPP* **18**(3): 752-762.

Hartmann, W. (1983). "Electronic music: A bridge between psychoacoustics and music." In *Music, Mind, and Brain: The Neuropsychology of Music*, M. Clynes, ed. New York: Plenum Press**:** 371-385.

Hartmann, W. M. (1996). “Pitch, periodicity, and auditory organization.” *Journal of the Acoustical Society of America* **100**(6): 3491-3502.

Hawley, M. J. (1993). *Structure out of Sound.* Ph.D. thesis, MIT Media Laboratory, Cambridge MA.

Hermes, D. J. (1988). “Measurement of pitch by subharmonic summation.” *Journal of the Acoustical Society of America* **83**(1): 257-264.

Holleran, S., M. R. Jones and D. Butler (1995). “Perceiving musical harmony: The influence of melodic and harmonic context.” *JEP:LMC* **21**(3): 737-753.

Huron, D. (1989). “Voice denumerability in polyphonic music on homogeneous timbres.” *Music Perception* **6**(4): 361-382.

Huron, D. (1991). “Tonal Consonance versus tonal fusion in polyphonic sonorities.” *Music Perception* **9**(2): 135-154.

Huron, D. and P. Sellmer (1992). “Critical bands and the spelling of vertical sonorities.” *Music Perception* **10**(2): 129-150.

Irino, T. and R. D. Patterson (1996). “Temporal asymmetry in the auditory system.” *Journal of the Acoustical Society of America* **99**(4): 2316-2331.

Izmirli, Ö. and S. Bilgen (1996). “A model for tonal context time course calculation from acoustical input.” *Journal of New Music Research* **25**(3): 276-288.

Jayant, N., J. Johnston and R. Safranek (1993). “Signal Compression based on models of human perception.” *Proc IEEE* **81**(10): 1385-1422.

Johnson-Laird, P. N. (1991a). "Jazz improvisation: A theory at the computational level." In *Representing musical structure*, P. Howell, R. West and I. Cross, ed. London: Academic Press.

Johnson-Laird, P. N. (1991b). “Rhythm and meter: A theory at the computational level.” *Psychomusicology* **10**: 88-106.

Jones, M. R. and M. Boltz (1989). “Dynamic attending and responses to time.” *Psychological Review* **96**(3): 459-491.

Juslin, P. N. (1997). “Emotional communication in music performance: A functionalist perspective and some data.” *Music Perception* **14**(4): 383-418.

Kashino, K. and H. Murase (1997). "Sound source identification for ensemble music based on the music stream extraction." In *Proc. Int. Joint Conf. on AI Workshop on Computational Auditory Scene Analysis*, Tokyo.

Kashino, K., K. Nakadai, T. Kinoshita and H. Tanaka (1995). "Application of Bayesian probability network to music scene analysis." In *Proc. Int. Joint Conf. on AI Workshop on Computational Auditory Scene Analysis*, Montreal.

Kashino, K. and H. Tanaka (1992). “A sound source separation system using spectral features integrated by the Dempster's law of combination.” *Annual Report of the Engineering Research Institute, University of Tokyo* **51**: 67-72.

Kashino, K. and H. Tanaka (1993). "A source source separation system with the ability of automatic tone modeling." In *Proc. ICMC*, Tokyo.

Kastner, M. P. and R. G. Crowder (1990). “Perception of the major/minor distinction: IV. Emotional connotations in young children.” *Music Perception* **8**(2): 189-202.

Katayose, H. and S. Inokuchi (1989). “The Kansei music system.” *Computer Music Journal* **13**(4): 72-77.

Klassner, F. I. (1996). *Data Reprocessing in Signal Understanding Systems*. Ph.D. thesis, University of Massachusetts Computer Science, Amherst, MA.

Klassner, F. I., V. Lesser and S. H. Nawab (1998). "The IPUS blackboard architecture as a framework for computational auditory scene analysis." In *Readings in Computational Auditory Scene Analysis*, D. F. Rosenthal and H. Okuno, ed. Mahweh, NJ: Erlbaum**:** 177-193.

Kronland-Martinet, R. and A. Grossman (1991). "Application of time-frequency and time-scale methods (wavelet transforms) to the analysis, synthesis, and transformation of natural sounds." In *Representations of Musical Signals*, G. De Poli, A. Piccialli and C. Roads, ed. Cambridge MA: MIT Press**:** 45-85.

Krumhansl, C. L. (1979). “The psychological representation of musical pitch in a tonal context.” *Cognitive Psychology* **11**: 346-374.

Krumhansl, C. L. (1990). *Cognitive Foundations of Musical Pitch*. Cambridge UK: Cambridge University Press.

