Music-Listening Systems

Article · May 2000

Source: CiteSeer

CITATIONS
94

READS
321

4 authors, including:

Barry Lloyd Vercoe
Massachusetts Institute of Technology
55 PUBLICATIONS **1,512** CITATIONS

SEE PROFILE

Music-Listening Systems

Eric D. Scheirer

B.A. Linguistics, Cornell University, 1993

B.A. Computer Science, Cornell University, 1993 (*cum laude*)

S.M. Media Arts and Sciences, Massachusetts Institute of Technology, 1995

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements of the degree of

Doctor of Philosophy

at the Massachusetts Institute of Technology

June, 2000

Author

Program in Media Arts and Sciences
April 28, 2000

Certified By

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

Accepted By

Stephen A. Benton
Chair, Departmental Committee on Graduate Students
Program in Media Arts and Sciences
Massachusetts Institute of Technology

Music-Listening Systems

Eric D. Scheirer

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning, on April 28, 2000,
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Abstract

When human listeners are confronted with musical sounds, they rapidly and automatically orient themselves in the music. Even musically untrained listeners have an exceptional ability to make rapid judgments about music from very short examples, such as determining the music's style, performer, beat, complexity, and emotional impact. However, there are presently no theories of music perception that can explain this behavior, and it has proven very difficult to build computer music-analysis tools with similar capabilities. This dissertation examines the psychoacoustic origins of the early stages of music listening in humans, using both experimental and computer-modeling approaches. The results of this research enable the construction of automatic machine-listening systems that can make human-like judgments about short musical stimuli.

New models are presented that explain the perception of musical tempo, the perceived segmentation of sound scenes into multiple auditory images, and the extraction of musical features from complex musical sounds. These models are implemented as signal-processing and pattern-recognition computer programs, using the principle of *understanding without separation*. Two experiments with human listeners study the rapid assignment of high-level judgments to musical stimuli, and it is demonstrated that many of the experimental results can be explained with a multiple-regression model on the extracted musical features.

From a theoretical standpoint, the thesis shows how theories of music perception can be grounded in a principled way upon psychoacoustic models in a computational-auditory-scene-analysis framework. Further, the perceptual theory presented is more relevant to everyday listeners and situations than are previous cognitive-structuralist approaches to music perception and cognition. From a practical standpoint, the various models form a set of computer signal-processing and pattern-recognition tools that can mimic human perceptual abilities on a variety of musical tasks such as tapping along with the beat, parsing music into sections, making semantic judgments about musical examples, and estimating the similarity of two pieces of music.

Thesis Supervisor: Barry L. Vercoe, D.M.A.
Professor of Media Arts and Sciences

This research was performed at the MIT Media Laboratory. Primary support was provided by the Digital Life consortium of the Media Laboratory, with additional support from Interval Research Corporation. The views expressed within do not necessarily reflect the views of supporting sponsors.

Doctoral Dissertation Committee

Thesis Advisor

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

Thesis Reader

Rosalind W. Picard
Associate Professor of Media Arts and Sciences
Massachusetts Institute of Technology

Thesis Reader

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

Thesis Reader

Malcolm Slaney
Member of the Technical Staff
Interval Research Corporation
Palo Alto, CA

Acknowledgments

It goes without saying, or it should, that the process of doing the research and writing a dissertation is really a collaborative effort. While there may be only one name on the title page, that name stands in for a huge network of colleagues and confidants, without whom the work would be impossible or unbearable. So I would forthwith like to thank everyone who has made this thesis possible, bearable, or both.

I have been lucky to reside in the Machine Listening Group of the Media Laboratory over a time period that allowed me to collaborate with so many brilliant young researchers. Dan Ellis and Keith Martin deserve special mention for their role in my research life. Dan took control of the former "Music and Cognition" group and showed us how we could lead the way into the bright future of Machine Listening. Keith and I arrived at the Lab at the same time, and have maintained a valuable dialogue even though he escaped a year earlier than I. I feel Dan's influence is most strongly in the depth of background that he encouraged me, by example, to develop. Many of the new ideas here were developed in collaboration with Keith, and the stamp of his critical thinking is on every page.

The rest of the Machine Listening Group students past and present were no less critical to my research, whether through discussion, dissension, or distraction. In alphabetical order, I acknowledge Jeff Bilmes, Michael Casey, Wei Chai, Jonathan Feldman, Ricardo García, Bill Gardner, Adam Lindsay, Youngmoo Kim, Nyssim Lefford, Joe Pompei, Nicolas Saint-Arnaud, and Paris Smaragdis. During the final year of my stay here, I had the good fortune to be exiled from the main Machine Listening office space—this is good fortune because I landed in an office with Push Singh. Push's remarkable insight into the human cognitive system, and his willingness to speculate, discuss, and blue-sky with me, have enriched both my academic life and my thesis.

Barry Vercoe, of course, deserves the credit for organizing and managing this wonderful group of people. Further, he has given us the greatest gift that an advisor can give his students—the freedom to pursue our own interests and ideas. Barry's willingness to shelter us from the cold winds of Media Lab demo pressure has been the catalyst for the group's academic development. Connie van Rheenen deserves special credit for the three years that she has been our administrator, resource, and den mother. I don't recall how it was that we managed our lives before Connie joined us.

The other members of my committee, Roz Picard, Perry Cook, and Malcolm Slaney, have been exemplary in their continuous support and encouragement. From the earliest stages of my proposal to the last comments on this thesis, they have been a bountiful source of inspiration, encouragement, and suggestions. I would also like to specially mention Prof. Carol Krumhansl of Cornell University. My first exposure to music psychology and the other auditory sciences came in her class. The two years that I spent as her undergraduate assistant impressed on me the difficulty and reward of a long-term experimental research program, and her analytic focus is unparalleled. Carol also encouraged me to apply to the Media Lab, a decision that I have never regretted for a second.

My parents have been uniquely unflagging in their ability to make me feel free to do exactly as I pleased, while still demanding that I live up to my ability. From the earliest time I can remember, they have instilled in me a love of learning that has only grown over the years.

Finally, there is no way to express the love and gratitude that I feel for my wife Jocelyn every moment of every day. Of all the people I have ever met, she is the most caring, most supportive, and most loving. I owe everything I have done and will do to her.

Music-Listening Systems

Eric D. Scheirer

Contents

CHAPTER 1 INTRODUCTION

This is a dissertation about listening to music. The kinds of listening that I will explore fall into two main areas: people listening to music, and machines listening to music. In order to teach machines how to listen to music, we must first understand what it is that people hear when they listen to music. And by trying to build computer machine-listening systems, we will learn a great deal about the nature of music and about human perceptual processes.

Music is one of the most striking activities that separates humans from animals. People everywhere and at every time throughout history have made music. The significance of this universality is undiminished—rather, accentuated—by music's seeming purposelessness. Music is also a rich source of inspiration for critical thinkers of all sorts. Everywhere that we can find historical records, we find that scholars and scientists have created religious, aesthetic, psychological, philosophical, cultural, scientific, spiritual, and (recently) computational theories regarding the analysis, function, and operation of music.

Music works its way into every aspect of human culture. Composers, performers, and listeners shape music to fill new cultural niches as soon as they arise. In the West, we bring music into movies, sporting events, religious ceremonies, nightclubs, living rooms, and shopping malls. There is no other form of expression that we use so broadly and in as many ways as music. Most relevant to the research that I present in this dissertation, interest in music has recently exploded on the Internet. Companies built around music have a pressing need for new kinds of music technology to support the growing passion for online music in the networked world.

Scholars and scientists have already written many books and dissertations about the perception of music. The highest-level question that motivates my work—what is it that people hear when they listen to music?—is shared by most of the research in the existing literature. But there is a major difference between the approach I will present here and most of the previous work in music psychology. This difference is an emphasis on *sound* and its centrality in the music-listening process.

Music is ineffably a phenomenon rooted in sound. Music is built from sound (Bregman, 1990, p. 455); without sound, there is no music. All the musics of the world arise through an elaborate interaction between the sound-generating properties of physical objects (termed *instruments* of music) and the sound-analysis properties of the human auditory system. The auditory system was not adapted to music; rather, music is the way it is because of the nature

of the human hearing process. Thus, I submit, we can only really understand human music-listening when we understand it as something to do with sound.

Psychoacoustics, the science of sound perception, connects the physical world of sound vibrations in the air to the perceptual world of the things we hear when we listen to sounds. My dissertation can be seen as an attempt to make music perception rest on a foundation of psychoacoustics. But this is not easy, because most previous inquiries into psychoacoustics only treat very simple sounds, like tones and buzzes and clicks. These test sounds are far removed from the elegant and emotion-laden sounds that we perceive as music. And so in order to build the right sort of psychoacoustic foundation, I will present some new theories of sound perception that are better-equipped to handle the complexities of real musical sound. These ideas will be presented through the development of new computational models for sound processing.

From a scientific perspective, it is crucial to develop stronger connections between music perception and psychoacoustics. A theory of musical listening cannot really be said to be a theory of perception at all unless it connects the musical perceptions to the sound. Drawing this connection firms up the theoretical rigor of music-perception research in general. But as well as this scientific advantage, there is a practical advantage to understanding music perception through psychoacoustics: the possibility of giving computers more advanced musical abilities.

Most music today is represented at one time or another in digital form; that is, as a sequence of bits whose values correspond to the sound-pressure levels in an analog acoustic waveform. It is easy to make computers process these bits in order to tell us simple things about the sound, such how much acoustic power it contains. But it has proven amazingly difficult to build computer systems that can understand the things that human listeners understand immediately and unconsciously when they hear music, such as the musical genre, or the instruments that are being played, or the beat in the music.

I submit that there is a straightforward reason for this difficulty. This is that *people hear the things they do in music because of the basic sound-processing capabilities they possess*. People have evolved sophisticated tools for processing and understanding sources of sound in the world. People use sound to sense danger and locate each other in the dark and in places that are visually obscured. The auditory capabilities that we use to process music are exactly the same—the same ears, the same cochleae, the same auditory brain—as those we use to get by in the world. Composers and musicians have learned, through a complex sociocultural evolution, to create special kinds of sounds that tickle these capabilities in interesting ways.

It's no good to try to approach the construction of music-listening computer systems through engineering alone. In order to understand the nature of music and how it is perceived, we must understand the functioning of the perceptual system. Further, from a more philosophical viewpoint, we must refer to human listening to know what a computer music-listening system should do in the first place. There is no ground truth that tells us what the computer is "supposed" to hear in the sound when it listens to music. Rather, the things that the computer should hear are just those things that people hear upon listening to the same music.

This means that in order to understand why music is the way it is, and in order to make computers able to process music more like people, we must first understand the way in which people hear music. The more we learn about human listening, the easier it will be to build machine music-listening systems that solve useful problems on our behalf. And it will be only when we understand human music listening as a psychoacoustic behavior that scientific results will translate naturally into new algorithms for music processing.

So my thesis makes contributions to three areas of inquiry that bear a complex overlapping relationship to one another. In order to understand music perception, we must understand it as a special kind of psychoacoustics. In order to understand this complex kind of psychoacoustics, we must build computer models that embody new theories of hearing. And

in order to build more musically-intelligent computers, we must understand the human perception of music. This three-way connection is at the heart of the results I will present in the rest of the dissertation.

1.1. Organization

My dissertation is divided into eight chapters. In chapter 2, ***Background***, I review relevant research from the fields of psychoacoustics, music perception, auditory scene analysis, and musical signal processing. This review will present the fundamentals of the science and engineering I am engaged in, and some reasons why I think the previous research has not been particularly successful. In chapter 3, ***Approach***, I explain more formally the problems I am trying to solve, the theory of listening that I am defending, and the implications of this theory for the construction of computational models. These two chapters present the context of the dissertation and set the stage for the presentation of new results.

Chapter 4, ***Musical Tempo***, presents a simple model that demonstrates the music-listening-system concept. The model that I introduce in this chapter is capable of listening to music and hearing the beat in it. I present the model in signal-processing terms, compare it to other psychoacoustic models, and demonstrate with a short listening experiment that the results produced by the model are similar to the results given by human listeners when asked to find the beat in a piece of music. The simplicity of this model will make it easy to reflect upon the way it embodies my approach.

In chapter 5, ***Musical Scene Analysis***, I present a much more complex model of a much more complex human perceptual behavior—the sense that music is made of multiple "instruments" or "voices" that are being played at the same time, overlapping but still individually salient. The model is based on new ideas about the psychoacoustic processing of complex scenes. In this chapter, I present the model and demonstrate extensively its capabilities to model human percepts on a variety of basic psychoacoustic stimuli. I also discuss the connections between this model and other sound-segregation models in the literature. This is the most technical chapter of the dissertation, in both computational and psychoacoustic details.

Chapter 6, ***Musical features,*** extends the models in Chapter 4 and 5 by introducing techniques for extracting perceptually-motivated features from the musical scene. I show how a number of simple features can be readily extracted from the psychoacoustic models.

Chapter 7, ***Musical Perceptions***, presents the results of two human listening experiments and two computer models that can predict these results. The experiments are about the ability of human listeners to react with "first impressions" immediately upon hearing a musical stimulus, such as that the music is fast, loud, and complex. The computer models presented in this chapter are thus music-listening systems that can make immediate judgments about music like people can. I also briefly demonstrate how this model might be applied to practical problems such as automatically classifying music into genre categories and performing music similarity-matching.

Chapter 8, ***Conclusion***, summarizes the contributions made in the dissertation and suggests directions for further research.

CHAPTER 2 BACKGROUND

Three areas of previous research bear the most direct relationship to my dissertation. 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 that 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.

In this chapter, I will present an in-depth discussion of previous research in these areas. I will not attempt to include all research in these disciplines, but only those I see as most current and most directly connected to the main direction of the new results that I will present in subsequent chapters. After this, in Section 2.4, I will discuss projects that cross this (somewhat arbitrary) division of boundaries. The projects in that final subsection are the ones closest to the research reported in the main body of my dissertation.

2.1. Psychoacoustics

Psychoacoustics as a field of inquiry has been in existence for more than a century. Some of the earliest studies recognizable as psychological science in the 19th century were concerned with the perception of the loudness and pitch of sounds. Even before scientific methods developed, philosophers engaged in speculation about the nature of sound. Psychoacoustic 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 at the same time.

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 early stages of hearing. 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 2.2 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 section; I do not view masking models as directly relevant to my research. Finally, there is a new area of psychoacoustic research dealing with spectral-temporal pattern integration that may lead in the future to a better understanding of how auditory scene analysis is grounded in low-level behavior. I will conclude this section with a review of these studies.

2.1.1. Pitch theory and models

"Pitch" is the perceptual correlate of the frequency of a simple tone (in Chapter 3, Section 3.1, I will expand more on the relationship between perceptual and physical attributes of sounds). 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 the research that examined such signals; Zwicker and Fastl (1990) and Hartmann (1996) have provided more recent reviews of the data. Zwicker and Fastl gave a list of no less than eleven different types of pitched sounds—even more have been discovered since. The primary task in building functional models of pitch processing is to explain this data by showing how, given an acoustic signal as input, the ear and brain generate the pitch percept. Stated another way, the goal of a functional pitch model is to explain why different sounds have the pitches that they do.

Place models of pitch

There are two main 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, place models do not easily explain certain pitch phenomena, such as iterated noise signals, that 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.

An advantage of central models is that they can naturally explain the phenomenon of *dichotic pitch* (Moore, 1997, pp. 202-203), in which part of a sound is presented to one ear and part to the other. Neither part, heard in isolation, is perceived to have a pitch, but when both ears hear the correct sounds at once, a clear pitch is perceived. It would seem that some sort of central mechanism is required to integrate the percepts from the two ears. However, whether the central mechanism must be strictly place-based, or could have important temporal-dynamics aspects, is not presently clear.

Temporal models of pitch

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

The first temporal model of pitch was presented by Licklider (1951b), who proposed an analysis technique 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 (1991), and others; it has since come to be called the *autocorrelogram* method of pitch analysis (see below) and is today the preferred model.

Meddis and Hewitt (1991) 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 of isolated sounds.

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 *et al.*, 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 temporal asymmetries in pitched signals.

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.

Recently, de Cheveigné (1993; 1998b) 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 has also shown that the cancellation model is a useful way to explain the perception of multiple-pitch stimuli (de Cheveigné, 1997).

Details of the autocorrelogram

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 there is now reasonably strong (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 (Section 2.1.2) have used the autocorrelogram as the front end.

Analysis of the dynamics of subband periodicity will form an important part of Chapters 4 and 5, so I will go a bit deeper here and explain the exact form that this important framework takes when it is implemented as a computational theory. The basic subband-periodicity model, as described by Meddis and Hewitt (1991), is shown in Figure 2-1. A monaural sound signal is processed by, first, passing it through a bank of filters that approximate the passive filtering process in the cochlea; second, applying smoothing and rectification to the signal as a model for transduction of the motion on the basilar membrane to neural impulses by the inner hair cells; and finally, determining the periodicity of the output of each cochlear subband through the use of autocorrelation or a similar mechanism.

The output of the first stage, the filterbank processing and rectification, is sometimes called the *cochleagram* because it is a first-order model of the average or ensemble firing rate of nerve fibers at various positions along the cochlea. A cochleagram for a test signal (the "McAdams oboe") is shown in Figure 2-2.[1]

The output of the second stage is termed the *autocorrelogram* when the periodicity detection is performed using autocorrelation. The autocorrelogram is the 3-D volumetric function mapping a cochlear frequency channel, temporal time delay (or *lag*), and time to a periodicity estimate in that frequency band at that lag and time (Slaney and Lyon, 1990). The autocorrelogram is most often visualized as a sequence of frames, each a constant-time slice through the full figure (Figure 2-3). Each row in one frame of the autocorrelogram is the autocorrelation of the recent history of the corresponding row in the cochleagram.

In order to model the perception of pitch, each frame is summed vertically to arrive at the *summary autocorrelogram* (SAC) as shown in the bottom panels of each frame in Figure 2-3. Meddis and Hewitt (1991) demonstrated that the peaks in the summary autocorrelogram account for the perception of pitch in a wide variety of stimuli. In the variant shown here, the samples of the lag (periodicity) axis are spaced logarithmically. Ellis (1996a) suggested this, arguing that it more accurately reflects human pitch perception. This variant has additional processing advantages that will become clear in Chapter 5.

[1] A more complete description of the sound appears in Section 5.4.1. Synthesis code for this example and the other test sounds used in Chapter 5 is included in Appendix B.

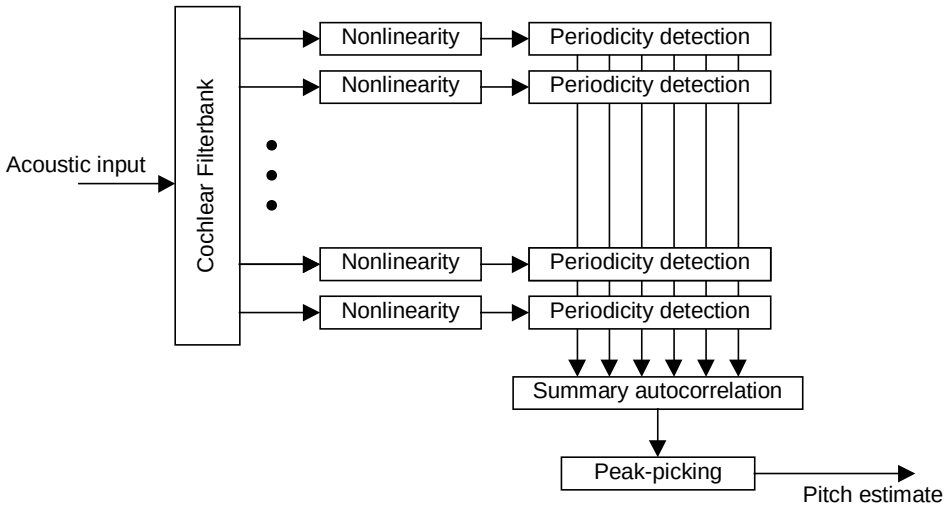


Figure 2-1**:** The subband-periodicity model of pitch detection, following Meddis and Hewitt (1991). A cochlear filterbank separates the input acoustic signal into subbands. A rectifying nonlinearity is used to track the envelope energy in each band. Periodicity detection (for example, autocorrelation) is applied to each envelope, and the resulting periodicity estimates are summed across frequencies to give a summary of the periodicity in the signal. The pick of the summary autocorrelogram function corresponds well to the human perception of pitch in the input signal.

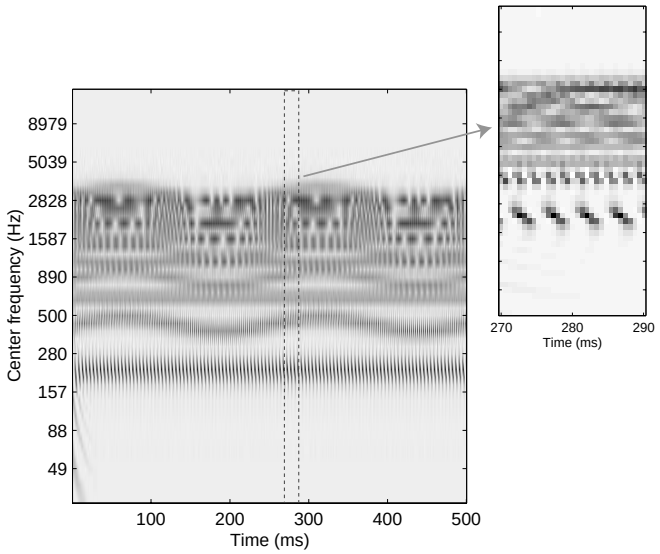


Figure 2-2**:** The output of a computational model of the cochlear filterbank and hair-cell transduction—the *cochleagram*—of part of the "McAdams oboe" sound (McAdams, 1984). This sound is comprised of the first ten harmonics of a 220 Hz fundamental with coherent vibrato (10% depth, 4 Hz) applied to the even harmonics only (this is a great deal of vibrato; it is used here to make the diagram clearer). The percept is that of a clarinet-like sound with pitch at 220 Hz and a soprano-like sound with pitch at 440 Hz. The main panel shows the broad frequency resolution of the filterbank; the vibrato in the second and fourth harmonics can be easily seen. The small panel presents a closer view of the time range around 280 ms; phase-locking in the middle frequencies and lack of phase-locking in the high frequencies is observed.

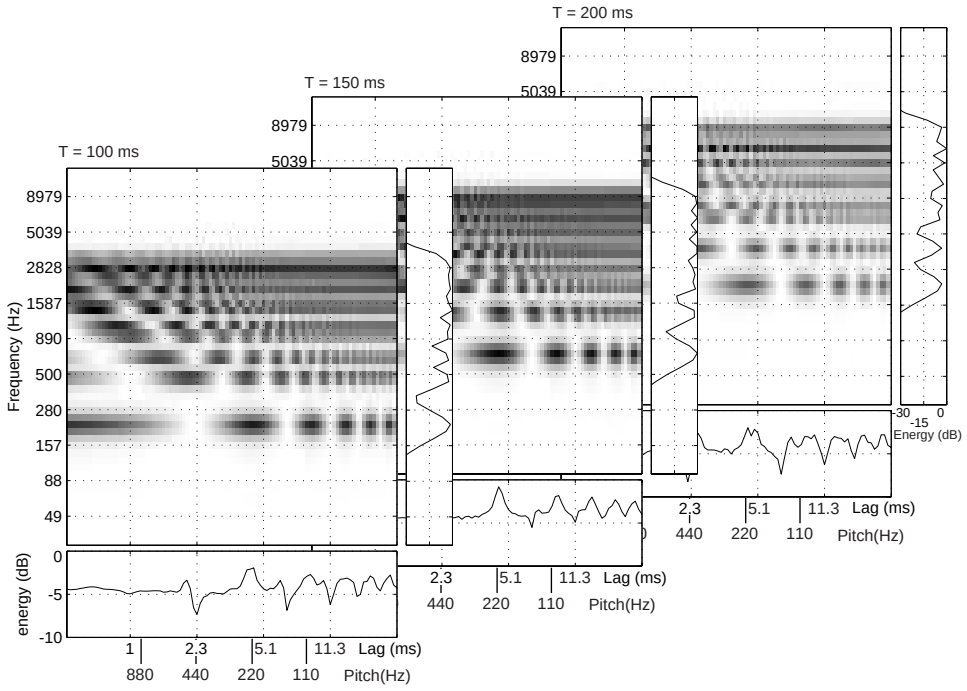


Figure 2-3: Three frames of the log-lag autocorrelogram of the McAdams oboe. For each frame, the main panel shows a constant-time slice through the autocorrelogram volume, calculated as described in the main text; the bottom panel shows the *summary autocorrelogram*, which is the sum across frequencies of the periodic energy at each lag (the sum is calculated in the linear-energy domain and then converted to dB scale for presentation); and the right panel shows the *energy spectrum*, which is the zero-lag energy in each channel. The three frames highlight different portions of the vibrato phase for the even harmonics; the first shows a point in time at which the even harmonics are sharp relative to the odd harmonics, the second at which the even harmonics are in tune with the odd harmonics, and the last when the even harmonics are flat. Readers familiar with pitch-based source separation techniques will observe the difficulty in distinguishing the in-tune from detuned partials using only the information in the summary autocorrelation.

An important thread of research that has recently been connected to pitch-modeling is the study of neural mechanisms for auditory processing. New reports suggest evidence of the information needed to extract pitch (Cariani, 1996; Cariani and Delgutte, 1996), temporal envelope (Delgutte *et al.*, 1997), modulation spectrum (Langner, 1992), and other musically-relevant information is present in the neural anatomy (although this is different from 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.

2.1.2. Computational 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 scenes, where multiple sources that change independently over time are present. Two sub-fields 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, called streams of sound.

Bregman summarized his pioneering work in a classic text that 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. In computer modeling, it has proven difficult to extract components robustly enough in the presence of noise or interfering sounds to build systems that can function for complex sound scenes.

The approaches reported over the last 15 years in the ASA literature have been strongly functionalist and computational in nature. Brown and Cooke (1994a) termed the discipline of constructing computer models to perform auditory source segregation *computational* auditory scene analysis (CASA).

In his dissertation and follow-on work, McAdams (1983; 1984; 1989) showed the important role of temporal modulation in the perceptual segregation of sounds. He demonstrated in psychoacoustic experiments 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 2.2) 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 (1951b) 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; this is the starting point of the system discussed in Chapter 5.

Mellinger's (1991) thesis 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 with a sinusoidal representation, where additive synthesis regenerates the components after they have been isolated; however, it is more difficult when using 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 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 that "guess" what is about to happen in the signal. A bottom-up signal analysis is 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 *et al.*, 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. His research was also the first to consider the perception of such complex sounds.

S. H. Nawab and his collaborators developed a robust testbed for the construction of PDCASA systems, called *IPUS* for *Integrated Processing and Analysis of Sound* (Klassner *et al.*, 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.

K. D. Martin, drawing on the work of Ellis, demonstrated in his dissertation (Martin, 1999) that autocorrelogram-based front-end processing can be used to support the model-based perception of sound. This is an important step forward, because in a prediction-driven approach, it is essential that the system maintain robust models of what sound-producing objects are present in the world and what sounds they are contributing to the overall acoustic scene. Martin showed how to connect autocorrelogram-based parameters to the physical properties of musical instruments, and thereby was able to construct a computer system capable of robustly recognizing instruments from their sounds. He attempted no segregation—although he used complex real-world examples of monophonic performance for evaluation—but it is clear how a system like his and a system like Ellis's might be integrated.

A second type of computational-auditory-scene-analysis system of recent interest is the model based on *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 shown to correspond to 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. I will return to this point in Chapter 3, Section 3.1.3 *et seq.*

2.1.3. Spectral-temporal pattern analysis

In the last fifteen years or so, several new psychoacoustic phenomena have been discovered that seem to point to connections between low-level aspects of hearing such as the basic transduction of sound into neural impulses, and higher aspects such as those discussed in the previous section. These have been called phenomena of *spectral-temporal pattern analysis* because they are exhibited only in psychoacoustic tests with sufficiently complex, patterned stimuli. I will discuss two important such phenomena: *comodulation release from masking*

and *profile analysis*. The experimental data on both phenomena have been recently reviewed by Hirsh and Watson (1996).

Comodulation release from masking (CMR, for "comodulation masking release") is a recently-discovered auditory phenomenon that is very important in drawing connections between theories of low-level psychoacoustics and theories of the analysis of auditory scenes. In its most basic form, CMR can be demonstrated by determining the masking level of a pure-tone signal by a fluctuating (random or periodic) narrow band of noise. That is, using psychometric tests, the threshold of audibility (in terms of SNR) of a pure tone in a modulated band of noise is measured. When additional ("flanking") bands of noise that are coherently modulated with the masker are added to the signal-plus-masker stimulus, the masking level increases. That is, the comodulated flanking noise "releases" the signal from masking. This is true whether the flanking bands are nearby (in frequency) to the signal band, or far away. However, a corresponding release from masking does not occur if the flanking bands are modulated independently from the masking band. The CMR phenomenon depends critically on the temporal correspondence of different frequency bands.

It is believed that CMR is evidence that the auditory system can compare sounds across cochlear channels in the formation of early auditory percepts. (The scene-analysis model that I will present in Chapter 5 depends on this assumption as well.) CMR was first discovered by Hall *et al.* (1984) and has since been the subject of dozens of studies. Reviews of the experimental data have been presented recently (Moore, 1997, pp.121-127); several processing models have also been developed (Berg, 1996; Verhey *et al.*, 1999). There is a general lack of agreement in the literature regarding whether the experimental CMR data are indicative of one main form of the phenomenon or several subforms.

Profile analysis (Green, 1996) is another phenomenon that apparently reveals cross-frequency processing in early stages of audition. In the basic form of this paradigm, two pure tones with the same pitch but slightly different loudness are presented sequentially, one after the other with a short delay between them. The just-noticeable-difference (JND) in level between the two tones is measured with standard psychometric procedures. Then, several other frequency components are added surrounding each tone, exactly the same for each. The only difference between the first stimulus and the second remains the slight change in loudness of a single component. In this condition, the JND becomes smaller—it is apparently easier to hear the differences between the target components when the unchanged components are present for contrast.

Further, when the flanking tones are added, the effect of the interstimulus delay time becomes smaller. That is, with tones in isolation, the more distance in time there is between the stimuli, the more difficult it is to hear differences in their level (the subject "forgets" the level). But with flanking tones, it is no more, or only slightly more, difficult to hear differences with long delays as compared to short ones.

It is thought that the results of profile-analysis experiments may be explained by some kind of spectral-shape hearing, as in the perception of vowels, or perhaps by a sort of subtractive comparison between the target and flanking tones.

The importance of CMR and profile analysis is that neither of these phenomena is directly compatible with the simple classical model of sound perception. In the classical model, the cochlear filterbank separates the sound into frequency components or spectral bands, and this frequency-based separation of sound is preserved throughout further processing. Both CMR and profile analysis seem to require a stage of cross-frequency processing in the early auditory system. The auditory-scene-analysis model I will present in Chapter 5 will be shown to explain one particular form of the CMR phenomenon.

It is striking that, as soon as we began the detailed study of stimuli with properties that were not pathologically simple, we immediately learned things about psychoacoustic processing that could not be reconciled with previous models of the ear and auditory system. It is likely

that there is more to be learned by continuing the study of still-more complex sounds. However, experimental work on such sounds is very difficult (since the degrees of freedom are so numerous), and so principled research can proceed only at a modest pace.

From a practical standpoint, previous advances in psychoacoustic knowledge have led directly to great advances in telecommunications technology. The early study of perceptual limits and intelligibility at Bell Laboratories was instrumental in the development of the US telephone network system (Allen, 1996). More recently, spectral models of loudness and masking behavior have led directly to advances in low-bitrate perceptual coding of sound and music (Brandenberg, 1998), enabling a multimillion-dollar industry in Internet music delivery. It is reasonable to believe that continuing advances in psychoacoustic knowledge will present attendant gains in our ability to engineer systems for the use and manipulation of speech, music, and other sounds.

2.2. 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 that is most relevant to my dissertation in six areas: pitch, tonality, and melody; tonal consonance and fusion; the perception of musical timbre; 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 does not fit directly into the mainstream of these sorts of endeavors; in fact, my own views (see Chapter 3, Sections 3.2 and 3.4) are generally contrary to those most often articulated by music psychologists. However, I believe that my dissertation is relevant to questions posed by psychologists as well as those posed by psychoacousticians, and so I will articulate the relationship of my research to the relevant sub-disciplines 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).

2.2.1. 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 subsumption 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 by Shepard (1964), as shown in Figure 2-4. 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

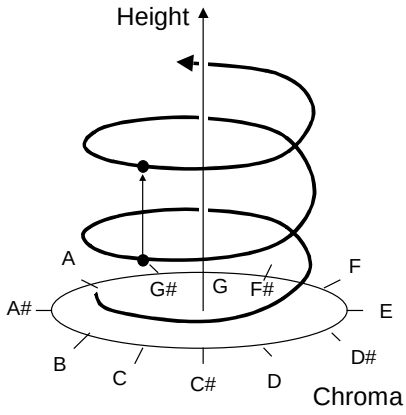


Figure 2-4: The Shepard (1964) helical model of musical pitch. Pitch in this model is expressed in two dimensions. The *chroma* direction, around the helix, relates notes that are chromatically near each other. The *height* direction, up the vertical axis, relates notes that have the same chroma across octaves.

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. L. 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 *et al.*, 1996), the relationships between melody, harmony, and key (Povel and van Egmond, 1993; Thompson, 1993; Holleran *et al.*, 1995) and the development of these processes in infants. Several researchers have also focused on the key-finding algorithm, attacking it as a practical engineering problem somewhat distinct from its origins in perceptual science (Ng *et al.*, 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 in Section 2.2.4). 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).

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 structuralist 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.[2] These issues 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.

The relationship between pitch considered as a perceptual property of sound and examined through psychoacoustic experiments (as discussed in Section 2.1.1), and pitch as a musical property that is subsumed into larger structures like melodies, has not been addressed in a principled way. Music psychologists generally treat the analysis of pitch as a "preprocessing" step and assume (usually implicitly) that the pitches of sounds that are heard in a musical scene are made available to a central cognitive mechanism. Such a stance has strong implications for the sorts of processing models that must be used; I will return to this point in Chapter 3, Section 3.3.

2.2.2. Perception of chords: tonal consonance and tonal fusion

Relatively few reports in the literature have examined the perception of vertical music structures such as chords. This is somewhat surprising; given the music-theoretic importance of the roots of chords (discussed by Thomson, 1993), and of the invariance of chords under inversion, it would be revealing to have experimental data confirming or disputing these

[2] "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 for their music.

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.

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 that 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.

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.

Helmholtz (1885), 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 conducted 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. These studies examine the relationship between 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 the principle of "avoid tonal fusion", and the principle of "seek 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 Huron's experiments using acoustic signals. To conduct such an analysis only from the written notation misses the crucial fact that a listener is not perceiving the notation directly, but is perceiving a sound that may or may not be somehow perceptually converted to a notation-like form.

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 the possible interactions between timbre, tonal fusion, and harmonic/melodic processing.

2.2.3. The perception of musical timbre

In all of the cases discussed in the previous section, 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 and Sellmer, 1992).

Integrating timbre into music-psychological studies, in the sense of controlling and understanding its effects on other phenomena has proven to be a difficult problem. There is, of course, an extensive literature on the perception of timbre as an isolated property of musical sounds (Martin (1999) has provided a thorough review). Most of the work in this literature focuses on geometric models of "timbre space" (Grey, 1977), in which the multidimensional-scaling paradigm is used to discover the underlying dimensions of variation in a set of stimuli. This has been a popular method of inquiry, but the psychological reality of the timbre-space model has never been compellingly demonstrated, in my opinion.

More promising is research based in ecological acoustics that attempts to discover how it is that listeners can learn about characteristics of physical objects from their sounds. Listeners seem to naturally make judgments about the underlying physical reality of complex sounds, even with impoverished stimuli. For example, Warren and Verbrugge (1984) showed that listeners can readily distinguish "bouncing" sounds from "breaking" sounds. Freed (1990) showed that listeners can identify the hardness of the mallet used to strike a vibraphone from its sound. Li *et al.* (1991) showed that listeners can accurately classify the gender of a person from listening to the sound of his/her footsteps. Finally, Cabe and Pittenger (2000) showed that listeners can keep track of a vessel being filled with liquid and estimate the point at which the vessel will be full, from sound cues only. Such results lead naturally to a *model-based* theory of timbre, in which the goal of sound perception is to associate properties of immanent sound models with their perceived realizations in sound, and thereby to understand the physical world. The recent dissertations of Casey (1998) and Martin (1999) take this approach to some degree in their construction of computational systems.

The difficulty in effectively integrating timbre into music-psychological theories stems from two basic problems. The first is that there is no standard music-theoretical approach to timbre. Most music-psychology studies take Western music theory and the organizational principles it provides (notes and chords, rhythms and melodies and harmonic structures, hierarchical relationships in music) as a starting point. The *result* of experimental study may be to criticize the tenets of music theory as lacking psychological reality, but the *concerns* and *questions* posed by music psychologists have nearly always been borrowed from the concerns

of music theorists. And so, since there is no well-organized Western theory of timbre and its use in music, music psychologists have had a difficult time finding a foothold on which to begin rigorous study.

The second problem, related to the first, is that timbre seems as though it is an aspect of music that is more intimately connected to the perception of sound than it is to cognitive/symbolic processing. (I say "seems as though" because I believe that, in fact, other aspects of musical perception are just as intimately connected, although they are not typically treated this way). Since we have no convenient symbolic representation of timbre, it becomes difficult to work timbre into theories of music perception that are based on structural manipulation. Thus it is possible for theories that purport to discuss the large-scale perceptual organization of music to never discuss timbre at all. This is surely backwards from the way most music is perceived by most listeners in most situations.

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.

Composers of the late 20th century have frequently been 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 "gray 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).

2.2.4. Music and emotion

There is a long-standing thread in music psychology that 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, 1956). 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). He focuses instead on the notion that the affective response is mainly a modulation of the listener's overall 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 rigorous emotional-psychology definitions.

Second, Meyer denies that the affective response to music is based on designative semiotics in the sense described above. Rather, he claims that all (or nearly 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 (1997) 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. 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 which physical parameters conveyed the emotional aspects of the performance. However, Juslin must be credited for acknowledging the importance of the acoustic performance to transport and mediate the emotional material.

Important recent work by Balkwill and Thompson (in press) has examined the cross-cultural perception of emotion in music. Balkwill and Thompson elicited musical performances from musicians of subcontinental India using traditional styles and instruments. They instructed the musicians to produce certain emotional qualities in the music. The resulting stimuli were played for Western listeners with no previous exposure to this musical culture; the listeners were still able to reliably distinguish the various emotions.

The authors argue that this indicates that universally-shared *surface* percepts of music (probably based on psychophysical cues in the sound) govern the perception of emotion in music to a significant extent, since it may be assumed that listeners untrained in a musical style are unable to make useful judgements about the high-level musical structures used in that culture. (If this assumption is false, it will require a rethinking of the concept of *musical structure*, since the understanding of structure is typically conceived as requiring familiarity with a particular style). It is possible, although I do not explore this directly, that the kinds of surface features I present in Chapter 6 would be sufficient to model the results of Balkwill and Thompson and thereby explain some aspects of the perception of emotion in music.

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 interdisciplinary collaboration. It would also be a useful example of a study in individual differences in the response to music, a topic that is largely unaddressed in the

music-psychology literature. These topics are not addressed directly in my dissertation, but are left for future study.

2.2.5. Perception of musical structure

One of the primary ways in which both musicians and non-musicians understand music is that they perceive musical *structure*. I use this term to refer 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 the approach I will discuss in Chapter 7. 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 (1990a) 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 that 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 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 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 explicit 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 musical structure.

2.2.6. 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 analysts to take a more perceptual stance by including aspects of real-time thinking, retrospection, or perceptual limitations in their theories of music.

The music theorist F. Lerdahl, in collaboration with the linguist R. Jackendoff, developed an extensive theory of musical "grammar" (Lerdahl and Jackendoff, 1983). This theory has as its goal the "formal description of the musical intuitions of a listener who is experienced in a musical idiom." (p. 1) 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." (p.2) That is, the *piece* of music has a cognitive 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 that 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 contrasts 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 that 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 (1979)) 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 argued in this article 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 itself a "creative act." However, his crucial point that listening is a *behavior* (and thus can only be properly addressed with progressive, temporally-situated 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 are 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 may be sensitive to these systematic performance-time deviations; if true, such a feature might be used to distinguish the music of 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)

Finally, 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 (represented as musical scores, not as sound) 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 received great popular acclaim (and argued by Cope himself) as an example of "musical artificial intelligence." I disagree with this view. To me, Cope's research in this vein serves primarily as a valuable 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, especially, how simple and easily-fooled is the perceptual apparatus that is used to categorize music by composer. A truly 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.

2.2.7. Musical experts and novices

As a final note, it is important to emphasize 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 "naïve" 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 and reproduce vocally, 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. This is an important motivation for my research.

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

2.3. 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 that could analyze and perform music

since the dawn of the computer age; the motivation of making computers able to participate musically with humans 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 Fast Fourier Transform is useful for analyzing musical sounds as well as many other sorts of signals. However, I am especially interested in signal-processing techniques that are specifically targeted to musical signals.

I review four main areas of research in this section. *Pitch-tracking* systems are computer systems that attempt to extract pitch from sound stimuli. They differ from the pitch models in Section 2.1.1 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 that 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 *et al.*, 1998); 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 result in insights about the music perception process in humans. They articulated this view in a paper (Desain and Honing, 1994) that argued on the basis of research into "foot tapping" systems that pursuing well-organized engineering solutions to certain problems can lead to more robust formal models of music-listening.

The summary in this section covers only half of the world of musical signal processing; the other half is that devoted to sound synthesis 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).

2.3.1. Pitch-tracking

Perhaps the most-studied engineering problem in musical signal processing is *pitch-tracking* – extracting pitch from acoustic signals.[3] 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

[3] 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 2.1.1, 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.

of error measurements. They concluded that no single method was superior overall, but that different techniques were useful for different applications. 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 in Section 2.1.1, but can still be efficiently implemented and applied in practical engineering situations. Van Immerseel and Martens (1992) constructed an "auditory model-based pitch extractor" that performs a temporal analysis of the outputs emerging from an auditory filterbank, haircell model, and envelope follower. They reported real-time performance and robust results on clean, bandpass-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.

2.3.2. 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. I prefer not to use it, however, because it conflates the issue of performing the musical analysis with that of printing the score (Carter *et al.*, 1988). It is the former problem that is more related to my dissertation, 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 2.1.2, and research into sound representations, which will be discussed in Section 2.3.3. Many researchers in musical-signal-processing equate the problems of music scene analysis and polyphonic pitch-tracking. I do not; scene analysis is a perceptual problem while polyphonic pitch-tracking is an engineering problem. Conflating these approaches—by assuming that the goal of the perceptual scene-analysis system is to produce a polyphonic transcription of the input sound—has much affinity with structural models of music perception, as I will discuss in Section 3.4.

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 *et al.*, 1982); these systems were envisioned as becoming part of an "intelligent editor of digital audio" (Chafe *et al.*, 1982). This early work soon gave way to a concerted attempt to build polyphonic pitch-tracking systems, largely using acoustic piano sounds for testing and evaluation. 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 that 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 that could perform relatively unconstrained polyphonic pitch-tracking of natural musical signals. He developed new digital-signal-processing techniques that 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 sinusoids (pure sounds that vary slowly 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 that 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.

M. J. Hawley's dissertation (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 working technology, so evaluation was not crucial.

My own master's thesis (Scheirer, 1995; Scheirer, 1998b) 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.

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.* (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.