Krumhansl, C. L. (1991a). “Memory for musical surface.” *Memory & Cognition* **19**(4): 401-411.

Krumhansl, C. L. (1991b). “Music psychology: tonal structures in perception and memory.” *Annual Review of Psychology* **42**: 277-303.

Krumhansl, C. L. (1997). "Effects of perceptual organization and musical form on melodic expectancies." In *Music, Gestalt, and Computing: Studies in Systematic and Cognitive Musicology*, M. Leman, ed. Berlin: Springer. **1317:** 294-320.

Krumhansl, C. L. and E. J. Kessler (1982). “Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys.” *Psychological Review* **89**(4): 334-368.

Kugel, P. (1990). “Myhill's thesis: There's more than computing in musical thinking.” *Computer Music Journal* **14**(3): 12-25.

Kuhn, W. B. (1990). “A real-time pitch recognition algorithm for music applications.” *Computer Music Journal* **14**(3): 60-71.

Kunst, J. (1978). *Making sense in music: an enquiry into the formal pragmatics of art*. Ghent BE: Communication and Cognition.

Leman, M. (1989). “Symbolic and subsymbolic information processing in models of musical communication and cognition.” *Interface* **18**: 141-160.

Leman, M. (1994). “Schema-based tone center recognition of musical signals.” *Journal of New Music Research* **23**(2): 169-204.

Leman, M. (1995). *Music and Schema Theory*. Berlin: Springer-Verlag.

Lerdahl, F. and R. Jackendoff (1983). *A Generative Theory of Tonal Music*. Cambridge MA: MIT Press.

Levitin, D. J. (1994). “Absolute memory for musical pitch: Evidence from the production of learned melodies.” *Perception & Psychophysics* **56**(4): 414-423.

Levitin, D. J. and P. R. Cook (1996). “Memory for musical tempo: Additional evidence that auditory memory is absolute.” *Perception & Psychophysics* **58**(6): 927-935.

Lewin, D. (1986). “Music theory, phenomenology, and modes of perception.” *Music Perception* **3**(4): 327-392.

Licklider, J. C. R. (1951a). "Basic correlates of the auditory stimulus." In *Handbook of Experimental Psychology,* S. S. Stevens, ed. New York: Wiley.

Licklider, J. C. R. (1951b). “A duplex theory of pitch perception.” *Experientia* **7**: 128-134.

Longuet-Higgins, H. C. (1994). “Artificial intelligence and musical cognition.” *Philosophical Transactions of the Royal Society of London (A)* **349**: 103-113.

Maher, R. C. (1990). “Evaluation of a method for separating digitized duet signals.” *JAES* **38**(12): 956-979.

Makhoul, J. (1975). “Linear prediction: A tutorial review.” *Proc IEEE* **63**(4): 561-580.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* New York: W.H. Freeman & Co.

Martin, K. D. (1996a). "Automatic transcription of simple polyphonic music: Robust front-end processing." MIT Media Laboratory Perceptual Computing Technical Report #399, Cambridge MA.

Martin, K. D. (1996b). "A blackboard system for automatic transcription of simple polyphonic music." MIT Media Laboratory Perceptual Computing Technical Report #385, Cambridge MA.

Martin, K. D. (1998). "Toward automatic sound source recognition: identifying musical instruments." In *Proc. NATO Computational Hearing Advanced Study Institute,* Il Ciocco IT.

Martin, K. D. and Y. E. Kim (1998a). "Musical instrument identification: A pattern-recognition approach." In *Proc. 136th Meeting of the Acoustical Society of America,* Norfolk, VA.

Martin, K. D., E. D. Scheirer and B. L. Vercoe (1998b). "Musical content analysis through models of audition." In *Proc. ACM Multimedia Workshop on Content-Based Processing of Music,* Bristol UK.

McAdams, S. (1983). "Spectral fusion and the creation of auditory images." In *Music, Mind, and Brain: The Neuropsychology of Music,* M. Clynes, ed. New York: Plenum Press**:** 279-298.

McAdams, S. (1984). *Spectral Fusion, Spectral Parsing, and the Formation of Auditory Images*. Ph.D. thesis, Stanford University CCRMA, Dept of Music, Stanford, CA.

McAdams, S. (1987). “Music: A science of the mind?” *Contemporary Music Review* **2**: 1-61.

McAdams, S. (1989). “Segregation of concurrent sounds. I: Effects of frequency modulation coherence.” *Journal of the Acoustical Society of America* **86**(6): 2148-2159.