Most recently, Goto and Hayamizu (1999) have constructed computer programs that can extract the melody and bass line from real, ecological musical signals. These programs are not polyphonic pitch-tracking systems *per se*, since they do not attempt to derive symbolic representations of the extracted lines. Rather, Goto's system performs frequency-domain segmentation and continuous transformation of the input signals, in order to produce output signals with certain desired properties. He has demonstrated that the system works (although principled evaluation is very difficult) on a variety of jazz and pop-music examples. This line of inquiry must still be considered preliminary, but seems extremely promising.

The recent development of *structured audio* methods for audio coding and transmission Vercoe *et al.* (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, 1985) or SAOL (Scheirer and Vercoe, 1999) 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 arbitrary sounds. 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.

2.3.3. 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 *et al.*, 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 by Helmholtz (1885), 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 that 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* (Patterson *et al.*, 1995) is a closer approximation used in perceptual studies. These representations also have various properties of time/frequency resolution and efficiency that 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 *et al.*, 1996) compared the properties of various time/frequency representations.

Musical-signal-processing systems have often been constructed around the idea of transforming a signal into a time-frequency distribution, and then analyzing the resulting spectral peaks to arrive at a representation that 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 and Quatieri, 1986) can be used to extract the sinusoids. These techniques have also been used in a more perceptually-motivated 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 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 of 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 that Ellis made in his dissertation (1996a) was the introduction of 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. Ellis provided algorithms for extracting wefts from autocorrelograms as well as details on their use in sound-scene analysis. He also discussed the general problem of trying to discover the true perceptual representations of sound.

2.3.4. Tempo and beat-tracking models

Another commonly-approached problem in musical signal processing is *beat-tracking* or *foot-tapping*; that is, attempting the construction of systems that can find the beat in a piece of music. A fair amount of experimental data on special, non-ecological examples has been collected in studies. I am not aware, though, of any perceptual data for real music except that contained in the short validation experiment in my own work (Chapter 4) 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 is therefore crucial in understanding of how listeners orient themselves to a new musical stimulus.

There is a large body of work originating in the music-psychology community that attempts to group musical *onsets* together into a rhythmic context. Such models subsume multiple onsets separated in time into a rhythmic clock, "hierarchy", grouping, or oscillatory framework.

Povel and Essens (1985) presented research on the association of "internal clocks" with temporal onset signals. They described an algorithm that could, given a sequence of inter-onset intervals as input, identify the clock a listener would associate with it. Their research was particularly concerned with the way that perceived accents lead to the internal clock. Although obviously related to music, their research purports to examine time intervals in general rather than being restricted to musical stimuli. Parncutt's recent work (Parncutt, 1994a) extends this type of model to include many aspects of the structural characteristics of temporal sequences, such as duration and phenomenal accent.

Desain and Honing have contributed many results to the computational modeling of beat-tracking. Their models (Desain and Honing, 1992a; Desain, 1995) typically also begin with inter-onset intervals and associate a rhythmic pulse with the interval stream. However, unlike the Povel/Essens and Parncutt models, these models are *process models*—they process the input sequentially rather than all-at-once. This is an essential aspect of a model of human rhythmic perception. Desain's "(de)composable" model calculates rhythmic expectations due to each of the possible inter-onset times in a rhythmic stream, and sums them to create an overall rhythmic expectation. (Note the similarity here to pitch models based on the histogram of interspike intervals, as discussed at the end of Section 2.1.1).

Large and Kolen have described a beat-tracking model (Large and Kolen, 1994) based on non-linear oscillators. Their model takes a stream of onsets as input, and uses a gradient-descent method to continually update the period and phase of an oscillator. In this manner, the oscillator phase-locks to the input stream. The resulting oscillation process seems to be a good match for the human perception of beat.

Longuet-Higgens and Lee have written several papers (for example, Longuet-Higgens and Lee, 1984) on the induction of rhythmic hierarchies from monophonic time sequences. They are more interested in the development of theories describing the relationship of rhythm, meter, and phrasing than on the boot-strapping process in which tempo and beat percepts arise. Tempo perception may be viewed as underlying their models.

These approaches, and several others (Rosenthal, 1992; Brown, 1993), require that robust onset-detection be a step preceding beat analysis. This entails important restrictions to their applicability. The models cannot operate on acoustic signals, but must be provided with pre-processed symbolic data such as event lists or MIDI. The extraction of onsets from multitimbral, polyphonic music is itself a difficult problem, and one that we have little perceptual data about. Thus, relying on event lists is a serious restriction of any model that claims to treat human rhythm perception. There has been little attempt to merge these sorts of models with real-time acoustic pattern recognition to allow them to work with acoustic data.

More recently, there has been some research attempting to extract rhythm and/or pulse information directly from acoustic signals. Goto has demonstrated a system that combines both low-level signal processing and high-level pattern-matching and "agent-based"

representations to beat-track and find rhythmic groupings in popular music (Goto and Muraoka, 1998). His method extracts drum patterns from a signal and uses a template-matching model to determine the beat from the drum track. This system runs in real-time on a parallel-processing computer and was used to control interactive-graphics displays from ecological music signals.

A further extension to this system used more sophisticated signal-processing (Goto, 1999) to extract onset times from signals without drums and make judgments similar to those made by the earlier system. It used a robust onset-detector and chord-change locator as the fundamental elements of bottom-up processing. Goto reported excellent results for both of these systems on rock-and-roll music, although it is not clear whether his methods were applicable to other musical genres or to signals with changing tempo.

N.P. Todd's work (Todd, 1994) has described algorithms that detect onsets in monophonic music under certain timbral constraints, and group them in a rhythmic framework using a multi-scale smoothing model. The onset model used is a simple one based on leaky integration. The resulting *rhythmogram* representation conceives of pulse, and in some cases, meter and phrase, perception as a very low-level process arising directly from the time- and loudness-integration properties of the auditory periphery. The model as presented can be implemented in an incremental manner, but Todd only tested it using toy examples (although, interestingly, a speech example was included).

All of the above-mentioned research uses what I have previously described as a *transcriptive* metaphor for analysis (Scheirer, 1996). That is, the music is first segmented, or assumed to already be segmented, into notes, onsets, timbres, and so forth. Post-processing algorithms are then used to group rhythms and track beats. As high-quality polyphonic music transcription algorithms are still years in the future—the state-of-the-art systems cannot transcribe pieces more complex than four-voice piano music, as discussed in Section 2.3.1—it seems logical for practical reasons to attempt to construct systems that can arrive at a musical understanding of a piece of music without going through a transcription step. Further, as the validity of the transcriptive metaphor as a framework for music perception has been challenged, it is scientifically appropriate as well.

Vercoe (1997) reported a system for beat-tracking acoustic music that 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. This was the first system to be reported that did not try to detect onsets robustly; the output of Vercoe's system is a continuous transformation of the input.

Beat-tracking is of great interest in the construction of general music perception systems; it has been implicated, especially in the work of M. R. Jones and her colleagues (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.

2.3.5. 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. Techniques from the broader pattern-recognition literature can often be leveraged quite effectively to produce solutions to problems in sound classification when they are considered only as engineering tasks.

I have conducted one well-organized study in sound-signal classification; M. Slaney and I built a system that 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 perception" 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 that 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.

In the last few years, this sort of research has gained momentum, as it has been discovered that for segmentation of broadcast video such as films or news programs, the soundtrack is often more useful than the video images. Several recent papers have demonstrated the use of pattern-recognition methods applied to the soundtrack in order to perform tasks such as scene-change detection (Liu *et al.*, 1998) and other video-structuring tasks (Minami *et al.*, 1998).

Smith *et al.* (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 that 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, 1999) 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 primitives and sentiments." The report of their system was still highly speculative. As music is an important vehicle of emotional communication between humans, the development of truly "affective" computer systems (Picard, 1997) requires more research of this sort.

2.4. Recent cross-disciplinary approaches

In this section, I examine some recent studies that report attempts to build 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 implementation 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 that 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 human 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." Again, the chords provided to the human listeners were synthetic, not ecological, and the model was provided with symbolic lists of components, not sounds, as the input.

Parncutt has continued to study models of 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. Neither Parncutt nor Terhardt have yet attempted to apply their models to real musical signals.

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 that operated on a time-frequency "cochleagram" image. By using this technique, solutions to problems in sound understanding can make use of techniques developed in the image-processing literature.

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 2.3.2 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 plane. Also notable was his 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 acoustic signals and, using the Kohonen-map framework, allowing a self-organizing architecture to represent tonality in music. His system consisted of two parts with interchangeable mechanisms for each:

1. An auditory model, which incorporated either the Terhardt model for pitch analysis or a cochlear filterbank process based on the model of Van Immerseel and Martens (1992) and a short-term within-channel autocorrelation (which he called 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 (Kohenen, 1995). 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 occurred. Certain structural equivalencies to theories of tonal music perception as described in Section 2.2.1 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 (see the next Chapter) differs 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 surely depends on our ability to segregate the sound scene into multiple sources and attend to them holistically or selectively. One of the important components of my research is a new model of musical source grouping using the autocorrelogram framework (Chapter 5). Leman and I share the goal of building what he terms *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 is based on signal-processing and pattern-recognition techniques. Finally, I 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 will show in Chapter 7 that my models 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 that 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.

2.5. Chapter summary

This chapter has covered a wide range of territory in reviewing pertinent research on music perception, psychoacoustics, and musical signal processing. Four threads and cross-discipline trends present clear starting points for my approach and the research I will present in the remainder of the dissertation.

The first important trend is what I term the *transcriptive* model of music processing. Both in music signal-processing and in psychological studies, researchers commonly embrace the viewpoint that first the signal is transcribed into a list of notes, and then the interesting work happens. This is true in the construction of automatic polyphonic transcription systems, which are often motivated as the means to some larger signal-processing end, and also (often implicitly) in the development of purported theories of music perception that are actually constructed around processing of the score. In contrast to this approach, I will present in Chapter 3 and use throughout the technical parts of my dissertation a new approach that I term *understanding without separation*. In this model, perceptual models, and useful music-analysis techniques, are developed as continuous signal-processing transformations from input to output.

The second important trend is the use of restricted musical domains for music research and analysis. Most previous studies have admitted only certain musical styles (two- or four-part music, music with restricted timbres, music only in the Western classical tradition) as candidates for explanation. In contrast, the experimental and computational work that I will present admits any signal that some human listener might consider music. In Chapter 7, I develop a stimulus set for model evaluation by randomly sampling a popular online database. This allows me to argue that the results I show are applicable to the entire (large) database, which has no particular restrictions on the types of content included.

The third important trend is the discussion of music as independent of sound, and sound as disconnected from musical (and other high-level) concerns. There are many computer-music studies, and many music-perception studies, that don't really treat music at all, but rather treat symbol-manipulation with a claim that doing so is representative of useful, or perceptual, musical processing. In the other direction, many (perhaps most) psychoacoustic studies use test sounds that are so simplified that it is difficult to tell what implications, if any, the results have on the perception of actual real-world sounds. Some of this separation of focus is natural as part of the reductionist approach to scientific inquiry, but ultimately, if theories of music perception are to be well-grounded, they must rest on principles of sound processing in the auditory system.

The fourth and final trend is a focus on non-ecological tasks and listening situations. Judgements that are typically elicited in psychoacoustic and music-perception tests ("is sound A higher or lower in pitch than sound B?" and "does note X serve as a good continuation for phrase Y?", respectively) have little to do with the everyday sorts of judgments made by real listeners. Similarly, the overwhelming focus in music signal-processing on pitch-tracking and polyphonic transcription is far removed from the kinds of practical problems that people naturally solve when listening to music. Further, I claim that the kinds of problems people solve naturally are actually *easier* to model computationally that are these artificial problems.

It might be argued that in the first case, the experiments actually serve to reveal preconscious perceptions that underlie more complex behaviors. However, at the very least such experiments and computational models should be augmented with the study of percepts that more closely follow the real behaviors of human listeners. This is the approach that I will present and follow in the rest of the dissertation.

CHAPTER 3 APPROACH

In this chapter, I will discuss my approach to the construction of music listening systems, and the theoretical basis of this approach in music psychology and psychoacoustics. I will begin by formally defining terms, in order to be as clear as possible about exactly the sorts of human behaviors I wish to model. In Section 3.2, I will present a definition of the perceptual problem to be solved: that of understanding the properties of the *musical surface*. Following that, I will discuss the use of computer modeling in the creation of psychological and psychophysical theories. Finally, in Section 3.4, I outline the basic theoretical stance of my approach, which I term *understanding without separation*, and compare it with other theories of psychophysical processing in the literature.

3.1. Definitions

It is crucial for a formal theory of perception to rest on careful definitions of terms. Especially in research areas that are relatively young (auditory scene analysis and musical signal processing), the use of terminology in the literature is not entirely consistent. This is particularly true in the literature on the perception of mixtures of sounds. I will attempt to define my terms carefully in this section and to follow these definitions throughout the rest of the dissertation.

3.1.1. The auditory stimulus

The first principle on which all other definitions rest is to take as given the *auditory stimulus*. Garner (1978) writes "in the beginning is a stimulus." I define the auditory stimulus to be *the complete mixture of sounds presented to the listener's ear*. The stimulus is not differentiated in any way *a priori*. When multiple sounds, each of which might be stimuli in its own right, are mixed, the result is still only a single stimulus. It is never right to say that multiple auditory stimuli are simultaneously presented; to say this begs many important questions about the listener's perception of the multiplicity of the sound. Regardless of how the sound mixture was created, a listener only ever receives a single auditory stimulus[4].

[4] Naturally, it is the case that most human listeners have two ears and perceive two sound signals at once. However, I prefer to consider the stereo percept a single stimulus, not

Following Gibson (1986), I emphasize that a stimulus is a *sound*, not the object in the world that makes it. A violin is not a stimulus; rather, a violin is a physical object that makes sounds that might be mixed with other sounds (through physical superposition). The mixed sound that impinges on the listener's eardrum is the stimulus. It may or may not be the case that from the sound stimulus, the listener can determine that there is a violin in the world. This is purely an empirical question of psychology.

A stimulus taken as a whole may be casually said to fall along a continuum from *simple* to *complex*. I won't use these terms rigorously. A simple stimulus is one that is not perceived to change very much over time, and that is perceived to have few components, each with stable perceptual attributes. A complex stimulus is one that is perceived to have many components, with rapidly changing attributes, and where the perceived organization of the attributes and the relationship between them is also perceived to be rapidly changing. In this way of speaking, then, a single pure tone is a very simple stimulus, as are most of the sorts of stimuli (synthesized vowels, glide tones, bandpassed noise) used in traditional psychoacoustic experiments. The sound of a single talker in a quiet environment is a moderately complex stimulus. A complete piece of music heard in a concert hall is usually a very complex stimulus, as is the stimulus presented during a walk around the city streets.

I will term a short, undifferentiated sound stimulus a *sound event*. A sound event is anything that is perceived in context as "the sound of one thing happening." The context is important; in one setting, a single note on a trumpet is considered an event; in another, an entire piece of music is considered an event. ("Suddenly, there was a burst of music from the radio.") A stimulus may be said to contain several sound events, but this is a casual use. When I speak more rigorously, I will use the term *auditory image* as discussed below.

I distinguish *natural*, or *ecological*, stimuli, from *artificial* or *laboratory* stimuli. A stimulus is natural if it is the kind of sound that is part of our everyday listening experience; it is any sound that a listener might hear outside the laboratory. This use of natural extends the simple definition "occurring in nature" to include musical sounds as they are found in the concert hall or on compact disc. A stimulus is artificial if it is not the kind of sound that would be heard outside the laboratory. This distinction is a continuum to some extent; certain synthesized sounds (such as those created with sampling synthesis) are laboratory creations intended to achieve part, but not all, of the complexity of ecological sound.

When I call Gould's recording of the *Well-Tempered Clavier* an ecological stimulus, I don't mean to say that you might happen to hear it walking through the woods! Rather, I mean that this stimulus has not been simplified to assist processing or experimentation in the same way that a clean performance synthesized from sampled sounds has been. A typical ecological recording of music has reverberation, noise, extraneous sounds (Gould sings as he plays, and his vocalizations are clearly audible to the listener in the recording), equalization, compression, and other kinds of "artifacts" that add richness for human listeners but are often difficult for machines to process. A major goal of my dissertation is to present final results only on complex ecological musical sounds, rather than on test sounds or on stimuli that have been specially prepared for easier processing. When I want to make the further distinction between "sounds occurring in the real world" and musical sounds, I will use the term *environmental* sounds to refer to the former. Thus, two important sorts of ecological sounds are environmental sounds and musical recordings and performances.

This dissertation is predominantly concerned with complex ecological musical stimuli. I will use laboratory stimuli only as examples and to evaluate psychoacoustic theories (particularly

---

two stimuli, because there are still empirical questions regarding how to relate the cross-channel properties of the two signals to the perception of the sound. The kinds of musical behaviors I am considering in this dissertation are easily enabled by monophonic signals as well as stereophonic ones. I do not consider stereo processing directly here.

in Chapter 5), not to evaluate the listening systems constructed as the larger goal. Environmental stimuli will only be considered in thought-experiments.

3.1.2. Properties, attributes and features of the auditory stimulus

Following Licklider (1951a), Terhardt (1991), and others who haven written with rigorous terminology about the hearing process, I will maintain a clear distinction between *physical properties* and *perceptual attributes* of sounds. However, since I am exploring sounds that are more complex than are typically considered elsewhere, I must define the terms for the perceived aspects of complex sounds especially carefully.

The physical properties of sounds are those that can be measured directly using scientific instruments. The perceptual attributes of sounds are those that a human listener associates with the sound. For some perceived attributes of some sounds, it is relatively easy to understand the physical *correlates* of the attributes, by which I mean the physical properties that lead to the particular percept. For example, as the frequency of a sine tone varies, the pitch of the sound varies in a simple way. However, for other sounds and other properties, the correlation between the physical and the perceptual is more difficult to understand. For example, the physical correlate (or set of correlates) for the perceived timbre of a sound is not known today, and there is an extensive literature exploring this topic.

I will now list and define some of the most commonly-studied perceptual properties of test sound and present theories of their physical correlates. This sets the stage for the discussion of the perceptual properties of complex sounds in the next section.

Pitch

Many sounds are perceived to have a *pitch*. The *pitch* of a sound is that attribute by which it may be positioned on a continuum from "low" to "high." I prefer to adhere to the strictest operational definition (Stevens, 1975, p. 230): The pitch of a sound is defined as the frequency of a sine-tone that is matched to the target sound in a psychophysical experiment (this is termed the "method of adjustment"). It is not often clearly enough said, even by authors who should know better, that this operational definition severely bounds the ways in which the term can be accurately used. For example, it is not the case that sounds have a "correct" pitch that listeners can be asked to "identify." The pitch of a sound is simply what a particular listener, in a particular experimental context, says that it is. Different listeners may attribute different pitches to the same stimulus; the same listener may attribute different pitches at different times.

Also, it is not possible for a stimulus to have more than one pitch. The question of the pitch (and other perceived attributes) of mixtures of sounds is something I will consider later in this section, and in fact throughout the dissertation. Given an operational definition of pitch, all we might say is that pitch judgments for simple stimuli have a lower variance (inter-listener, inter-trial) than pitch judgments for complex ones. When complex sounds are presented to listeners as stimuli in operational pitch-judgment trials, the result is a probabilistic histogram of "pitch likelihood." If a pitch must be associated with a complex sound, I would rather call the entire probability density function the pitch than to impart multiple pitches to a single stimulus. (Also see the discussion of "auditory images" below).

This operational definition also entails that the same physical stimulus may be perceived to have different pitches to different listeners, or to a single listener at different times. There is nothing wrong with this—it is simply in the nature of some sounds to be ambiguous in this way. For example, consider the sounds shown in Figure 3-1. The first is unambiguously a complex tone with pitch of F_0. The second is unambiguously a complex tone with pitch of $2F_0$. Thus, if the amplitude of the low partial is ramped so that it slowly fades in (as in the third stimulus), at some point the stimulus changes from a $2F_0$ to an F_0 pitch (the exact time and nature of the change depends on the speed of the cross-fade and other factors that are not

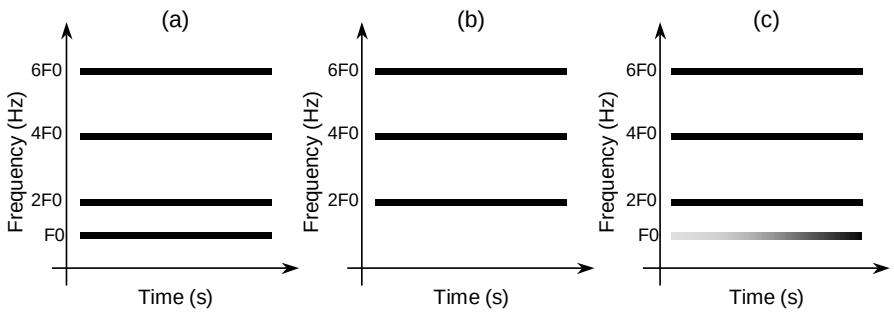


Figure 3-1: Three sounds that demonstrate the ease of constructing stimuli that are perceived to have ambiguous pitch. Each panel shows a schematic of a spectrogram representation, in which the intensity of a point on the time-frequency plane indicates the amount of sound energy at that frequency at that time. In (a) is a sound that is usually perceived to have pitch F_0—it is a four-component harmonic series that makes up a complex tone with two overtones missing. In (b) is a sound that is usually perceived to have pitch at $2F_0$—it is the first three harmonics of the fundamental $2F_0$. In (c) is the sound creating by cross-fading from (a) to (b)—the three upper components are steady, but the low component at F_0 fades in gradually. At the beginning of stimulus (c), the sound is instantaneously the same as (b), while at the end, it is the same (a). Thus, the percept changes discontinuously from having pitch at F_0 to having pitch at $2F_0$. Even given the simplicity of such a stimulus, little is known about the exact manner of the perceptual shift, the way on which it depends on the length of the stimulus in time, and how it varies across listeners.

important for this discussion). There is an intermediate ambiguous stimulus in which some listeners will judge the pitch as $2F_0$ and some as F_0.

There is clearly no "right" model for the pitch during the time in which it is ambiguous. It is an empirical question, relatively unexplored, to what degree such ambiguously pitched stimuli occur naturally. It is another one, also little-studied, how they are perceived by listeners. It is known that some speech sounds (particularly the glottal stimulation known as "creaky voice" (Ladefoged, 1982, p. 226)) are ambiguously pitched, which is a source of great difficulty in building and evaluating pitch-estimating algorithms for voiced speech (Hermes, 1993).

Finally, some authors have written about the concept of "pitchiness" in discussing the way in which it is easy or difficult to assign a pitch to a stimulus. A very pitchy sound, like a harmonic complex tone, has a clear, unambiguous, and immediately-perceived pitch. For a less-pitchy sound, like that of repetition noise, the percept is not as immediate and seems phenomenally weaker in strength. There is no work of which I'm aware that has studied the pitchiness of sounds empirically or has tried to define this concept rigorously.

The physical correlate of pitch is often *frequency*. Frequency is a physical property of a sound that can be measured with scientific equipment. As the frequency of a sound changes, we can see through experiment that the pitch of the sound is perceived to change in predictable ways. However, we cannot *measure* the pitch directly as we can frequency; we can only *model* the pitch by creating mappings from properties that we *can* measure to experimentally-derived estimates of pitch. This is what it means for pitch to be a perceptual quality, and frequency to be a physical quantity.

We may consider the perceived properties of sounds to create equivalence classes among sounds; that is, there is (to a given listener) a set of sounds that all have a particular pitch, a set of sounds that all have a particular tempo, and so forth. The problem of determining the mapping from physical properties to perceived qualities then becomes one of determining aspects of sound that are invariant within each equivalence class. Gibson, in developing the field of ecological psychology, wrote extensively on the role of physical invariants in visual perception (Gibson, 1986).

The strangeness of the set of sounds that all evoke a particular pitch, as listed by Zwicker and Fastl (1990, p. 125), is evidence that the processing performed by the auditory system must be quite complex. Such a wide variety of sounds leads naturally to the hypothesis that there is no simple (or perhaps even single) method for predicting the pitch of a sound given its physical attributes. Further, it suggests that any accurate computational model of pitch is necessarily going to be messy, hard-to-understand, and full of special cases.

Loudness

The *loudness* of a sound is that perceptual attribute that allows it to be positioned on the continuum from "soft" to "loud." Via the method of adjustment, loudness can be defined operationally as the amount of energy in a reference sine tone (or wideband noise) that is adjusted to be equally loud as a given stimulus. Loudness is typically correlated with the physical property *power*. The conceptual relationship between loudness and power is very similar to that between pitch and frequency. Just as with pitch, each sound has only one loudness (or one loudness probability distribution) and this loudness is not a veridical property of the sound.

Pitch and loudness are the two most well-studied attributes of sounds. There are thousands of papers in the psychoacoustic literature treating each, and robust models for predicting them from physical measurements of sounds, at least for moderately simple stimuli. As we go farther afield to consider properties that are less rigorously grounded in psychophysical experiment, the terminology becomes less standard and used rather more irregularly. There is little known about the effect of cognitive factors on pitch or loudness judgments. In one notable series of studies, Fucci *et al.* (Fucci *et al.*, 1993; Fucci *et al.*, 1996) found that listeners who disliked rock music reported it to be consistently louder than subjects who liked it.

Timbre

Following Martin (1999), I will define the *timbre* of a sound to be the quality or set of qualities that allows a listener to identify the physical source of a sound. I will not use the word in any more rigorous way than this. *Timbre* is a justly-maligned term that is used in different contexts to mean different things. There is no simple set of physical properties that corresponds to timbre, and no clear operational definition. It is likely that the set of properties used to determine sound-source identity differs in different circumstances (Martin, 1999). Timbre is a simpler concept to understand for simple sounds (isolated tones, for example), than for complex sounds and mixtures of sounds. Given a definition of *timbre* that, like this one, rests upon source identification, it is interesting to say that the qualities that allow us to decide that one sound is a rock-and-roll power trio, while another is a symphony orchestra, should be considered the timbre of the power trio and the symphony orchestra respectively.

There is a long-standing hypothesis (Grey, 1977) that the physical properties corresponding to the timbre of musical instruments can be identified through the psychological paradigm called *multidimensional scaling*.[5] Numerous studies have been done that purport to identify such physical properties. (It is important to recognize that this is only a hypothesis, not part of the definition of timbre). However, such studies have generally not undertaken the construction of computer systems that measure the resulting qualities from sounds and use them as the

[5] Grey writes: "[The researcher] may start with the perceptual judgments of similarity among a diverse set of (naturalistic) stimuli, and then explore the various factors which contributed to the subjective distance relationships. These factors may be physical parameters of the stimuli, which then would lead to a psychophysical model; yet, multidimensional scaling techniques may also uncover any other factors involved in judgment strategies."

basis for automatic classification of timbre. This, to me, is the key scientific step in such modeling work; without it, all that is presented is a post-hoc evaluation of individual results on small experiments.

The testable prediction that is made (often implicitly) by such a research model is that it is these *particular* properties that are really used by a listener to identify objects from their sounds. It is incumbent upon those researchers who wish to assert the continued utility of the multidimensional-scaling paradigm for timbre research to conduct such computational studies to confirm that these properties contain sufficient information to support the behaviors imputed to them.

In point of fact, computational systems that are capable of classifying instrument identities with accuracy comparable to that of humans (see the review in Section 6.6 of Martin, 1999) have *not* used the sorts of physical properties of sounds that come from multidimensional scaling research. To me, this calls into question the continued utility of this research paradigm for anything other than exploratory studies.

Tempo

The *tempo* of a sound is the perceptual sense that the sound is recurrent in time at regular intervals, where the interval length is between about 250 ms and 2 s. A good operational definition of tempo would be: the frequency of a click-track adjusted to have the same perceived speed as the stimulus. I am not aware of any experiments that use the method of adjustment for evaluating the perceived tempo of stimuli. Like pitch and loudness, tempo is a perceptual attribute that cannot be measured directly from a sound. A sound does not have a "real tempo" that a listener might judge "incorrectly." The tempo of a sound to a listener is just whatever the listener thinks it is. It is not presently well-understood what the physical correlate of tempo is. For simple stimuli such as click trains, the tempo corresponds to the frequency of clicks. However, listeners also perceive tempo in signals that have little overall change in amplitude envelopes, such as music that has been dynamic-range compressed for radio airplay. I have developed a new model for the perceptual processing of tempo from physical properties of the sound stimulus; it is presented in Chapter 4.

Other attributes

There are a wide variety of perceptual attributes that have neither a clear operational definition nor known physical correlates. It is easy to conduct experiments in which listeners are asked to rate sounds on many perceptual scales (for example: *complexity*, ranging from "more complex" to "less complex"; *pleasantness*, ranging from "pleasing" to "displeasing"; *familiarity*, ranging from "familiar" to "less familiar"). In general, consistent and interpretable results can be obtained through this experimental paradigm (see Chapter 7). But it is very difficult without more detailed hypotheses to map these attributes back to their origins in physical attributes of the sound, and seemingly impossible to conduct operational studies via the method of adjustment. To do the latter would require a set of reference sounds *defining* the range of complexity, pleasantness, etc.

However, just because we don't know the perceived physical correlates of a particular quality doesn't mean that there aren't any. Quite the opposite, in fact: formally, any equivalence relation defines sets of stimuli for which all members are equal under that relation. Given a certain "amount" of complexity c, for a particular listener at a particular time there exists a set of stimuli C such that all the stimuli in the set C have complexity c. Whatever physical properties can be used to discriminate sounds belonging to C from those not belonging to C

must be philosophically considered to be the perceptual correlates of the percept of "having complexity *c*."[6]

One of my primary goals in this dissertation is to open a line of inquiry that considers these "extra" sorts of perceptual attributes more rigorously. This is an important contribution to the ethological study of human listening behavior, because it is *these* sorts of judgments that are most frequently made by actual listeners engaged in actual listening to actual music. Studying such attributes forms an important connection between the low-level perception of sound, in which important properties are extracted from the signal but often unavailable to the conscious mind, and the high-level cognitive processing of sound in the way it is most often embodied in naturalistic human behavior. These issues will be considered in more depth in Chapter 6.

3.1.3. Mixtures of sounds

When a complex stimulus is heard, it is naturally and automatically perceived as the union of a set of perceptual constituents. That is, the listener perceives that the sound is not emanating all from one source, but originates from many different sources. This is the case both when this percept is veridical, as when we walk along the street, and when it is not, as when we listen to the radio (on the radio, all of the sound is originating from the speaker; there is really only one sound source present.) Following McAdams (1983) and Yost (1991), I define *auditory image* to mean the percept that corresponds to one of these perceived sources. That is, a complex stimulus is perceptually segmented into a number of auditory images. There is relatively little known about the manner in which this happens; this question is a primary topic of the present dissertation. There is also relatively little known about the physical properties that cause (or allow) the segmentation to occur. The key perceptual aspect of an auditory image is that the listener can imagine hearing it in isolation, separated from the rest of the mixture.

A note in passing for readers already familiar with the literature on this topic: The term *auditory image* is used differently by Patterson (Patterson *et al.*, 1992), who uses it to refer to a particular processing model for sound. Patterson's Auditory Image Model is a variant of the autocorrelogram processing model presented in Section 2.1.1. I believe that the term *auditory object*, which is sometimes where I use *auditory image*, is misleading and unsuitable because it conflates the physical aspect of sound-producing objects in the world with the perceptual aspect of the perceived images. If the term *object* is used for both, the perceptual-physical distinction is diminished.

In natural sound environments, our hearing system seems remarkably adept at deriving the veridical segmentation of sounds. That is, when we are in a location where there are multiple sounding objects, we seem generally able to associate one auditory image with each object.[7] This relationship is deeper than the one that mediates the physical properties of sound and

---

[6] This argument holds only so long as perceived complexity is itself a well-defined equivalence relation on stimuli. If it is not (for example, if a listener judges sound *A* and sound *B* to be equally complex, and also sound *B* and sound *C* to be equally complex, but not for sound *A* and sound *C*), then naturally there are not necessarily coherent physical correlates. On the other hand, there is no necessity for complexity as measured on a perceptual task by a particular listener to correspond neatly with some analytic conception of complexity; the relation described here still holds as long as the perceptual judgment is consistent.

[7] This is not to say that our perception of what objects exist in the world is static and fixed. Depending on context and attentional stance, we can choose to hear "a car" or "the wheels and the engine and the brakes and the muffler." I only mean to say that as we choose *some* segmentation of the world, we can generally associate sounds with objects in the way that matches the physical sound generation mechanisms.

their perceived attributes. It relates the *interpretation* of the perceived attributes to the actual, physical world of objects. Theories such as Gibson's "ecological perception" (Gibson, 1986) are concerned with this relationship; see also Casey (1998) for one computational approach and Windsor (1995) for an aesthetic approach to understanding this relationship.

It is unsurprising that this problem is solved effectively by the human auditory system, because this is the environmental characteristic that has most directly influenced its evolution. The human auditory system, in its processing capabilities, reflects invariants in the real world of environmental sounds and uses them to perceive such scenes veridically whenever possible.

In contrast, when listening to complex musical stimuli, we are much less able to understand the actual world of objects from the sound. A symphony orchestra may be made up of dozens of instruments making sound simultaneously, yet it seems that we perceive no more than a half-dozen auditory images. Different listeners have different abilities to "hear out" the instruments from a mixture, but no one (I submit) has this ability to such an extreme as to be able to perceive the 24 different violins in the violin section as separate and distinct entities. The relationship in music between the veridical world, the physical properties of the stimulus, and the music as perceived is a very complex one.

Another way of stating the two preceding paragraphs is that it seems that complex environmental stimuli have a particular kind of *perceptual linearity* that is not shared by musical stimuli. That is, up to reasonable limits, the percept that corresponds to the sum of two environmental sounds is the union of the percepts of the parts. In contrast, musical stimuli often fuse together in complicated ways, such that two sounds, each of which would is perceived in isolation to contain one auditory image, may still only have one auditory image when they are mixed together. Both for environment sounds and for musical sounds, it is an empirical/experimental question how many auditory images are in the sum of a set of sounds, given the physical properties of the individual sounds in the set. Like other perceived attributes, it is logically impossible for the number and properties of auditory images in a complex stimulus to be judged "incorrectly"—only the perception of the listener matters.

A complex stimulus that is perceived to have a number of auditory images is often termed an *auditory scene*. The perceptual process applied to an auditory scene, particularly with regard to determining the perceived number and nature of auditory images, is termed *auditory scene analysis* (Bregman, 1990). I will sometimes refer to the complex musical stimulus as a *musical scene* in contexts when I am discussing auditory scene analysis as applied to music.

3.1.4. Attributes of mixtures

Individual auditory images in a mixture may be perceived to possess perceptual attributes such as pitch. What this means is the following. A listener hears a complex auditory scene and perceives that it contains several auditory images. She is able to imagine that she is hearing one of the auditory images in isolation, and is able to consistently associate a particular reference signal with this imaged stimulus. The pitch of the associated reference signal may then be termed the pitch of the auditory image. In such a case, the processing relationship between the perceptual attribute and the physical stimulus may be very complex indeed. As well as pitch, the other qualities of simple stimuli discussed above might also be associated with individual auditory images.

It is not always the case that a sound possessing certain perceptual attributes when presented in isolation will have the same attributes when presented as part of a mixture. That is, even if a sound is perceived to have pitch *p* on its own, when the sound is mixed with others, the auditory image corresponding to the sound (assuming that there is one) might have some pitch different than *p*. It may also be the case that a sound possesses some property like pitch in isolation, but then when the sound is presented as part of a mixture, the listener is not longer

able to consistently match the image corresponding to that sound to a reference signal at all. In this case, I would say that the pitch has been lost.

A particularly strong view of sound-source perception (Yost, 1991) argues that the fundamental purpose of auditory perception is to determine the attributes of auditory images, and impart them to objects in the world. For each auditory image, the listener imagines that it is separated from the mixture, and associates perceptual attributes with the imagined source-in-isolation, and thereby determines the number of objects in the world and their identities. Yost does not extend this viewpoint to musical scenes; to do so might be problematic, since many of the auditory images in a musical scene do not correspond directly to any particular object in the world.

At the risk of being repetitive, the judgment of a listener is not open to criticism or evaluation. It is not the case that the perceived attributes that derive from the imagined source are "correct" or "incorrect" based on any criteria whatsoever. It may be the case that the isolated source is imagined to have the same attributes as the simple source did before it became part of the mixture, or it may not be the case. This does not make the perception "wrong"—the perception is what it is, regardless of the separate physical reality of the sound-producing objects. For example, it may be the case that a particular listener is unable to hear out the notes of a chord. This listener is therefore unable to imagine these notes in isolation and is unable to associate perceptual qualities with them. The goal of a scientist studying the perception of music must be to understand the implications of these inabilities in the subject's judgments and behaviors when presented with musical stimuli. It is not appropriate to dismiss such a listener as being "unable" to make the "correct" or "educated" judgments about music.

A valuable contrast to make when discussing properties of mixtures was presented by Garner (1978). He distinguished a *simple holistic attribute* from a *configuration attribute*. In this terminology, a simple holistic attribute is an attribute of a mixture that is in some sense a sum of the perceptual components of the mixture. For example, the number of auditory images in a mixture, or the list of pitches in a melody, is a simple holistic attribute of these stimuli. A configuration attribute is an attribute that does not depend on the exact properties of the components, but rather the relationship between them: "[T]he whole is indeed more than, at least other than, the sum of the parts...." (Garner, 1978, p. 123).

An attribute like tempo is clearly of this nature. No individual sound event has a tempo; it is only through the juxtaposition of many events into a musical scene that tempo arises (although I do not mean to assert that tempo is *perceived* as a feature through the juxtaposition of perceived objects, only that tempo does not *arise* until there is more than one event in a musical scene). The question of the degree to which important musical percepts, such as the percept corresponding to the physical stimulus of a "chord," are holistic and to which they are configurational, is relatively unexplored. This topic is also put aside for future research.

3.1.5. The perceived qualities of music

In the preceding several pages, the kinds of properties that I have discussed are those that have been primarily studied by psychoacousticians. This sort of study has the advantage of being extremely rigorous with regard to acoustic properties of the stimuli, and the disadvantage that these properties are still rather far from our everyday phenomenal experience of listening to music. When we listen to music, the perceptions of notes, chords, melodies, harmonies, rhythms, and emotions are much more important (at least consciously) than are perceptions of pitches, loudnesses, and timbres.

It is even more important, as I begin to discuss music, to keep in mind the difference between the physical and the perceptual. This is especially true because for music there exist several other levels of interpretation. In addition to the musical objects in the world producing sound waves in the air, and the phenomenal percepts of music in the mind, music can be considered

as a sequence of symbols that a composer wrote upon a page (see Section 3.4), or as a mythopoetic object with historical effects and aesthetic implications, or as the particular set of physical actions that a performer must execute in order to create it. These are all different levels of interpretation and must not be confused.

Consider a musical chord; that is, several notes played at the same time. On the page, the representation of a chord is unproblematic; it is a graphical figure containing symbols that represent notes for a performer to play (Figure 3-2).

A music theorist may look at the chord and interpret it in some particular way. For example, that it is a C dominant chord, a musical tension that could be resolved by an F major chord. (Another theorist might interpret it as a C_7 chord, the home chord of a blues in C major). When a performer plays this chord, perhaps on a piano, the representation is still unproblematic—we can use acoustical equipment to measure the physical properties of the resulting sound in a variety of ways. It can be very difficult to do this measurement accurately, and there will be disagreements between acoustical physicists regarding the proper measuring techniques and the implications of particular results, but the physical signal exists and can be measured.

When we consider the perception of the chord by a human listener, though, it is imperative to remember that this demands a different interpretative stance than do the music-theoretical or physical manifestations of the chord. As perceptual scientists, we do not know what the perceptual analogue of the chord is until we have devised experimental means of testing hypotheses about it. We do not know that the perceived chord is a single perceptual entity; we do not know that it is multiple entities. We do not know the nature and dimensions of the features of the perceived chord. We do not know if the percept is the same for all listeners, or for most listeners, or for some interesting set of listeners, or if the percept differs in some important way from one listener to another.

An important interpretative stance by psychophysicists was the development of the two parallel vocabularies for the physical world and the perceptual world: frequency-pitch, power-loudness, spectrum-timbre (even though the last is not strictly believed to be true any more). It is unfortunate that we do not have such parallel terms for the immanent components of music. The term *chord* is used variously to refer to the notation on the page, the action by the performer, the analytic object of artistic interest, the acoustical signal, and the perception that corresponds to this acoustical signal. To have a name for the chord-as-perceived, melody-as-perceived, and cadence-as-perceived would make the distinction between the poetic, physical, and perceptual interpretations more clear. Perhaps some of the great confusion regarding the relationship between music perception and music theory (Clarke, 1986) could have been avoided altogether.

3.2. The musical surface

What is the surface structure of music as it is perceived by the human listener? When we consider listening to music as a complex perceptual/cognitive behavior, it is clear that there are many interesting musical behaviors that seem to stem from the ability of the perception system to rapidly analyze, segment, and process an incoming musical stimulus. Therefore, it is unfortunate that these aspects of listening are not well understood, and surprising that there has been relatively little research treating them.

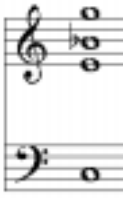


Figure 3-2: A chord.

I define the *musical surface* to be the set of representations and processes that result from immediate, preconscious, perceptual organization of a musical stimulus and enable a behavioral response[8]. There are then three primary questions that immediately concern us. First, what sorts of representations and processes are these? Second, what sorts of behaviors do they afford the human listener? Third, what is the interaction between the representations and the processes as the listening evolves in time?

I wish to be quite clear that I view the behavioral definition of *musical surface* to be essential, and further, that I consider an ethological approach to such behaviorism the most pressing research issue right now. That is, I am predominantly interested in the *actual* behaviors of the listening human as manifested in *natural* settings and situations. To review the literature on cognitive music psychology as discussed in Chapter 2, Section 2.2, it is quite clear that little research has taken this approach directly. Of course, the goal of most music psychologists has always been to arrive, in the long term, at a sophisticated model of the music-listening experience. But their choices of experimental paradigms have almost always required the listener to make unnatural judgments on impoverished stimuli.

As an example, let us consider the sorts of musical behaviors that a typical non-musically-trained listener engages in as part of everyday life. Imagine that the listener is confronted with five seconds from the middle of some piece of music that he had never heard previously. (For example, imagine that the subject has turned on a radio that is tuned to a random station.) Such a listener can tap his foot, or otherwise move rhythmically, in response to a musical stimulus.[9] He can quickly articulate whether the piece of music is in a familiar style, and whether it is a style that he likes. If he is familiar with the music he may be able to identify the composer and/or performers. He can list some instruments that he hears in the music. He can immediately assess stylistic and emotional aspects of the music, including whether or not the music is loud, complicated, sad, fast, soothing, or anxious. He can make complicated sociocultural judgments, such as suggesting a friend that would like the music, or a social occasion for which it is appropriate. All of these judgments interrelate to some degree: whether a piece of music is fast and loud surely affects whether it is judged to be soothing.

Naturalistic real-world settings exist that provide opportunities to see these behaviors in action. Perhaps the most significant today is "scanning the radio dial." The listener rapidly switches from one radio station to another, assessing within scant seconds whether the music he hears is (a) a piece that he likes, (b) a style that he likes, (c) suggestive that the radio station will shortly play some other piece of music that he likes, and/or (d) a piece/style/station that would be liked by other people in the vicinity. Based on these judgments, he immediately decides whether to stay with this station or to move onto the next.

A preliminary report on "scanning the dial" behavior and its implications regarding the use of music to mediate social relationships was recently presented by Perrott and Gjerdigen (1999). They found that college students were able to accurately judge the genre of a piece of music (about 50% correct in a ten-way forced choice paradigm) after listening to only 250-ms samples. This remarkable result forces us to confront the issue of musical surface directly. The kind of musical information that is available after only 250 ms is quite different than the kind of information that is treated in the traditional sort of music-psychology experiment (notes, chords, and melodies). 250 ms is often shorter than a single note in many genres of

[8] Although this is not to say a "conditioned" response in the sense of classical behaviorism.