McAdams, S., M.-C. Botte and C. Drake (1998). “Auditory continuity and loudness computation.” *Journal of the Acoustical Society of America* **103**(3): 1580-1591.

McAdams, S. and K. Saariaho (1985). "Qualities and functions of musical timbre." In *Proc. ICMC,* Burnaby BC, CA.

McAulay, R. J. and T. F. Quatieri (1986). “Speech analysis/synthesis based on a sinusoidal representation.” *IEEE ASSP* **34**(4): 744-754.

McCabe, S. L. and M. J. Denham (1997). “A model of auditory streaming.” *Journal of the Acoustical Society of America* **101**(3): 1611-1621.

Meddis, R. and M. J. Hewitt (1991a). “Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification.” *Journal of the Acoustical Society of America* **89**(6): 2866-2882.

Meddis, R. and M. J. Hewitt (1991b). “Virtual pitch and phase sensitivity of a computer model of the auditory periphery. II: Phase sensitivity.” *Journal of the Acoustical Society of America* **89**(6): 2883-2894.

Meddis, R. and M. J. Hewitt (1992). “Modeling the identification of concurrenet vowels with different fundamental frequencies.” *Journal of the Acoustical Society of America* **91**(1): 233-244.

Mellinger, D. K. (1991). *Event Formation and Separation in Musical Sound*. Ph.D. thesis, Stanford University Dept. of Computer Science, Palo Alto CA.

Minksy, M. (1985). *The Society of Mind*. New York: Simon & Schuster.

Minsky, M. (1989). "Music, mind, and meaning." In *The Music Machine: Selected Readings from Computer Music Journal*, C. Roads, ed. Cambridge MA: MIT Press**:** 639-656.

Minsky, M. and O. Laske (1992). "Foreward: A conversation with Marvin Minsky." In *Understanding Music with AI: Perspectives on Music Cognition*, M. Balaban, K. Ebcioglu and O. Laske, ed. Cambridge MA: MIT Press.

Moorer, J. A. (1977). “On the transcription of musical sound by computer.” *Computer Music Journal* **1**(4): 32-38.

Narmour, E. (1990). *The Analysis and Cognition of Basic Melodic Structures*. Chicago: University of Chicago Press.

Nawab, S. H., C. Y. Espy-Wilson, R. Mani and N. N. Bitar (1998). "Knowledge-based analysis of speech mixed with sporadic environmental sounds." In *Readings in Computational Auditory Scene Analysis,* D. F. Rosenthal and H. Okuno, ed. Mahweh, NJ: Erlbaum**:** 177-193.

Ng, K., R. Boyle and D. Cooper (1996). “Automatic detection of tonality using note distribution.” *Journal of New Music Research* **25**(4): 369-381.

Oppenheim, A. V. and R. W. Schafer (1989). *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice Hall.

Parncutt, R. (1989). *Harmony: A Psychoacoustical Approach*. Berlin: Springer-Verlag.

Parncutt, R. (1997). "A model of the perceptual root(s) of a chord accounting for voicing and prevailing tonality." In *Music, Gestalt, and Computing: Studies in Systematic and Cognitive Musicology*, M. Leman, ed. Berlin: Springer. **1317:** 181-199.

Patterson, R. D., M. H. Allerhand and C. Giguere (1995). “Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform.” *Journal of the Acoustical Society of America* **98**(4): 1890-1894.

Picard, R. W. (1997). *Affective Computing*. Cambridge MA: MIT Press.

Pielemeier, W. J., G. H. Wakefield and M. H. Simoni (1996). “Time-frequency analysis of musical signals.” *Proc IEEE* **84**(9): 1216-1230.

Piszczalski, M. and B. A. Galler (1977). “Automatic music transcription.” *Computer Music Journal* **1**(4): 24-31.

Piszczalski, M. and B. A. Galler (1983). "A computer model of music recognition." In *Music, Mind, and Brain: The Neuropsychology of Music*, M. Clynes, ed. New York: Plenum Press**:** 399-416.

Plomp, R. and W. J. M. Levelt (1965). “Tonal consonance and critical bandwidth.” *Journal of the Acoustical Society of America* **38**(2): 548-560.

Povel, D.-J. and R. van Egmond (1993). “The function of accompanying chords in the recognition of melodic fragments.” *Music Perception* **11**(2): 101-115.

Rabiner, L. R., M. J. Cheng, A. E. Rosenberg and C. A. McGonegal (1976). “A comparative performance study of several pitch detection algorithms.” *IEEE Trans ASSP* **24**(5): 399-418.

Roads, C. (1996). *The Computer Music Tutorial*. Cambridge, MA: MIT Press.