[9] To be sure, it is an empirical question whether this or the other behaviors discussed here are *actually* exhibited by listeners when they are engaged with music, and whether they are actually immediate and preconscious. I do not mean to posit these behaviors, but to suggest them as reasonable working hypotheses. Experimental results for some tasks, including rhythmic tapping, are presented in Chapters 4 and 7.

music; therefore, listeners must be making this decision with information that is essentially static with regard to the musical score (although certainly not stationary in the acoustic signal).

A crucial motivation for the research I present here is that *this is a kind of musical behavior that is fundamentally inexplicable with present models of music perception*. It is not at all clear what sort of "cognitive structures" might be built that could support this sort of decision-making. The stimuli are too short to contain melodies, harmonic rhythms, or much hierarchical structure. On the other hand, the frequency content, in many styles, is not at all stationary even within this short duration. Thus, it seems quite possible that listeners are using dynamic cues in the short-time spectrum at least in part to make these judgments. This sort of description makes genre classification, at least at these short time-scales, seem very much like timbre classification. This viewpoint is in concert with the writing of many recent composers on the relationship between timbre and orchestration (Erickson, 1985).

Exploration of the concept of musical surface is a primary theme of this dissertation. This concept is essential if we wish to connect music psychology to psychoacoustics as discussed in the introduction (Chapter 1). That is, the musical surface, as well as enabling important musical judgments directly, can also be seen as the bridge that allows music psychology to rest upon a coherent psychoacoustic theory. There are several other working hypothesis that I can articulate regarding the musical surface; these will be discussed in passing as the models and experimental results are presented.

- The musical surface is the crossover point between low-level, bottom-up perceptual processing and high-level, top-down (or circulant) cognitive processing. That is, the processes involved in constructing the musical surface from acoustic input are primarily automatic and bottom-up, while processes that make use of the musical surface to exercise cognitive abilities (memory, structural analysis, expectation) may make use of expectations, goals, and strategic thinking as well.
- Related to this, if a theory in which there is some central symbolic processing agency is presumed, the musical surface is the crossover point between sensory processing and symbolic processing. Processing previous to, and resulting in, the formation of the musical surface is like signal processing, while subsequent processing that makes use of the musical surface to enable cognitive behaviors is like the manipulation of symbols. (I do not presume such a central processing system; see Section 3.4).
- The musical surface contains only entities (properties and objects) that are calculated from the acoustic input. Memories and expectations that do not arise from the acoustic signal actually impinging on the listener should not be considered part of the musical surface.

Of course it is the case that top-down, cognitive information is used in making musical judgments. That is, for each of the cases discussed above, the exact nature of the behavior will depend on many factors, only some of which are directly related to the signal. To take one simple example, the temporal locations at which a listener taps his foot (that is, the exact form that the tapping behavior takes) depend not only on the musical surface, but on the listener's musical ability, his attentive state, his motor skills, his musical tastes, and his mood. All of this is incontrovertible.

But what is also clear, and what I am trying to emphasize most in this dissertation, is that the behavior cannot depend on these factors *alone*. Auditory processing of the musical signal must be involved. It is an empirical question precisely how much of any particular behavior can be explained with low-level models, and how much requires high-level models in addition. The role of the musical surface (as part of an entire cascade of representations, from very high to very low) in enabling music behavior has been largely unaddressed, and it is this problem that I am trying to rectify.

To avoid confusion, there is no connection intended between the notion of surface structure in generative linguistics and the notion of the musical surface that I am promoting here.

3.3. Representations and computer models in perception research

An essential part of my approach is the development of computer models as implementations of perceptual theories. Of course there has been research on musical signal processing algorithms for some time, as reviewed in Chapter 2, Section 2.3. But there has been relatively little discussion of the interpretative relationship between psychoacoustic theories and computer signal-processing programs, particularly for theories of the perception of complex scenes.

In the early days of modern psychoacoustic science (roughly from the 1920s until the 1970s), research considered only very simple sounds. Researchers made progress by proposing theories about the perceptual processing of these sounds, based on experimental evidence, and using their theories to make new hypotheses that could be tested with new experiments. The range of stimuli considered by a theory was very narrow, and as a result a single experiment or small body of experimental work would often suffice to formally test its implications.

As a new focus of research began to consider mixtures of simple sounds, and sounds that were not stationary over time, this situation changed. The number of degrees of freedom of the stimuli admitted to these theories is very large, and so it takes a great deal of difficult experimental work to test the theories in a principled way. It takes the larger part of a career to conduct the experiments, and lengthy monographs to report the results and discuss their implications. The works of Bregman (1990) and Warren (1999) are prime examples of this.

Even the complex stimuli used in perceptual ASA experiments such as those reported by Bregman are still very simple when compared to real-world sounds, however. If we wish to develop theories that are capable of explaining the perceptions of complex stimuli such as actual recordings of real music, or mixtures of multiple speakers in the presence of reverberation and ambient noise, it may be necessary to engage in larger leaps of deduction. It is impossible to perform well-controlled listening experiments that can fully test all degrees of freedom in a perceptual theory for ecological sounds in full complexity.

In this scenario, the role of computer modeling to implement and test theories of complex sound perception becomes more critical. Computer models allow researchers to frame and test hypotheses regarding the ways that sound is processed by the auditory system. Even if it is not possible to explore the actual relationship of a model to current conceptions of the auditory physiology, it is still possible to understand what aspects of mid-level representation and processing are necessary and sufficient to account for a given set of observed perceptual effects. The implementation of a processing theory as a computer program is a more rigorous way to present such a theory than is a simple explanation of experimental results. The construction of computer models challenges the researcher to explore hidden assumptions more thoroughly than does arguing for the validity of paper-and-pencil models.

On the other hand, the development of computer models has its own theoretical difficulties. Most notably, these are a tendency to conflate the computer program that embodies a model with the computational theory that is being implemented, and an inadequate consideration of the role of representational choices in the description of a computational theory. I will discuss these issues in this section.

3.3.1. Representation and Music-AI

Desain *et al.* (1998) have presented a valuable critique of computational approaches to the modeling of music cognition. They make two key points. First, they argue, this paradigm too

often degenerates into a loose "if you want to understand the theory, here, look in my program" approach. They correctly observe that a computational theory of music cognition—indeed, any computational theory of perception or cognition in any modality—must be validated through comparison to meaningful experimental data. I would go one step further on this point, and argue that a computational model must generate novel *testable* predictions about the perception of music if it is to be considered with any seriousness as a theory; that is, a contribution to the scientific discourse.

Desain *et al.* argue that simply comparing the behavior of two systems does not, itself, validate a model—"the model builder might have implemented a large lookup table and listed all the stimulus patterns and the appropriate responses" (p. 156). They say that the way to deal with this issue is to get in under the hood of the system, to understand what aspects are theory-driven, and what aspects are implementation-driven, as seen in Figure 3-3. I would respond that the ability to make predictions separates the sorts of systems that Desain *et al.* appropriately disparage from models that really embody computational theories. If we can understand the input-output function of the model for some interesting class of signals, then this class of admissible inputs and the input-output mapping become a theory that can be tested in a perceptual experiment (by treating the model outputs as theoretical predictions). This can be done, I claim, without recourse to opening the model.

Evaluating computational models

There are three problems with trying to open the model in the way that Desain *et al.* suggest. Because of them, it is not possible to resort to this sort of evaluation as part of the interpretative process of science.

The first problem is that it can't be done in a principled way. Any theorist who opens a model to see how it works and how it implements a computational theory is bound to bring her own interpretative biases along. The second point of Desain *et al.* tries to address this: by applying results from the theory of computer science, formal semantics, program verification (Gries, 1981), and so forth, we might be able to develop the ability to examine computational theories in a formal way. I don't think this is true, because it's too difficult to develop a good formal semantics that ground a computer model with respect to the psychological theory that it represents.

This sort of grounding would require the development of new formal logics in which

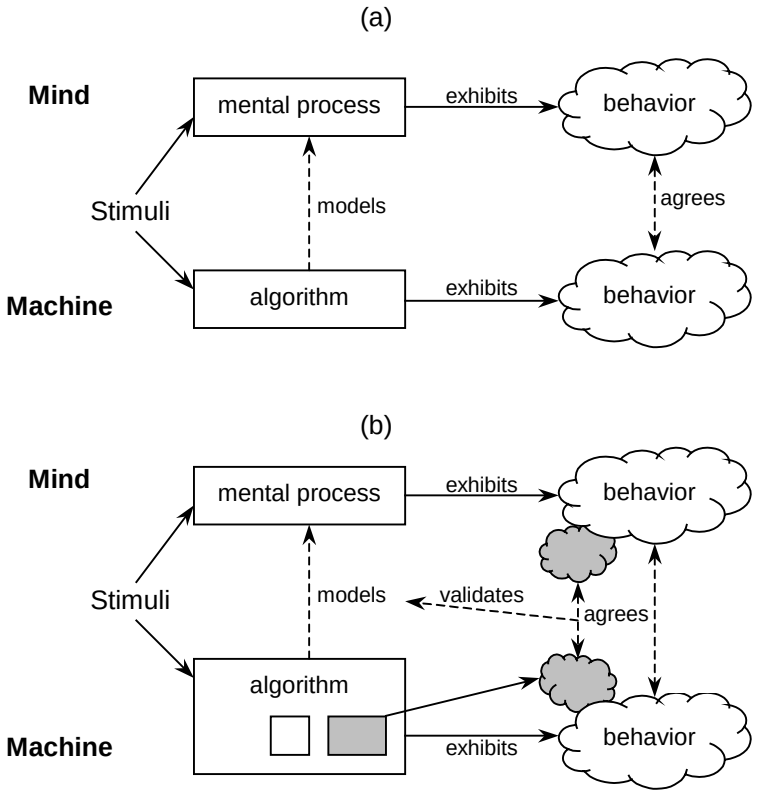


Figure 3-3: Psychological validation of a computational model, as presented by Desain *et al.* (1998). In (a), the behavior of a computer algorithm is tested against that of a human mental process; the same stimulus is presented to each and the behaviors that each exhibit are compared. If the algorithm is a model of the mental behavior, then the results should agree. However, Desain *et al.* argue, this is only a necessary test, not a sufficient one for validating the algorithm as a model of a mental process. A stronger test is shown in (b). The algorithm is "opened up" to reveal that it has several subprocesses. Each of these subprocesses is responsible for part of the observed machine behavior. If each such machine subbehavior can be seen through experiment to have an analogous human subbehavior, then this is a very strong sign that the algorithm is indeed modeling the human mental process.

statements can be interpreted as claims and results in the scientific domain at hand (statements about psychoacoustics). Then, advanced compilers and program verification tools would

relate computer programs in real languages like C++ to structures built in these formal semantics, so that each computer program can be "translated" into a set of statements and predictions about psychoacoustics, music perception, and so on. But unfortunately, neither of these steps is particularly close to reality at this time. The required sort of logical semantics have never (to my knowledge) been developed for any complex inferential system, only very simple ones like text-processing systems and blocks worlds. Thus, restricting rigorous evaluation of computational theories to domains that can be so formalized would force the same sorts of restrictions as those arising from strict experimentalism.

For the foreseeable future, the relation of theorists to their models will have to remain on the intuitive level rather than being formalized. This doesn't mean that we can't evaluate models, but it means that we must be "creative" in doing it: we have to look at models and think hard about what kinds of experiments would confirm or disprove the hypotheses they suggest. The computational model and the psychoacoustic experiment play overlapping and complementary roles in advancing our knowledge about the world.

A second problem with trying to open models is that it can be difficult to lodge criticisms and comparisons at an appropriate level of description. An reviewer of a earlier paper describing the tempo model presented here (Chapter 4) complained that it failed to embed known data on the effects of accent structure on beat perception. This is exactly the danger of opening the hood in the way Desain and Honing recommend: that we open it with a preconception, based on readings of the literature, that we "should find" one element or another. In this case, the reviewer wanted to see an element in the model that corresponded to the data with which he or she was familiar regarding accent structures.

But of course, Povel and Okkerman's (1981) discussion of accent structures makes no claims that this is a *processing element* of their theory. Rather, they present a *structural description* that explains post-hoc the outcome of certain perceptual experiments. Thus, my beat-tracking model should be able to explain and produce similar results when given the same acoustic input stimuli, *even without an explicit processing element to account for them.* We might call the relevant data an "emergent behavior" of a certain set of systems that process sounds with unusual accent structures.

This is not to say that it is an *a priori* given that my beat-tracking system *will* produce the right emergent behavior. Rather, it is a hypothesis to be tested:

> H1: The accent-perception data of Povel and Okkerman (1981) can be explained with a signal transformation model of beat inference from acoustic signals.

It is an empirical question (once the model is described accurately enough) whether this hypothesis is confirmed or denied; in either case, we have an addition to the scientific discussion regarding beat perception and accent structures.

The use of "emergent" in the preceding structure will make some readers think about neural networks, since this is the processing model that most often uses this terminology. But this is just because, due to the black-box nature of neural-network models, it is generally *impossible* to open the hood and see what's inside. Thus, the search for emergent behavior is the *only* way to evaluate such a model. But this does not mean that the emergent-behavior philosophy only applies to self-organizing models; a signal-processing or structural model can be evaluated both from the inside, from an "architectural" point of view, and from the outside, by examining the emergent behavior of the system in relation to complex or unusual stimuli.

The final problem with trying to evaluate models by opening them and examining their subpredictions and internal representations is that, for stimuli of interesting complexity, it is far too difficult to do. For the sorts of musical sounds I consider throughout the main parts of this dissertation, we have little psychoacoustic or behavioral data that actually suggests what internal representations are being used. To be sure, we are able to introspect about the process of listening to complex sounds (such as the sound produced by a symphony

orchestra), but we are a very long way indeed from being able to say anything coherent about the actual representations that are being used.

One counter-argument to this point is that by studying simple stimuli, we can learn about the representations that are used in more complex stimuli. This is the traditional approach of science, to break a difficult problem down into easier subproblems. The dilemma here is that, for the kinds of signal-processing techniques that have been applied to simple musical sounds, we find that they do *not* scale to more complex sounds. Thus, if we are to engage in a reductionist approach that might allow representations to be examined, we still do not know what is the right way to reduce and derive subproblems while still leaving ourselves the ability to induce answers for interesting real-world problems.

This difficulty, like the previous one, points to the necessity of *indirect* evaluation. We are confronted with the need to look at emergent and surface-level behaviors of systems, because this is the only level at which we can hope to find a basis of comparison. (More discussion of this point will be presented much later, in Section 6.2).

The role of representations

The essay by Desain *et al.* further fails to consider a fundamental aspect of any computational system, that of the input/output representations. In their schematic (Figure 3-3), they show an arrow going from "stimuli" to "mental process" and another one from "stimulus" to "algorithm." But of course, it's not the stimulus in an abstract sense that goes to either one of these processors, it's some concrete set of inputs. Typically, in the top panel, the human system, the stimulus is an acoustic sound. Thus, this top part leaves out the crucial *transduction* stage, wherein the perceptual apparatus takes in a signal and forms a representation. And the diagram is misleading, because in the vast majority of models of music perception and cognition, it is a different stimulus that is applied to both processors. The acoustic stimulus undergoes some kind of preprocessing before it is presented to the algorithm; or worse, the preprocessing is *implicit* and some theory-specific representation is used to substitute for the stimulus itself.

Most frequently in music perception research, the theory-specific representation that has been used is the traditional Western musical notation. The viewpoint to which this leads is misleading; I have argued the reasons why in great detail elsewhere (Scheirer, 1996), but I will recapitulate the main points here. Most models of music perception use what I term 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 and Jackendoff, 1983; Narmour, 1990).

It is a trap set by centuries of musicology and music theory that we believe that symbolic models that 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 perceptual theorists are making a very strong assumption about the early stages of perceptual organization when they assume the mid-level representation of music is "like" a score. This assumption is as yet largely unjustified.

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. (p. 68)

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 an analytic 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 in an actual performance as they are notated here. 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 demonstrated 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 that has great ramifications for the analytic results that 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.

To emphasize, I don't wish to argue that this representation of *Walkin'* is wrong in the sense that I prefer some other sequence of symbols as a better transcription. Rather, it is wrong in the sense that presenting a sequence of symbols *at all* involves the use of a number of unexamined assumptions. These assumptions play an important role in determining the nature of the perceptual theories that build on the symbols. It would be a more rigorous approach to make perceptual theories relate to the acoustic signal itself rather than to any particular notation.

Naturally, it can simplify the process of conceiving and describing perception theories to use notations that are less unwieldy than the acoustic signal. But when we do so, it should be *essential* that we consider in a careful and principled way the assumptions embodied in those notations, and their connection to the acoustic signal. The acoustic signal must always be considered the fundamental basis of music perception.

I am not the first to present an argument like this one warning about 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)

Nicholas Cook (1994) wrote an incisive essay arguing against what he terms "scriptism" in music studies (both music theory and music psychology), in which he makes a similar point.

The study of the true perceptual organization of music, as opposed to the study of the historical remnants of music theory, is the fundamental organizing principle of the research I

outline here. I hold firm to the position that most stages of music perception have nothing to do with notes for most listeners. 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.

Naturally, in the course of building models and describing theories, I make assumptions and simplifications myself, but I try to make the essential step of recognizing that they are *theory-specific* assumptions. They may be criticized or falsified, which then damages the theory of which they are a part. By not talking about the representations in (for example) their beat-tracking work, Desain and Honing make it seem as though the representations are "natural" and somehow theory-independent. They are not; in fact, a bit of analysis of the operation of their systems makes it clear how crucial their representational assumptions are. All of the models I develop here work from the acoustic signal only. I hope to make clear how difficult it is to map from acoustic data to mid-level representations for complex sounds, and what an essential part of a perceptual theory this mapping process forms.

3.3.2. On components

As computer models have been applied to the study of the perceptual organization of sound scenes, another representational approach has become common. This is the use (as outlined in Section 2.1.2) of a sinusoidal analysis of the input signal as a first stage. This leads to a mid-level representation composed of a set of sinusoidal components. These components, called "tracks" by Ellis (1994) and "synchrony strands" by Cooke (1993), as well as various other things, are used as an input to the next stage of processing. Thus, the sinusoidal analysis provides a mid-level representation for further analysis of the stimulus.

The impact and influence of this computational theory was probably due to the conjunction of two factors. The first was the development of new digital-signal-processing techniques for forming sinusoidal representations of sound (reviewed in Quatieri and McAulay, 1998), which provided a new suite of tools for signal-processing researchers to work with. The second was the organization of Bregman's work on perceptual auditory scene analysis. Bregman (1990) clearly presents the hearing process as consisting of three stages.

1. Identification of components. Through some unspecified process, the incoming acoustic waveform is transformed into a set of components. Components are typically sinusoids in Bregman's work, although he treats wideband and narrowband noises at some times as well.

2. Simultaneous "grouping" of components into events. Based on "Gestalt grouping rules" such as proximity, harmonicity, and common fate, at each instant in time the components are partitioned into a set of sources.

3. Sequential "streaming" of events to form auditory streams. Bregman presents a number of rules that depend on the properties of the events and describe the manner in which humans form auditory streams from the events.

In Bregman's characterization, the first stage seems generally to precede the second and third stages. The second and third stages may operate in parallel or in sequence.

The early computational auditory scene analysis (CASA) systems (Brown and Cooke, 1994a; Ellis, 1994) can be considered attempts to directly implement the grouping and streaming rules Bregman suggested. As a cognitive psychologist, Bregman articulated his theory of hearing with computational metaphors that seem amenable to direct implementation. But when computational studies began, it immediately became clear that handling sinusoidal components—extracting them, processing them, and maintaining good representations of them—is the most difficult part of such a system to build. Recent approaches similar in spirit have developed increasing elaborate ways to extract sinusoids from complex sound scenes (Mani, 1999), and/or a more complex ontology of sound components (Ellis, 1996a).

It is worthwhile to consider the psychological status of components like sinusoids. Just because they are a convenient representation for explaining human behavior (which was Bregman's overt goal), doesn't mean that they are useful computational entities, or that the hearing system actually extracts them during the listening process. As with the discussion of accent structures above, components are a convenient formalization in which to describe (a) the way a test sound scene is constructed through synthesis, and (b) the results of psychoacoustic experiments on these test sounds. There are no experimental results of which I am aware that directly demonstrate the psychological reality of sinusoidal components in the perception of complex sound scenes.

I consider the focus on components in CASA systems to be misdirected. It has proven very difficult to get good "componentization" of complex scenes, and the higher-level stages of previous systems are sensitive to imperfections in the component extraction. Notably, modern computational psychoacoustics does not often use this model. Particularly in the contemporary pitch literature, the recent emphasis has been on the extraction of perceptual properties directly from auditory representations of sound (Meddis and Hewitt, 1991). In computational studies of double-vowel separation (de Cheveigné, 1997), the analysis is similarly conducted directly from the autocorrelogram.

We are left with something of a gap in the literature. On one hand, we have CASA systems, which attempt to process complex, time-varying acoustic scenes, but which use an awkward "sound component" model of signal processing. On the other, we have computational models of double-vowel perception (as discussed in Section 2.1.1), which seem to be more plausible models of sound processing in the auditory system, but haven't been extended to stimuli other than the most trivial. The model that I will present in Chapter 5 is an attempt to bridge this gap—to present a starting point for a theory of the perceptual segmentation of complex sound scenes that doesn't include a componentization step.

3.4. Understanding without Separation

This is not a dissertation about polyphonic music transcription. The goal of my research is not to recover the score of a piece of music from a performance, or to separate sounds in the sense of creating multiple output sounds that can be summed to reconstruct a scene. Rather, it is to show how incoming sound data may be transformed by simple signal-processing and pattern-classification techniques directly into judgments about the musical qualities of a stimulus. In this section, I will discuss the theoretical implications of this stance.

Many previous authors have described computational and geometric models of the perception and cognition of music. With very few exceptions, however, these theories have been not been grounded in actual sound processing. Rather, as discussed in Chapter 2, a simplifying assumption is made, in which the sound to be transformed by some unspecified perceptual agency into a structural representation suitable for supporting the theory at hand. Cognitive psychology, with its emphasis on mental structures and representations, is much better at forming and testing hypotheses built around high-level representations than it is about examining critically the notion of high-level representation itself.

What I will demonstrate in Chapters 4-7 is that to make many kinds of interesting musical judgments, high-level structural representation of the musical scene is unnecessary. In fact, given that the systems I will demonstrate are presently the only ones that can make the judgments they can, a stronger argument is that the use of structural representations is a *fundamentally misleading* approach to the construction of perceptual-cognitive models of musical listening.

Most people do not, I claim, maintain score-like representations of music that are then used to perform cognition. *There are no mental entities that correspond to events happening in the*

music being heard. Rather, the process of music listening is one of continuous transformation, in which different agencies of musical expertise (in the language of Minsky (1985)) monitor the input signal and produce various output signals in the forms of behavioral responses. These responses may be overt, as in foot-tapping behavior, or emotional, as in tension and relaxation reactions, or entirely immanent, as in recognition and classification of instruments, genres, and composers.

This argument is so unusual as to seem absurd on the surface. Of course it must be the case—the reader may say—that I have a score, or at least some sort of central representation of a piece of music in my head. After all, I can imagine music in my head, and reproduce it upon request, and recognize themes when they recur, and describe the effect of applying Coltrane changes to a 32-bar standard jazz form. Perhaps a "simple transformation" can do something simple like tap a (virtual) foot to the music—goes the argument—but you'll never be able to demonstrate a similar system that does X (where X is the reader's favorite "high-level" musical behavior).

My reply to this is that the fact that we can do such things (which is not in question) does not constitute evidence that there are structural representations in the mind. I believe that, given time, we will be able to demonstrate systems that *do* perform all of these tasks without overt centralized representation. Structuralism is an incorrect conclusion drawn mostly from introspection (which, as is well established, is often wrong about what's really going on in the mind) and uncritical acceptance of music theory as a starting hypothesis for music psychology (Cook, 1994).

Structuralism is a fundamental tenet of cognitive science. Cognitive science has as its very goal the explication of the mental structures and processes that are used to analyze and transform internal representations. As a basic assumption (not as a working hypothesis), the notion that the external world is somehow converted into symbolic models, and these models are the basis for rational behavior, is taken as granted. Krumhansl writes:

> Listening to music, we hear the sounds not as isolated, disconnected units, but integrated into patterns. Our perceptual experience goes beyond sensory registration of single musical events. Sound elements are heard in context, organized in pitch and time, and are understood in terms of their functions within that context....The listener appreciates the organization of a musical work by assigning significance to the sounded elements according to their roles in the musical context, and by integrating them into the broader pattern. (Krumhansl, 1990, p. 3)

Note the variety of assumptions made regarding the hearing process and the way it relates to music perception. There are "sounds" (as reified, immanent entities) that can be treated as "units," "events," and "elements." Each sound "element" is perceived in terms of its pitch and time and is "understood" in terms of "functions." The only behavior of the listener considered directly is that of "appreciation" of the "organization" of a musical work. Crucially, these statements are not working hypotheses to be tested; rather, they are assumptions treated as the fundamental starting point of research. But this description of music-listening seems very far indeed from the real experience that most (particularly non-musically-trained) listeners have with real music. In fact, from this description, it seems unlikely that non-musicians should enjoy or appreciate music at all!

The immediate response by a cognitive scientist to this argument is that while untrained listeners may not be consciously aware of their symbol-manipulation abilities, they really must be using these abilities deep down in order to engage in the behaviors we can observe on the surface. But there is no proof of this assertion: there are no structuralist computational models of music perception that can do *anything* robustly with a wide variety of audio input signals. In all of the functioning structural models that I'm aware of, either the input domain is greatly simplified and restricted (as in polyphonic transcription research), or the system takes symbolic input like MIDI as the starting point. In each case, a hand-waving argument is presented that the system is really part of a speculated bigger system—in the first case, that

the lessons learned will someday be used to construct truly general polyphonic pitch-trackers, and in the second, that an audio segmentation preprocessor will someday be able to produce the right representations for symbolic processing. The argument that a transformational model cannot demonstrate behavior *X* has no weight, until a structural model *can* demonstrate behavior *X* given a wide variety of audio signals as input.

This criticism is echoed by the writing of several theorists of ecological psychology, who use a similar argument to criticize the modern tradition of psychology more generally. For example, E. S. Reed (1996) writes:

> It is striking that to the extent our psychology has been conceived of as "scientific," it has tended to shrink from dealing with everyday concerns. Scientific psychologists have ceded the territory of the every day to popular psychologists and even to outright quacks, whom the scientists profess to disdain. But instead of mounting a challenge to these inadequate accounts of common life, the vast majority of scientific psychologists take refuge in a self-justifying myth that science equals experimental control equals avoidance of the messiness of the real world (p. 7).

It is precisely the "messiness of the real world" that I am primarily concerned with in this dissertation. But it is important to recognize that this does not imply that I believe music is somehow mystical, or impossible to explain scientifically. I do believe that a rigorous scientific basis is essential for founding theories of the perception, appreciation, and creation of music—but structuralism is not the right basis.

Much of the structuralist stance in music-perception research can be seen as stemming from the continued dominance of the music sciences by ideas and attitudes from music analysis and theory. Music analysis has a long tradition of privileging a view of music listening as rational, symbolic, and refined that is quite out-of-step with the actual listening experience of most listeners (Cook, 1998). Western culture still grants that music analysts with their excruciatingly detailed analyses of score-based minutia are the final authority in musical interpretation. Music psychologists posit a listening process that reflects the music analyst in microcosm: the process by which a listener interprets a heard piece of music mirrors the process by which an analyst interprets the score of the music. Perhaps the listener is not as *sophisticated* as the analyst, but that is the position to which we all aspire, in this model, as music-lovers.

For example, Sloboda (1985) writes the following in discussing the results of an experiment by that failed to find support for a structural theory of emotion in music:

> It is natural to suppose that, as one becomes increasingly sophisticated musically, one becomes attuned to finer emotional nuance. A hierarchy of emotional cues seems likely, with primitive cues (such as speed and loudness) available at all levels of musical sophistication, and with more subtle cues (such as aspects of tonal relations) available only to those with deeper analytic powers with respect to the music. A common experience among many music lovers (including myself) is a belated appreciated of the emotional diversity and subtlety in the music of a composer such as Mozart. The inexperienced listener may find Mozart pale, insipid, and all vaguely 'jolly', especially when set beside the kaleidoscopic turbulence of the romantic composers. Closer knowledge of Mozart (and maybe of the emotional world) results in his music becoming richly and sublimely expressive. The ability to 'read' the emotional language of music is an acquired skill, and we should, perhaps, not be too surprised that a group of 'ordinary' subjects do not show much awareness of the finer details of this language. (p. 63)

It is hard to interpret such an argument as saying anything other than the reason the experiment failed to find evidence for the proper emotional relationships is that it used subjects that were not sophisticated enough to hear the proper emotional relationships. That is, that there is a single measure of ability to "read *the* emotional language," apparently given by the theory under test, and that we may pick and choose subjects in testing this theory according to their ability to measure up accordingly. It is difficult to see how the original

theory could ever be falsified under such testing conditions. Further, this argument denies the ultimate subjectivity of the emotional response to music, instead suggesting, as Cook (1998) writes, that:

> If you aren't a composer or performer, or at any rate someone with a musical training, then you are a non-musician.....Classical aesthetics doesn't recognize you as a stakeholder. (p.83)

A similar philosophical approach stressing the importance of sound separating and "parsing" can be found in the computational auditory-scene analysis (CASA) literature. In traditional CASA systems, the goal of sound-processing has been to extract multiple constituent sounds from a mixture. The output sounds can then be analyzed independently to compute their features. The sounds that are the output should be the same in some perceptually important way as the sounds that were the constituents of the mixture. A primary motivating factor for this approach is its potential application to automated speech recognition (ASR). ASR systems today perform passably well on clean speech without interference; this makes it attractive to imagine "cleaning up" signals with noise or interference in order to use them as input to unmodified ASR systems.

In Chapter 5, I do present a method for analyzing a complex sound scene into multiple auditory images. The goal of the method I develop is not to separate the sound, but to *partition* the sound data so that feature analysis and further continuous transformation can be undertaken. The difference between the goal represented here and the goal represented by most previous research into computational auditory-scene analysis (CASA) systems is represented schematically in Figure 3-4.

It is apparent that sound understanding is easier than sound separation, since less time must be spent on achieving high-quality synthesis of output sounds, and that it is more similar to the human hearing process, since it is unlikely that human listeners maintain multiple independent time-domain signals as an intermediate representation of complex signals.

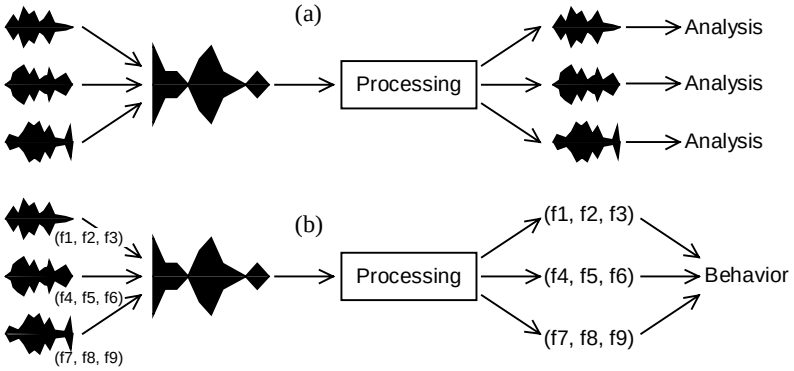


Figure 3-4**:** Two different models for computational auditory scene analysis. In (a), a *sound separation system* analyzes a sound mixture to discover the constituent sounds that comprise it and analyze them. The goal is to extract and reconstruct the exact constituent sounds that went into making the sound mixture. In (b), a *sound understanding system* analyzes a sound mixture to discover the *features* of the sounds that comprise it, and use these features to support interesting behaviors. The goal is to describe the sound scene at a sufficient level of detail for solving problems.

The advantage of the understanding-without-separation approach is most apparent in the case when one constituent signal destroys information in another through masking or cancellation. In a sound-separation system, it is very difficult to deal with this situation properly, since the obliterated sound must be invented wholesale from models or *a priori* assumptions. In a separationless approach, the required action is one of making feature judgments from partial evidence, a problem that is treated frequently in the pattern recognition and artificial

intelligence literature. Rather than having to invent a answer, the system can delay decision-making, work probabilistically, or otherwise avoid the problematic situation until a solution presents itself.

It is worth noting the similarities between this theoretical stance and that of Brooks (1999). Brooks is a AI researcher and roboticist who worked on vision systems early in his career. More recently, he has been arguing in favor of a *reactive* or *representationless* approach to building robots, as an alternative to the predominant *representational* form of such systems. In the representational model of systems-building, a computer vision system tries to segment and understand the world visually, and then the robot plans a course of action using the high-level representation that the vision system produced. In a reactive system, the robot's sensors are wired directly (through a series of transformational stages) to its actuators. There is no central plan, map, or agency in such a system—the robot operates in continuous immediate reaction to the environment.

Even using such simple mechanisms, Brooks' "creatures" can implement extremely complex behavior—his robot "Herbert" can wander through a building, enter open offices, find empty soda cans, and pick them up and return them to a recycling facility. Most interestingly, observers of such systems typically *impart* a systemic planning agency to the robot—that it, it *looks* like the robot is planning, deciding, and forming representations of its environment. Since it is not, this is valuable evidence that simple inspection of a system cannot reveal its representations aspects clearly. (Philosophically, the idea that representations and intentions are something imparted from without by an observer, rather than imbued within an organism or other active system, is termed the *intentional stance*; Dennett (1987) among others has written extensively about this.)

The parallels between Brooks' arguments and mine are striking. Brooks argues that the focus on representationalism has led AI research into a path of making oversimplifying assumptions and hand-waving through an argument that someday, the vision system will be able to produce such a representation. This argument is the same as mine—that a focus on score-based processing has led psychologists and music-signal-processing researchers to make oversimplifying assumptions, with associated hand-waving that someday, high-quality polyphonic transcription will be able to produce this representation. Brooks has provided evidence that the "reactive" philosophy of systems-building more directly leads to robust systems that embody behaviors than does a centralized vision-based approach. The systems I will demonstrate in Chapters 4 through 7 parallel this—they operate more robustly and produce more interesting musical behaviors than other systems in the literature.

Brooks' writing on the philosophy of his systems is directly applicable to my approach, with only a few word changes. For example, he lists four characteristics of his constructions (Brooks, 1999, pp. 138-139):

> **Situatedness:** The robots are situated in the world—they do not deal with abstract descriptions, but with the here and now of the world directly influencing the behavior of the system.
>
> **Embodiment:** The robots have bodies and experience the world directly—their actions are part of a dynamic with the world and have immediate feedback on their own sensations.
>
> **Intelligence:** They are observed to be intelligent—but the source intelligence is not limited to just the computational engine. It also comes from the situation in the world, the signal transformation within the sensors, and the physical coupling of the robot with the world.
>
> **Emergence:** The intelligence of the system emerges from the system's interaction with the world and from sometimes indirect interactions between its components—it is sometimes hard to point to one event or place within the system and say that is why some external action was manifested.

My music-listening systems are *situated* in that they do not deal with abstract descriptions of musical sounds, but with the sounds themselves. Just as Brooks says that "[t]he world [is] its own best model" (*ibid*, p. 81), I believe that the sound signal is its own best description. My music-listening systems might be considered to be *musically intelligent* in simple ways, but the intelligence does not come from a set of rules that describe "how music works." Rather, the intelligence comes from continuous transformations of the signal, and the way that these transformations are manifested as musical behaviors. The intelligence *emerges* from the interaction of many (usually highly parallel) processing elements and the rich properties of the input signal. There is no function in which it is decided what is the "right thing to do" as the signal is being processed.

The major aspect of difference for the systems that I present here is Brooks' characterization of his systems as essentially *embodied*. Brooks' form of embodiment crucially means that when the robots make changes to the world, they perceive the changes *via* the world. This is a difficult problem for music-listening systems, since so many aspects of musical listening are immanent rather than manifested in external behaviors. In some ways, this makes the results of the beat-tracking system more satisfying that the high-level description system, since it is much clearer what the observed manifestation of signal understanding should be—an act of tapping along with the signal. Some other system could listen to the taps and make judgments on that basis. Beat-tracking is unusual in this regard; it is unclear what sorts of behaviors should be evoked by the perceptual act of recognizing the musical genre and how they could affect other perceptions.

Perhaps it is better to see the kinds of systems I produce as the perceptual components of larger systems. In this manner of thinking, it is not possible at all to evaluate the systems in their own right, but only in the functionality that they provide to a larger intelligent system (perhaps an automated composer, or interactive synthetic performer). Brooks' theory places great weight on the capability of the perceptual subsystems in his creatures; thus, it is attractive to imagine constructing robust perceptual models using the same philosophy and integrating them into his creatures.

This argument takes us a bit further away from the view of computer system as perceptual model. We are beginning to lose the notion that music-listening systems should be evaluated by comparing their behavior to that of human listeners. Rather, the suggestion is that evaluation is according to engineering criteria. For Brooks, this is natural, as he works in the field of artificial intelligence, with an additional interest in the engineering of computational vision systems.[10] He has no specific interest in the perceptual abilities of humans except insofar as learning about humans will help him build better robots. But I believe that the basic tenets of his theory form a strong candidate for a kind of perceptual psychology that draws together aspects of the mind that are normally considered "sensory" with those that are

[10] As an aside, it is surprising how differently weighted are "machine vision" and "machine listening" within the AI community. Machine vision has always been a central problem of AI—at times, it has been *the* central problem of AI. Every AI lab and program maintains a concentration in machine vision. And yet there is essentially no organized parallel in the hearing sciences. Only a few enclaves of researchers could be said to take an analogous approach in building sound systems, and these researchers are not part of the mainstream AI community. The construction of vision systems is viewed as a problem of fundamental importance in the attempt to build intelligent machines, yet the construction of hearing systems is viewed as a marginal aspect of the improvement of speech-recognizers. This disparity is striking and unfortunate—as students of sensory systems that are complementary in functioning organisms, hearing and vision researchers have much to learn from each other. It is also likely that the problems with which sound research must always deal (for example, the fact that sounds are embedded in time) would provide interesting fodder and inspiration for AI research.

normally considered "cognitive." As mentioned earlier, the tradition of ecological psychology has also developed theories of this sort, although typically not computationally.

It is not surprising that many of the problems and difficulties suffered by cognitive psychology (including music cognition) parallel those suffered by "good old fashioned" artificial intelligence, as the historical connections between these disciplines are well-understood (Gardner, 1985). Reactive models are a way to avoid the dangers of excess symbolization, both in the construction of operational systems and in the conception of psychological theories.

3.4.1. Bottom-up vs. Top-Down Processing

An important theoretical contribution that Ellis made in his dissertation (1996a) was a careful consideration of the relationship between bottom-up and top-down processing. His discussion continues a lengthy chain of debate on this topic in the sensory-systems literature. The debate originates most explicitly with Pamela Churchland's critique (1994) of Marr (1982) and other "pure vision" theories of seeing. Marr presented, in an influential and wide-ranging book, the first coherent computational theory of vision. Analogous to the CASA research on which this dissertation draws, his was a project that drew from (and had implications for) psychophysical theories and theories of perception as well as the construction of perceptual-computing systems.

Marr's primary theoretical contribution was to highlight the kinds of theories that can be built with computer models. In doing so, he highlighted the notion of the *mid-level representation*, and thus made explicit an idea that had been present implicitly in the perceptual-computing literature for some time. Namely, that perceptual processing can be seen as a chain of computational stages in which the voluminous and close-to-the-signal sensory data are resolved with more and more abstraction in succeeding stages. As the processing goes from one stage to the next, the representations become less verbose, more semantic (in the sense of providing direct affordance of behavior), and less directly related to the specific sensory input. The specific mid-level representations that Marr proposed were edge maps and illumination gradient functions (for visual processing), but his theoretical model is more general than this. For example, the naïve model of speech understanding forms such a processing chain: a sound signal enters the auditory system, and is converted to a spectrogram, which is converted into a sequence of phonemes, which is converted into a sequence of words, which are understood.

Marr's viewpoint is attractive for a number of reasons. First, it makes the problem separable, or reducible into independent sub-problems. This matches both the traditional reductionistic approach to science, and the standard approach to computer programming ("divide and conquer," as it is sometimes called). Under the assumption that a problem can be separated into independent stages, the stages can be researched and modeled independently. As long as the representations match up in the middle, the whole system should work when we put it together. This is the reason for the primary important of the mid-level representation in Marr's theory.

Second, all of the processing (perception) is done *locally*. That is to say, to process a local region of the visual image, we need only be concerned with the sensory information coming from that region. We can process all of the sections in parallel, and then put the representations together at the end. For speech, this corresponds to the notion that to model the perception of speech at one point in time, we are only concerned with that particular temporal location in the signal. Earlier and later segments of the signal do not affect the perception of that local part. This is another sort of separability, and has an equally simplifying effect on the construction of computational models.

Churchland's critique of Marr was based on psychophysical evidence that had been known for some time, including the perception of bistable quartets and motion correspondence, and the effect of semantic information on shape perception. These phenomena suggest that vision

cannot simply proceed in a chain of isolated connections; in some way "global" information, or at least non-local information, comes to influence the formation of even low-level representations. Churchland called such influence a "top-down" flow of information, complementing Marr's strictly bottom-up approach. Analogues in the auditory domain were drawn by Slaney (1998) in a critique of Bregman, and Ellis (1996a) in a critique of early work in CASA.

There is also psychoacoustic evidence of the importance of non-local processing in hearing. For example, Warren in a classic and important set of experiments, demonstrated what is now known as *phonemic restoration*. If a short segment of a speech signal is silenced (gated out), so that the sound has a hole in it, listeners are readily aware of the gap and distracted by it. However, if the gap is replaced with a noise (such as the sound of a cough), listeners no longer notice the gap, and in fact are typically unable to accurately place the location of the cough in the sentence. That is, listeners have no ability to tell which segment of sound was removed and replaced by the noise. Warren argued that the missing sound was perceptually *restored* to the signal based on higher-level (for example, word-level) cues in the speech. This restoration process must occur strictly before the final analysis or inspection of the speech features, because the lack of speech features for the missing segment cannot be perceived.

More striking still is the extension known as *semantic* restoration. In this experimental paradigm, the segment removed from the speech signal is the one that creates maximal semantic uncertainty in the local meaning. For example, the listener hears one of the following sentences:

- *The *eel is on the axle.*
- *The *eel is on the orange.*
- *The *eel is on the shoe.*

(The * indicates the part of the signal that is obliterated.) Each of these sentences starts out the same. In fact, to assure that there were no intonational cues, Warren (1970) created these stimuli by copying the same sound for the beginning of each sentence. Remarkably, subjects when presented with these sentences restore the sound that gives the sentence its most semantically logical form in each case. That is, the subject hears *wheel* in the first case, *peel* in the second, and *heel* in the third. It is to be emphasized that the subject does not feel as though he is deducing the answer in retrospect. Rather, the phenomenal percept is the same as before: the subject does not even realize that part of the signal has been removed, and cannot say which part of the sentence is synchronous with the cough.

The implication is that the phonemic-restoration process can make use of extremely high-level information—the meanings of words and their networks of semantic associations—in order to fill in the gaps. This is an extremely lengthy (in the sense of a Marrian processing chain) flow of information from a high to a low level of representation.