Roads, C., S. T. Pope, A. Piccialli and G. de Poli, Eds. (1997). *Musical Signal Processing*. Studies on New Music Research. Lisse, NL, Swets & Zeitlinger.

Robinson, K. (1993). “Brightness and octave position: Are changes in spectral envelope and in tone height perceptually equivalent?” *Contemporary Music Review* **9**(1&2): 83-95.

Rosner, B. S. (1988). "Music perception, music theory, music psychology." In *Explorations in Music, the Arts, and Ideas: Essays in Honor of Leonard B. Meyer*, E. Narmour and R. A. Sobie, ed. Stuyvesant: Pendragon Press.

Rossi, L., G. Girolami and M. Leca (1997). “Identification of polyphonic piano signals.” *Acustica* **83**(6): 1077-1084.

Sandell, G. J. (1995). “Roles for spectral centroid and other factors in determining "blended" instrument pairings in orchestration.” *Music Perception* **13**(2): 209-246.

Scheirer, E. D. (1995). *Extracting expressive performance information from recorded music*. M.S. thesis, MIT Media Laboratory, Cambridge, MA.

Scheirer, E. D. (1996). "Bregman's chimerae: Music perception as auditory scene analysis." In *Proc. International Conference on Music Perception and Cognition*, Montreal.

Scheirer, E. D. (1997). "Pulse tracking with a pitch tracker." In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk, NY.

Scheirer, E. D. (1998a). "The MPEG-4 Structured Audio Orchestra Language." In *Proc. ICMC*, Ann Arbor, MI.

Scheirer, E. D. (1998b). “Tempo and beat analysis of acoustic musical signals.” *Journal of the Acoustical Society of America* **103**(1): 588-601.

Scheirer, E. D. (1998c). "Using musical knowledge to extract expressive performance information from recorded signals." In *Readings in Computational Auditory Scene Analysis*, D. F. Rosenthal and H. Okuno, ed. Mahweh, NJ: Lawrence Erlbaum**:** 361-380.

Scheirer, E. D. and M. Slaney (1997). "Construction and evalution of a robust multifeature speech/music discriminator." In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich.

Schellenberg, E. G. (1996). “Expectancy in melody: Tests of the implication-realization model.” *Cognition* **58**(1): 75-125.

Schmuckler, M. A. and M. G. Boltz (1994). “Harmonic and rhythmic influences on musical expectancy.” *Perception & Psychophysics* **56**(3): 313-325.

Schneider, A. (1997). ""Verschmelzung", tonal fusion, and consonance: Carl Stumpf revisited." In *Music, Gestalt, and Computing*, M. Leman, ed. Berlin: Springer-Verlag**:** 117-143.

Serafine, M. L. (1988). *Music as Cognition: The Development of Thought in Sound*. New York: Columbia University Press.

Shepard, R. N. (1964). “Circularity in judgments of relative pitch.” *Journal of the Acoustical Society of America* **36**(12): 2346-2353.

Shepard, R. N. (1982). “Geometrical approximations to the structure of musical pitch.” *Psychological Review* **89**(4): 305-333.

Slaney, M. (1994). "Auditory toolbox." Apple Computer, Inc. Technical Report #45, Cupertino CA.

Slaney, M. (1997). "Connecting correlograms to neurophysiology and psychoacoustics." In *Proc. XIth International Symposium on Hearing*, Lincolnshire UK.

Slaney, M., M. Covell and B. Lassiter (1996). "Automatic audio morphing." In *Proc. ICASSP*, Atlanta GA.

Slaney, M. and R. F. Lyon (1990). "A perceptual pitch detector." In *Proc. ICASSP*, Albequerque.

Slaney, M. and R. F. Lyon (1991). "Apple Hearing Demo Reel." Apple Computer, Inc. Technical Report #25, Cupertino CA.

Slaney, M., D. Naar and R. F. Lyon (1994). "Auditory model inversion for sound separation." In *Proc. ICASSP*, Adelaide AU.

Smith, G., H. Murase and K. Kashino (1998). "Quick audio retrieval using active search." In *Proc. ICASSP*, Seattle.

Smith, J. D. (1987). “Conflicting aesthetic ideals in a musical culture.” *Music Perception* **4**(4): 373-392.

Smith, J. D. (1997). “The place of musical novices in music science.” *Music Perception* **14**(3): 227-262.

Smoliar, S. W. (1991). “Review of *The Analysis and Cognition of Basic Melodic Structures: The Implication-Realization Model* by Eugene Narmour.” *In Theory Only* **12**(1-2): 43-56.