The construction of computer models that can make use of top-down information in this way is very difficult. A model of processing that only contains a single flow of information, with layers of representation proceeding from more detailed to more abstract, is well in-line with traditional approaches to computer programming and signal processing. It is much harder to build signal-processing systems that can use multiple layers of representation in concert to form judgments. The only serious attempt is a recent thread of work using blackboard systems, with various approaches taken by Klassner (1996), Ellis (Ellis, 1996a), and Mani (1999). The blackboard system is a knowledge-representation model developed in the artificial-intelligence literature (Engelmore and Morgan, 1988) that represents multiple hypotheses, as well as evidence for and against them, in a large data structure called the *blackboard*. As computation proceeds, representations at any stage may be reprocessed to affect representations at other stages, both "above" or "below" them in the representation

chain. There are a number of strategies known successful for ordering the processing in such a scheme.

While not strictly a blackboard system, another successful system to integrate multiple levels of representation was Martin's (1999) timbre-classification system. Martin demonstrated that this system could classify musical instruments by their sounds with accuracy competitive with skilled humans. Martin also drew interesting connections between human abilities in taxonomic classification and the processing-stage models represented by layered-processing theories. That is, if the entire pyramid of representations is available to the conscious mind, and representations from any level of processing can support judgments, then the result is a framework that very similar to the one that perceptual psychologists have suggested as the basis of taxonomic categorization and classification.

In fact, Martin demonstrated that his system made the same sorts of errors in categorizing musical-instrument sounds as did humans: it was more successful at judging the *family* of an instrument (that is, whether the instrument was one of the brasses or one of the strings) than at judging the particular identity of the instrument (that is, whether the instrument was a violin or a viola). As did humans, Martin's system made more successful judgments at higher levels of abstraction than at lower, more specific levels. This is striking, albeit indirect, evidence in favor of something like this layered-abstraction model as a part of the human perceptual-cognitive apparatus.

A different sort of interaction between bottom-up and top-down processing is represented by the present state-of-the-art in automated speech-recognition systems. In these systems, low-level acoustic processing of the signal interacts with high-level probabilistic modeling of likely utterances determined from a language model. These systems are different in flavor than the other systems discussed in this section because they are not psychoacoustically- or perceptually-motivated. There is no particular connection between most of the stages in an ASR system and the present theories of human speech perception (which suggests a reason that human performance is so much better than ASR performance in difficult conditions). Nonetheless, from an engineering viewpoint, ASR systems are well worth examining as an example of methods allowing low-level and high-level information to interact in a principled way to solve interesting problems.

It is common, when trying to develop sound-processing systems, that we focus an inordinate amount of attention on the bottom-up aspects. I think this is a mistake, but a natural one—in order to feel like the system is "doing something right," we often pay most attention to the lowest level of performance (for example, componentization or pitch-tracking) and try to make that stage "perfect." But in light of the kinds of insults to signals that are readily accepted by the human hearing system, there must be a great amount of leeway in the accuracy of the representations at low levels. The slack in such a system must be borne by pattern-recognition at higher levels.

This is difficult system-building terrain, falling in the area of "AI systems" rather than "digital audio processing," and so we often shy away from it in order to focus on things with which we feel more comfortable (signal processing). I think that the emphasis on components can be partly understood in these terms. The proper way to build music-analysis systems that can process complex audio scenes robustly is to develop new kinds of pattern recognition that can deal with messy, incomplete, noisy, and obscured sound data.

3.5. Chapter summary

In this chapter, I have presented a wide-ranging collection of material, starting from basic definitions and proceeding to an articulation of my computational and philosophical approach. I assert that this approach is the right one for the development of theories of musical hearing

and for the construction of machine-listening systems. I have also discussed the role of computational representations in computer models and psychoacoustic and music-psychological theories.

It is worth contrasting the approach I have presented in this chapter with the summary of previous research that I presented in Chapter 2. To do so will give a clear picture of where the rest of my dissertation is headed.

1. I propose that rather than focus on the study of musical transcription, we should build systems and theories based on the principle of *understanding without separation*. In this model, perceptions of sound, and digital-signal-processing programs, are developed as continuous transformations from input (sound) to output (judgment). I assert that this approach is both easier, since we then don't have to deal with the practical impossibilities of separating sounds, and closer to human perception, since the human listener surely doesn't maintain independent time-domain waveforms as an intermediate representation of sound.

2. I propose that, to be maximally relevant to the full spectrum of listening situations and practical applications, we must embrace a *broader range of musical sound*. Our experimental materials and processing inputs must reflect the actual musics that real people listen to, not music-theoretical biases about what kinds of music are sophisticated or well-structured.

3. I propose that the study of music-listening in humans and machines must focus on *sound processing*; and further, that present theories of sound processing in humans and machines must be updated to deal with the messiness of real sounds in the real world. We must maintain strict attention to the relationship between perceptual theories and the primitive sound-processing functions and representations they assume. We must also try to build computational systems that are capable of dealing automatically with a wide range of complex sound signals, even if the initial capabilities of such systems seem limited compared to those that only apply to a more limited domain or that assume fancy preprocessing.

4. I propose that the study of music-listening in both humans and machines must focus on the *ecological* perception of sounds; that is, the sorts of judgments made by everyday listeners about everyday sounds as part of the natural music-listening experience. It is still an open question exactly what the nature of such judgments is, but it is a pressing question for both theoretical and practical reasons. We must try to elicit these judgments in perceptual experiments, and model them in computational approaches.

Now that I have articulated the approach and philosophical stance that underlies my research, the stage is set for me to introduce new results. The remainder of the dissertation is taken up with the presentation of several signal-processing models of sound, and their use in creating a feature-based theory of immediate musical perception.

CHAPTER 4 MUSICAL TEMPO

The beat of a piece of music is one of the most important immediately-perceived aspects of the sound. Every musical culture organizes sound in time through the creation of rhythmic emphasis. Further, nearly every listener, whether skilled or not according to traditional criteria, can find the beat in a piece of music and clap her hands or tap her feet to it. In this chapter, I will present a new signal-processing model of the perception of beat and tempo by human listeners.[11] The model is constructed as a direct transformation from complex input sounds into percepts of beat and tempo. I will demonstrate that the model performs similarly to human listeners in a variety of musical circumstances, and that it is has certain similarities to existing theories of sound perception that make it attractive as a psychoacoustic model of tempo perception. The model serves as a simple demonstration of the principle of *understanding without separation* as discussed in Chapter 3, Section 3.4.

Automatic extraction of rhythmic pulse from musical excerpts has been a topic of active research in recent years. Also called *beat-tracking* and *foot-tapping*, the goal is to construct a computational algorithm capable of producing behaviors that correspond to the phenomenal experience of *beat* or *pulse* in a human listener.

Rhythm as a musical concept is intuitive to understand, but somewhat difficult to define. Handel writes "The experience of rhythm involves movement, regularity, grouping, and yet accentuation and differentiation" (Handel, 1989, p. 384) and also stresses the importance of the phenomenalist point of view: there is no ground truth for rhythm to be found in simple measurements of an acoustic signal. The only ground truth is what human listeners agree to be the rhythmic aspects of the musical content of that signal. This places tempo and rhythm into the domain of *perceptual attributes* of sound as discussed in Chapter 3, Section 3.1. A pressing question for psychoacoustians and music-signal-processing researchers is to discover the physical properties that correlate with this perceptual attribute.

As contrasted with rhythm in general, *beat* and *pulse* correspond only to "the sense of equally spaced temporal units" (Handel). Where *meter* and *rhythm* are terms associates with qualities of grouping, hierarchy, and a strong/weak dichotomy, *pulses* in a piece of music are only periodic at a simple level. For my purposes, I will define the *beat* of a piece of music as the sequence of equally-spaced phenomenal impulses that defines a tempo for the music. To use

[11] Much of the material in this chapter was previously published as a stand-alone paper (Scheirer, 1998a).

an operational definition, if a listener is allowed to adjust the phase and frequency of a metronome so that its clicks seem well-matched to a musical stimulus, the frequency of the metronome is the perceived *tempo* of the signal. The locations in time of the metronome's clicks indicate the temporal location of *beats* in the music. This chapter is only concerned with beat and tempo. The grouping and strong/weak relationships that define rhythm and meter are not considered.

It is important to note that there is no simple relationship between polyphonic complexity—the number of notes played at a single time, and the number of instruments on which they are played—in a piece of music, and rhythmic complexity of that music. There are pieces and styles of music that are texturally and timbrally complex, but have straightforward, perceptually simple rhythms. There also exist types of music with less complex textures but that are more difficult to rhythmically understand and describe.

The former sorts of musical pieces, as contrasted with the latter sorts, have a "strong beat," and my primary concern is with them in this chapter. For these kinds of musical examples, the rhythmic response of listeners is simple, immediate, and unambiguous, and every listener will agree on the rhythmic content. Because of this, the beat of a piece of music is a constituent of the perceptual musical surface, as defined in Chapter 3, Section 3.2. Rhythmically complex music, in which different listeners might disagree on the beat or tempo, or in which some listeners might have a difficult time deciding where the beat is, is discussed only briefly at the end of the chapter.

The chapter is organized as follows. In Section 4.1, I will describe a psychoacoustic demonstration that indicates that only the temporal envelopes of the subbands of a musical signal convey the beat of the signal. From this demonstration, it appears that pitch and pitch-like information is not a necessary constituent of a theory of tempo perception. Following that, I will describe the signal-processing techniques used to implement the model. The basic principle is the detection of periodic energy fluctuations within a subband representation. In Section 4.4, I will provide evidence that the algorithm performs similarly to human listeners on a variety of musical examples, first qualitatively, through simple demonstration on real musical signals, and then quantitatively, with a formal listening test.

Finally, in Section 4.5, I will discuss the implications of the model, focusing on its significance in the understanding-without-separation approach, a comparison with other models in the literature, and speculations on the connections to psychoacoustics. This section includes a brief reimplementation of the model as a variant of the subband-periodicity model of pitch.

4.1. A Psychoacoustic Demonstration

One of the key difficulties with most previous models of rhythmic perception, as described in Chapter 2, Section 2.3.4, is that they operate from an onset sequence, assuming that the input signal has been parsed into notes or other time-positioned events. As I discussed at length in Chapter 3, this stance is theoretically suspect, since there is no good evidence that such accurate segmentation is performed by the human listener, or even that it is possible at all. From an engineering standpoint, the difficulty with this assumption is the complexity of grouping harmonic partials together to form notes, and determining the onset times of those notes. Even if simplifying assumptions about the pitch and timbral content are made, identifying attack and release times is no easy task (Scheirer, 1998b).

A psychoacoustic demonstration on beat perception shows that certain kinds of signal manipulations and simplifications can be performed without affecting the perceived tempo and beat of a musical signal. Consider the signal flow network shown in Figure 4-1. An *amplitude-modulated noise* can be constructed by vocoding a white noise signal with the

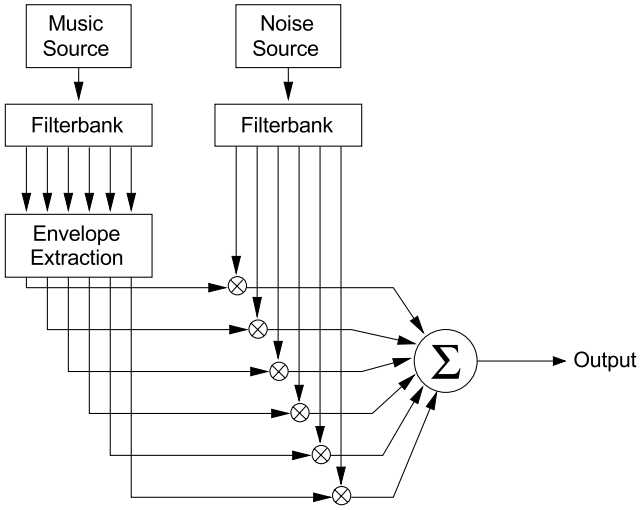


Figure 4-1: Creating a "modulated noise" signal by vocoding a noise signal with a music signal. An input musical sound is divided into frequency bands and the envelopes of the subband signals calculated. In parallel, a broadband noise signal is divided into subbands using the same filterbank. Each filtered-noise band is modulated by the envelope of the corresponding band from the musical signal, and these modulated noises are added together. The output sound, for most sorts of music and many sorts of frequency filterbanks, is perceived to have the same rhythm as the input music signal, indicating that the amplitude envelopes of the bands are a sufficient representation for rhythmic analysis.

subband envelopes of a musical signal. This is accomplished by performing a subband analysis of the music, and also of a white-noise signal from a pseudo-random generator. The amplitude of each band of the noise signal is modulated with the amplitude envelope of the corresponding band of the musical filterbank output, and the resulting noise signals are summed together to form an output signal.

For many frequency filterbanks and envelope calculations, the resulting noise signal has a rhythmic percept that is very similar to the original music signal. Even if there are very few, very broad bands (for example, four three-octave bands covering the audible spectrum), and the subband envelopes are low-pass filtered at 10 Hz, the tempo and beat characteristics of the original signal are instantly recognizable (Sound Example 3-1)[12].

The only thing preserved in this transformation is the amplitude envelopes of the filterbank outputs. Therefore, only this information is necessary to extract tempo and beat from a musical signal. Algorithms for beat extraction can be created that operate only on the envelope signals, and notes are not a necessary constituent for hearing rhythm. This is a vast reduction of the input data from the original signal. Shannon (1995) has reported a similar effect for the perception of speech.

Some other simplifications are not possible without changing the rhythmic perception of the stimulus. For example, if only one band is used, or the subband envelopes are linearly combined before modulating the noise (Figure 4-2), a listener can no longer perceive the rhythmic content of many signals (Sound Example 3-2). Thus, it seems that separating the signal into subbands and maintaining the subband envelopes separately is necessary for accurate rhythmic processing.

[12] Audio examples for my dissertation can be found on my Media Lab WWW site at http://sound.media.mit.edu/eds/thesis/ .

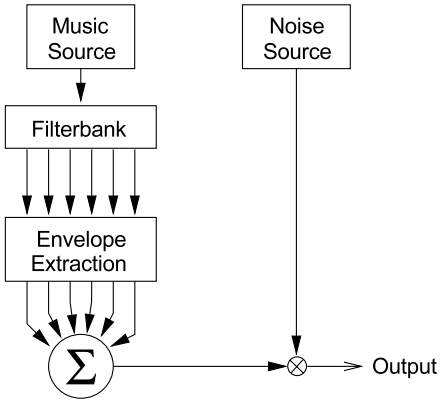


Figure 4-2: A noise signal that does not have the same rhythmic characteristics as the musical input, indicating that the sum of the amplitude envelopes is not a sufficient representation for rhythm analysis. Certain types of nonlinear combination by frequency channel are evidently present in the beat perception facility.

Stated another way, the algorithm in Figure 4-2 is a method for generating new signals whose representation under a filterbank-envelope-and-sum process is the same as a given piece of music. However, since these new signals are generally not perceptually equivalent to the originals, the filter-envelope-sum framework must be *inadequate* to represent data in the musical signal that is important for rhythmic understanding. This fact immediately leads to a psychoacoustic hypothesis regarding rhythmic perception: some sort of cross-band rhythmic integration, not simply summation across frequency bands, is performed by the auditory system.

4.2. Description of a Beat-tracking Model

The beat-tracking algorithm that I will present in this section bears most resemblance to the method of Large and Kolen (1994), in that it uses a network of resonators to phase-lock with the beat of the signal. However, my method is somewhat different. The resonators I use are analytically much simpler than theirs, a bank of resonators is used rather than gradient descent with a single adaptive resonator, and more pre- and post-processing of the signal is necessary, as the present model operates on acoustic data rather than an event stream.

A rhythmic beat is described in terms of its *frequency* and *phase*, just as a periodic sound waveform is. The frequency of the beat in a rhythmic musical signal is the tempo or perceived rate of the music, and the phase of the beat indicates where the downbeat of the rhythm occurs. That is, if the times at which a pulse occurs are defined to have zero phase, then the points in time exactly in-between pulses have phase of π radians.

While human pitch-recognition is only sensitive to signal phase under certain unusual conditions (Moore, 1997, pp. 97-100), rhythmic response is crucially a phased phenomenon. Tapping on the beat is not at all the same as tapping against the beat, or slightly ahead of or behind the beat, even if the frequency of tapping is accurate. Even when no overt behavior such as tapping is exhibited, it seems from introspection that during the presentation of signals with strong beats, listeners are aware of the current phase of the beat most of the time.

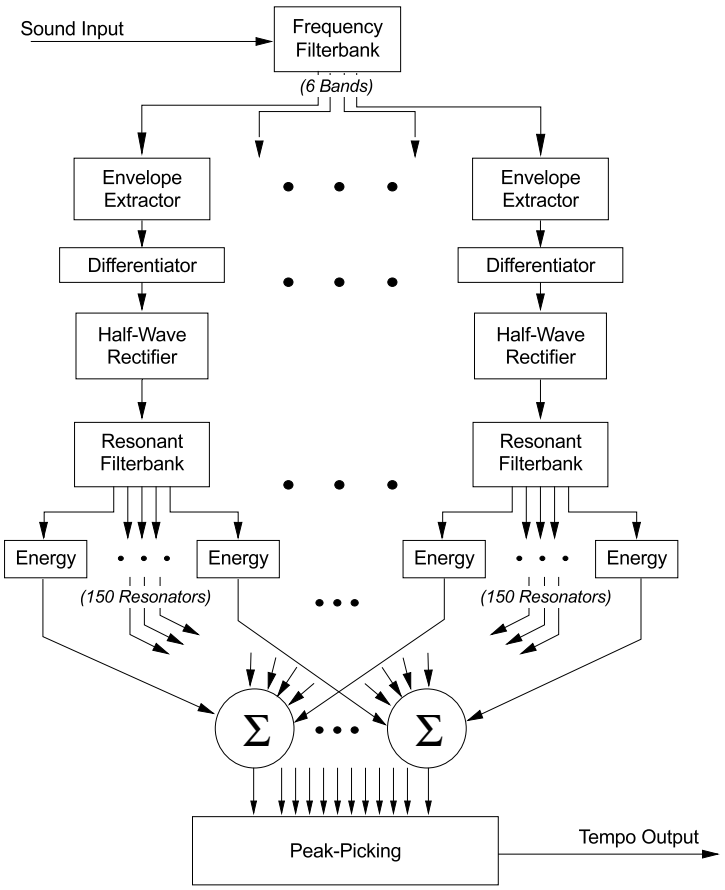


Figure 4-3: Schematic view of the beat-extraction algorithm.

Figure 4-3 shows an overall view of my tempo-analysis algorithm as a signal flow network. I will describe its method of operation briefly here, and then present more details piece-by-piece in the subsequent sections. These techniques were developed empirically via experimentation; however, in Section 4.5 I will discuss their relationship to other models of rhythm perception and other psychoacoustic behaviors.

As the signal comes in, a filterbank is used to divide it into six bands. For each of these subbands, the amplitude envelope is calculated and the derivative taken and half-wave rectified. Each of the envelope derivatives is passed on to another filterbank of 150 *tuned resonators*. In each resonator filterbank, one of the resonators will phase-lock—the one for which the resonant frequency matches the rate of periodic modulation of the envelope derivative.

The outputs of the resonators are examined to see which ones are exhibiting phase-locked behavior, and this information is tabulated for each of the bandpass channels. The tabulations are summed across the frequency filterbank to arrive at the frequency (tempo) estimate for the signal, and reference back to the peaks in the phase-locked resonators is used to determine the beat phase of the signal.

4.2.1. Frequency analysis and envelope extraction

As discussed in Section 4.1, envelopes extracted from a small number of broad frequency channels are sufficient information to rhythmically analyze a musical signal, at least for human listeners. Further, I have found through empirical studies of the use of various

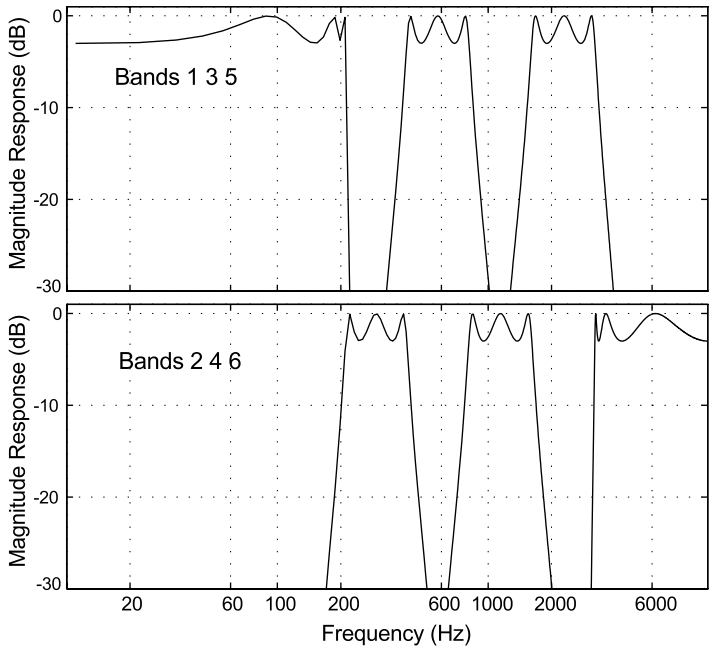


Figure 4-4: Magnitude response of the frequency filterbank used in the system, plotted in two pieces for clarity. The upper plot shows the first, third, and fifth bands; the lower, the second, fourth, and sixth. Each filter is a sixth-order elliptic filter, with 3 dB of passband ripple and 40 dB of stopband rejection

filterbanks with this algorithm that it is not particularly sensitive to the particular frequency tunings or filter implementations used. I expect that psychoacoustic investigation into rhythmic perception of amplitude-modulated noise signals created with various vocoder filterbanks would confirm that the same is true of human rhythmic perception.

The filterbank implementation used in the final version of the algorithm has six bands; each band has sharp cutoffs and covers roughly a one-octave range. The lowest band is a low-pass filter with cutoff 200 Hz; the next four bands are band-pass, with cutoffs at 200 Hz and 400 Hz, 400 Hz and 800 Hz, 800 Hz and 1600 Hz, and 1600 Hz and 3200 Hz. The highest band is high-pass, with cutoff frequency at 3200 Hz. Each filter is implemented using a sixth-order elliptic filter, with 3 dB of ripple in the passband and 40 dB of rejection in the stopband. Figure 4-4 shows the magnitude responses of these filters.