Smoliar, S. W. (1992). “Elements of a neuronal model of listening to music.” *In Theory Only* **12**(3-4): 29-46.

Smoliar, S. W. (1995). “Parsing, structure, memory and affect.” *Journal of New Music Research* **24**(1): 21-33.

Steedman, M. (1994). “The well-tempered computer.” *Philosophical Transactions of the Royal Society of London (A)* **349**: 115-131.

Summerfield, Q., A. Lea and D. Marshall (1990). “Modelling auditory scene analysis: strategies for source segregation using autocorrelograms.” *Proceedings of the Institute of Acoustics* **12**(10): 507-514.

Temperley, D. (1997). “An algorithm for harmonic analysis.” *Music Perception* **15**(1): 31-68.

Terhardt, E. (1974). “Pitch, consonance, and harmony.” *Journal of the Acoustical Society of America* **55**(5): 1061-1069.

Terhardt, E. (1978). “Psychoacoustic evaluation of musical sounds.” *Perception & Psychophysics* **23**(6): 483-492.

Terhardt, E. (1982). "Impact of computers on music: an outline." In *Music, Mind, and Brain: The Neuropsychology of Music*, M. Clynes, ed. New York: Plenum Press**:** 353:369.

Therrien, C. W. (1989). *Decision, estimation and classification: An introduction to pattern recognition and related topics.* New York: Wiley.

Thompson, W. F. (1993). “Modeling perceived relationships between melody, harmony, and key.” *Perception & Psychophysics* **53**(1): 13-24.

Thompson, W. F. and R. Parncutt (1997). “Perceptual judgments of triads and dyads: Assessment of a psychoacoustic model.” *Music Perception* **14**(3): 263-280.

Thomson, W. (1993). “The harmonic root: A fragile marriage of concept and percept.” *Music Perception* **10**(4): 385-416.

Van Immerseel, L. M. and J.-P. Martens (1992). “Pitch and voiced/unvoiced determination with an auditory model.” *Journal of the Acoustical Society of America* **91**(6): 3511-3526.

van Noorden, L. (1983). "Two-channel pitch perception." In *Music, Mind, and Brain: The Neuropsychology of Music*, M. Clynes, ed. New York: Plenum Press**:** 251-269.

Vercoe, B. L. (1984). "The synthetic performer in the context of live performance." In *Proc. ICMC*, Paris.

Vercoe, B. L. (1988). "Hearing polyphonic music on the connection machine." In *Proc. First AAAI Workshop on Artificial Intelligence and Music*, Minneapolis.

Vercoe, B. L. (1996). "Csound: A Manual for the Audio-Processing System." MIT Media Lab Technical Report Cambridge MA.

Vercoe, B. L. (1997). "Computational auditory pathways to music understanding." In *Perception and Cognition of Music*, I. Deliège and J. Sloboda, ed. London: Psychology Press**:** 307-326.

Vercoe, B. L., W. G. Gardner and E. D. Scheirer (1998). “Structured audio: The creation, transmission, and rendering of parametric sound representations.” *Proceedings of the IEEE* **85**(5): 922-940.

Vercoe, B. L. and M. S. Puckette (1985). "Synthetic rehearsal: Training the synthetic performer." In *Proc. ICMC*, Burnaby BC, Canada.

Vos, P. G. and E. W. Van Geenen (1996). “A parallel-processing key-finding model.” *Music Perception* **14**(2): 185-224.

Wang, A. (1994). *Instantaneous and Frequency-Warped Signal Processing Techniques for Auditory Source Separation.* Ph.D. thesis, Stanford University CCRMA, Stanford, CA.

Wang, D. (1996). “Primitive auditory segregation based on oscillatory correlation.” *Cognitive Science* **20**: 409-456.

Warren, R. M. (1970). “Perceptual restoration of missing speech sounds.” *Science* **167**: 392-393.

Warren, R. M., C. J. Obusek and J. M. Ackroff (1972). “Auditory induction: perceptual synthesis of absent sounds.” *Science* **176**: 1149-1151.

Weintraub, M. (1985). *A theory and computational model of auditory monaural sound separation*. Ph. D. thesis, Stanford University Dept. of Electrical Engineering, Palo Alto, CA.

Wightman, F. L. (1973). “The pattern-transformation model of pitch.” *Journal of the Acoustical Society of America* **54**(2): 407-416.

Wold, E., T. Blum, D. Keislar and J. Wheaton (1996). “Content-based classification, search, and retrieval of audio.” *IEEE Multimedia* **3**(3): 27-36.

Zwicker, E. and H. Fastl (1990). *Psychacoustics: Facts and Models*. Berlin: Springer-Verlag.