The envelope is extracted from each band of the filtered signal through a rectify-and-smooth method. The rectified filterbank outputs are convolved with a 200 ms half-Hanning (raised cosine) window. This window has a discontinuity at time *t=0*, then slopes smoothly away to 0 at 200 ms. It has a low-pass characteristic, with a cutoff frequency about 10 Hz ("frequency" in this case referring to envelope spectra, not waveform spectra), where it has a -15 dB response, and 6 dB/octave smooth rolloff thereafter.

The window's discontinuity in time means that it has nonlinear phase response; it passes slow envelope frequencies with much more delay than rapid ones. High frequencies, above 20 Hz, are passed with approximately zero delay; 0 Hz is delayed about 59 ms and 7 Hz advanced about 14 ms. Thus, there is a maximum blur of about 73 ms between these envelope frequencies. This spectral smearing is probably not enough to affect the tempo measurements, since the cross-band integration doesn't happen until after the within-band periodic energy estimation (see below).

This window performs energy integration in a way similar to that in the auditory system, emphasizing the most recent inputs but masking rapid modulation. Todd (1994) has examined the use of similar temporal-integration filters that are directly motivated by known psychoacoustic properties. After smoothing, the envelope can be decimated for further analysis; the next stages of processing operate on the envelopes downsampled to 200 Hz. There is little energy left in the envelope spectra at this frequency (since the smoothing filter is rejecting most envelope energy above 20 Hz), but it aids the phase-estimation process (see below) to maintain oversampled resolution for the envelopes.

After calculating the envelope, its first-order difference function is calculated and half-wave rectified; this rectified difference signal is what will be examined for periodic modulation. The derivative-of-envelope function performs a type of onset filtering process (see, for example, Smith's work on difference-of-Gaussian functions for onset segmentation in (Smith, 1994)) but the explicit segmentation, thresholding, or peak-peaking of the differenced envelope is not attempted. The modulation detectors in the next stage of the algorithm are sensitive to imperfections in an onset track. The half-wave rectified envelope difference avoids this pitfall by not making strict decisions, instead having broad (in time) response to onsets in the input signal. This process is similar to detecting onset points in the signal bands, and then broadening them via low-pass filtering.

Figure 4-5 shows the envelope extraction process for one frequency band in each of two signals, a 2 Hz click track and a polyphonic music example. The lowest band is shown for the click track, and the second-highest for the music track.

4.2.2. Resonators and tempo analysis

After the envelope has been extracted and processed for each channel, a filterbank of comb-filter resonators is used to determine the tempo of the signal. While comb filters are often used in reverberators and other sorts of audio signal processing, they also have properties that make them suitable resonators in the phase-locking pulse extraction process.

We can understand the resonant properties of comb filters by stimulating them with pulse trains of various rates. Depending on the relationship between the rate of the input pulse train and the "best resonance" of the comb filter, the long-term magnitude of response differs. Suppose that we stimulate a comb filter having delay T and gain α ($|\alpha| < 1$) with a right-sided pulse train of height A and period κ (that is, the sequence whose value is A at time 0, κ, 2κ, ... and 0 elsewhere). The comb filter resonates best to this input sequence when $T = \kappa$.

Let x_t and y_t be the input and output signals at time t; the equation of the filter is then

$$y_t = \alpha\, y_{t-T} + (1-\alpha)\, x_t \tag{4-1}$$

and so, by expanding the recurrence and plugging in the input signal,

$$\begin{aligned} y_0 &= (1-\alpha)A \\ y_\kappa &= \alpha(1-\alpha)A + (1-\alpha)A = (1-\alpha)A(1+\alpha) \\ y_{2\kappa} &= (1-\alpha)A(\alpha^2+\alpha+1) \\ &\vdots \\ y_{n\kappa} &= (1-\alpha)A(\textstyle\sum_{i=0}^{n}\alpha^i). \end{aligned} \tag{4-2}$$

Thus,

$$\lim_{n\to\infty} y_\kappa = \frac{(1-\alpha)A}{1-\alpha} = A. \tag{4-3}$$

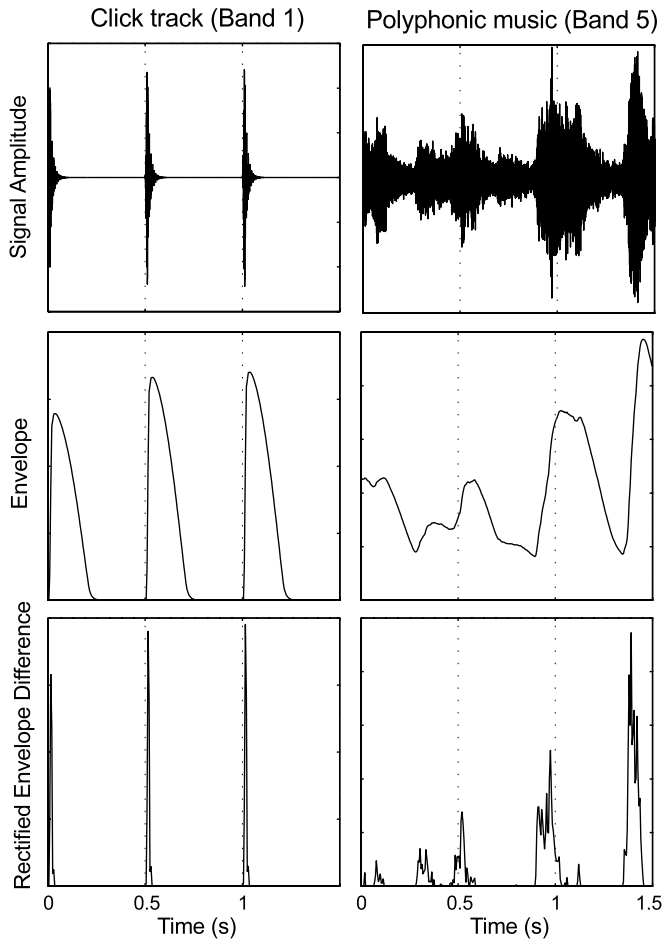


Figure 4-5: Envelope extraction process, for a 2 Hz click track (left) and a polyphonic music example (right). The top panels show the audio waveforms; the middle panels, the envelopes; and the bottom, the half-wave rectified difference of envelopes. The lowest filterbank band is shown for the click track, the second-highest for the music. Note that for the clean signal, the envelope difference corresponds closely to the signal onsets. However, this is not the case for the polyphonic music signal. If we picked onsets from the envelope signal, we would discard information that is present in the rectified envelope difference shown. This information, once discarded, cannot be recovered at later stages of processing.

On the other hand, if $T \neq \kappa$, the convergence is to a smaller value. Let λ be the least common multiple (common period) of T and κ; then there is only reinforcement every T / λ periods, and by a similar logic as the above,

$$\lim_{n\to\infty} y_{n\lambda} = \frac{(1-\alpha)\mathrm{A}}{1-\alpha^{T/\lambda}}, \tag{4-4}$$

and since $|\alpha| < 1$ if the filter is to be stable, and $T / \lambda \geq 1$,

$$1-\alpha^{T/\lambda} \geq 1-\alpha. \tag{4-5}$$

So a filter with delay matching (or evenly dividing) the period of a pulse train will have larger (more energetic) output than a filter with mismatched delay.

This is true not only for pulse trains, but for any periodic signal, as can be seen by doing a similar analysis in the frequency domain. The comb filter with delay *T* and gain α has magnitude response

$$\left|H(e^{j\omega})\right| = \left|\frac{1-\alpha}{1-\alpha e^{-j\omega T}}\right| , \tag{4-6}$$

which has local maxima wherever $\alpha\, e^{-j\omega T}$ gets close to 1. This occurs near the *T*-th roots of unity, which can be expressed as

$$e^{-j2\pi n/T}, 0 \le n < T . \tag{4-7}$$

These frequency-domain points are exactly those at which a periodic signal of period *T* has energy. Thus, the comb filter with delay *T* will respond more strongly to a signal with period *T* than any other, since the response peaks in the filter line up with the frequency distribution of energy in the signal.

For each envelope channel of the frequency filterbank, a filterbank of comb filters is implemented, in which the delays vary by channel and cover the range of possible pulse frequencies to track. The output of these resonator filterbanks is summed across frequency subbands. By examining the energy output from each resonance channel of the summed resonator filterbanks, the strongest periodic component of the signal may be determined. The frequency of the resonator with the maximum energy output is selected as the tempo of the signal. This is shown schematically in Figure 4-6.

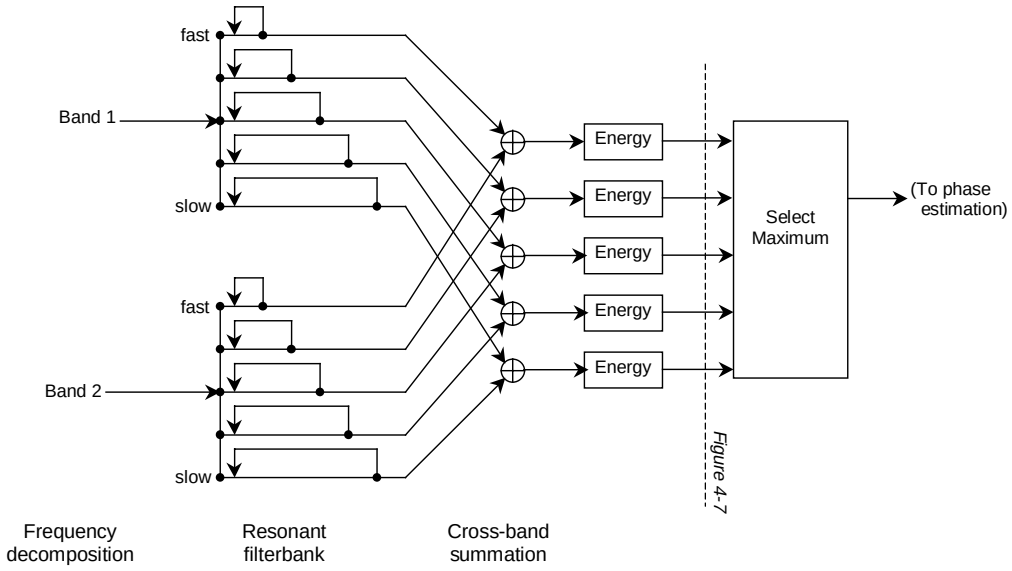


Figure 4-7

Figure 4-6: The resonant filterbanks that are used to match the tempo of the signal. Within each subband, a bank of comb filters consists of several parallel filters, each tuned to a different resonant frequency. This schematic shows two bands with five comb filters each; in the actual implementation, there are six bands with 150 resonators each. In this schematic, the lengths of the delay lines in the comb filters indicate the delay of the filter. The gain on each comb filter is tuned so that all filters have equal half-energy time in their impulse responses. The outputs from corresponding comb filters are summed across frequencies, and the resonant frequency with the strongest response is determined. The period of this resonator is the tempo of the signal, and the summed resonator outputs are used to determine the phase of the beat.

The α parameter for each comb filter is set differently, so that each filter has equivalent half-energy time. That is, a comb filter of period T has an exponential curve shaping its impulse response. This curve reaches half-energy output at the time t when $\alpha^{T/t} = 0.5$. Thus, α is set separately for each resonator, at $\alpha = 0.5^{t/T}$. A half-energy time of 1500-2000 msec seems to give results most like human perception.

Figure 4-7 shows the summed comb-filterbank output for a 2 Hz pulse train and for a polyphonic music example. The horizontal axis is labeled with "metronome marking" in beats per minute; this is a direct mapping of the delay of the corresponding comb filter. That is, for the 200 Hz power envelope signal, a feedback delay of 100 samples corresponds to a 500 msec resonance period, or a tempo of 120 bpm.

In the pulse train plot in Figure 4-7, a clear, large peak occurs at 120 bpm, and additional smaller peaks at tempi that bear a simple harmonic relationship (3::2 or 4::5, for example) to the main peak. In the music plot, there are two peaks, which correspond to the tempi of the quarter note and half note in this piece. If the width of the upper plot were extended, a similar peak at 60 bpm would be visible.

As can be seen in Figure 4-7, the differences in timing between the various comb filters is very small. If there are 150 comb filters spanning the range from 60—180 BPM (that is, 1-3 Hz), then they have an average difference of 13.3 ms. This is why the envelopes must be oversampled; if the envelope signals were only conveyed at a low sampling rate, such as 50 Hz, fractional-delay comb filters would have to be used to achieve such tight spacing of resonant frequencies. The implementation is simplified by oversampling the envelopes so that integer-delay filters can be used instead.

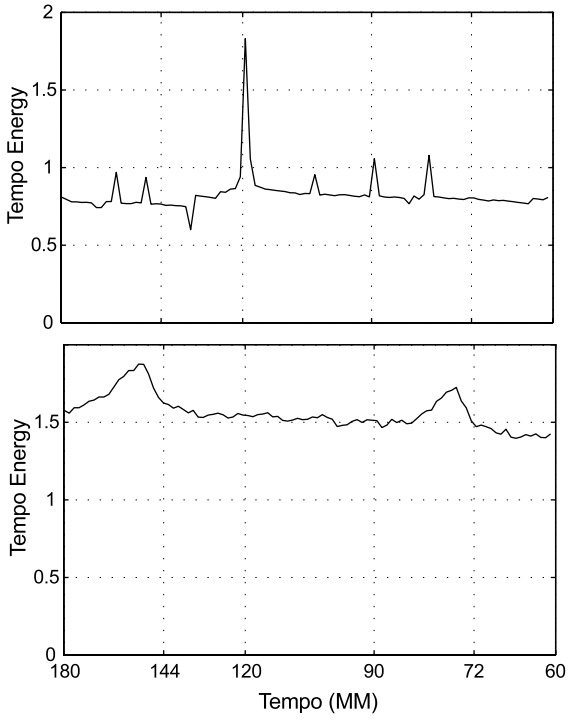


Figure 4-7: Tempo estimates, after tracking 5 sec of a 2 Hz (120 BPM) click track (top) and of a polyphonic music example (bottom). These curves are the values computed as the cross-band energy output of the resonant filterbank; that is, the energy at the "Select Maximum" stage of Figure 4-6. The x-axes correspond to the sampling of the time delays of the comb filters—there is one point on the x-axis for each comb filter in the resonant filterbank. The x-axes are labeled in beats per minute, that is, 120 BPM = 2 Hz. The polyphonic music shows more overall energy, but the tempo is still seen clearly as peaks in the curve.

4.2.3. Phase determination

It is relatively simple to extract the phase of the signal once its tempo is known, by examining the output of the resonators directly, or even better, by examining the internal state of the delays of these filters. The comb filters in the resonator filterbank are implemented with lattices of delay-and-hold stages. The vector *w* of delays in each filter can be interpreted at a particular point in time as the "predicted output" of that resonator. That is, the *w* vector contains the next *n* samples of envelope output that the filter would generate in response to zero input, where *n* is the period of the filter.

The sum of the delay vectors over all frequency channels for those resonators corresponding to the tempo determined in the previous step is examined. The peak of this prediction vector is taken as the estimate of when the *next* beat will arrive in the input. The ratio $\omega = 2\pi(t_n - t) / T$, where t_n is the time of the next predicted beat, t the current time, and T the period of the resonator, is the phase ω of the tempo being tracked. The phase and period may be used to predict beat times as far into the future as desired.

The present implementation analyzes the phase in this way every 25 ms and integrates evidence between frames in order to predict beats. Since re-estimation occurs multiple times between beats (because 25 ms is much shorter than the beat period), the results from each phase analysis can be used to confirm the current prediction or adjust it as needed. Currently, this prediction/adjustment is done in an ad-hoc manner. If several successive frames make the same beat prediction within a certain tolerance all of these estimates are averaged to arrive at the final prediction. This stage would be the appropriate one for the inclusion of high-level information, non-deterministic elements, or more sophisticated rhythmic modeling; see section 4.5.3.

Figure 4-8 shows the phase peaks for a 2 Hz pulse train, and for a polyphonic music example. In the upper plot, as the tempo is 120 bpm, the x-axis covers the next half-second of time; and for the lower plot, the estimated tempo is 149 bpm (see Figure 4-7), so one period is approximately 400 ms.

4.2.4. Comparison with autocorrelation methods

There are analytical similarities between this bank-of-comb-filters approach and previous autocorrelation methods for modeling the perception of tempo (Vercoe, 1997). Insofar as they are both ways of detecting periodic energy modulations in a signal, they are performing similar calculations. However, there are several advantages to expressing these operations as multiple comb filters over expressing them as autocorrelation.

Predominantly, comb filtering implicitly encodes aspects of rhythmic hierarchy, where autocorrelation does not. That is, a comb filter tuned to a certain tempo τ has peak response to stimuli at tempo τ, but also lesser response to stimuli with tempi at multiples (2τ, 3τ), fractions ($\tau/2$, $\tau/3$), and simple rational relationships ($3/2\,\tau$, $3/4\,\tau$, etc.). The autocorrelation only has this shared response for fractional tempi, not multiples or rationally-related tempi. An autocorrelation model asserts that a click track at 60 bpm gives no sense of tempo at 120 bpm, which seems intuitively wrong. The comb filter model asserts instead, that there is such a sense, but a reduced one when compared to a click track to 120 bpm.

Autocorrelation methods are zero-phase, which means that some other method of determining signal phase must be used. Vercoe (1997) claimed the use of a "phase-preserving narrowed autocorrelation," but neither he nor the source he cites (Brown and Puckette, 1989) explains what this means. The comb filtering method shown here is phase-preserving, and so provides a way of simultaneously extracting tempo and phase, as discussed in the previous section. The fact that the tempo and phase representations arise together gives us additional

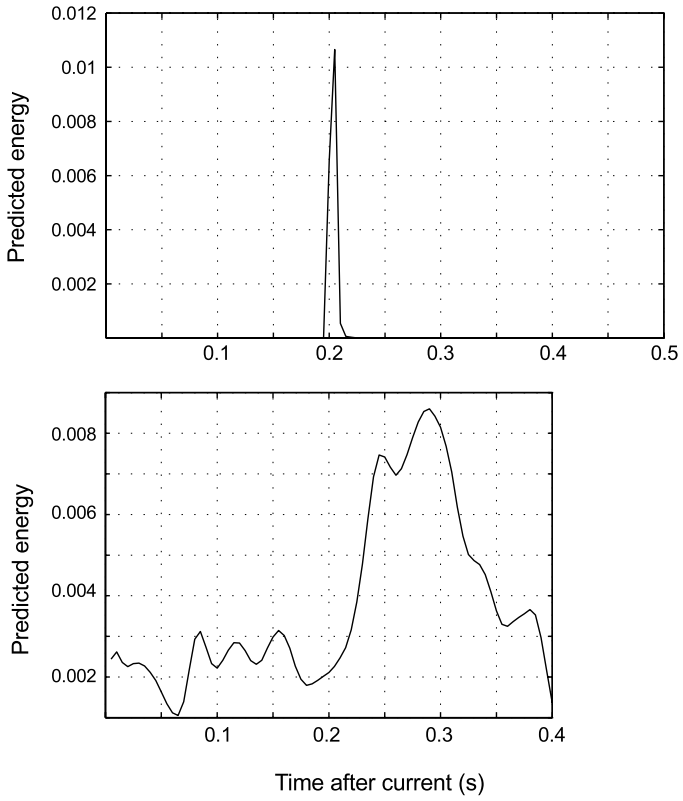


Figure 4-8: Phase estimates, after tracking 5 sec of a 2 Hz click track (top) and a polyphonic music example (bottom). The x-axis in each case covers the next full period of the resonator tracking the tempo, and the peak of the curve shows where the next beat is predicted to occur: about 210 msec in the future for the upper case, and 290 msec for the lower.

advantages for constructing further processing algorithms based on the output of the beat-tracker, as will be presented in Chapter 6.

One advantage of autocorrelation schemes is that they are more efficient in memory usage than banks of comb filters. The various lags can all access the same delay line—this is why the autocorrelation is zero-phase—whereas each comb filter must maintain a delay line of its own. In return for the extra memory usage, the comb filters provide estimates of output energy at each phase angle of each lag, where the autocorrelation accumulates it and only presents the summary.

Ultimately, it is representationally satisfying to have the frequency and phase of the signal explicitly encoded in the processing units of the algorithm. In an autocorrelation methodology, the rhythmic oscillations of the signal are only represented as post-processed summary results; whereas in the comb filtering method, the filter states themselves explicitly represent the rhythmic content—that is, there is an element of the processing network that phase-locks to and oscillates in synchrony with the signal. This is the key similarity between the present technique and that of Large and Kolen (1994).

4.3. Implementation and Complexity

The algorithms described above have been implemented in C++ code; the resulting program causally processes audio files captured from compact disks or other audio recordings, or coming in via a live microphone input. In this section, the parameters available for controlling the speed and accuracy of the program are described.

4.3.1. Program parameters

The current implementation of the system has a number of parameters that can be used to control the accuracy/speed relationship of the algorithms. The program will run in real-time on a modern desktop workstation such as a Pentium III, depending on the settings of these parameters and the sampling rate of the incoming audio stream. It is also clear from to the highly parallel structure of Figure 4-3 that the algorithm could efficiently make use of a multiple-processor architecture. This has not yet been accomplished, however.

There are four major areas where the performance and accuracy of the system can be tuned, and control over three of them has been implemented. The algorithm has been tested for audio at sampling rates from 8 kHz to 44.1 kHz and gives roughly equivalent qualitative performance in all of these.

Frequency filterbank

As discussed in section 4.1, there is likely a fair amount of latitude in choosing a frequency filterbank for decomposing the incoming audio stream without affecting human rhythmic perception. The speed of the system will vary a great deal with the complexity of these filters (since there is a fair CPU load for implementing high-order filters in real-time on high-bandwidth audio), and their number (since for each of the frequency channels, a full resonator filterbank structure is implemented).

The performance of the beat-tracking program using filterbanks other than the 6-channel 6th-order IIR filterbank described above has not been tested.

Envelope sampling rate

The decimation rate of the channel envelopes affects the speed and performance of the system. There are two major implications for using a slow envelope sampling rate. First, there are many resonator frequencies that cannot be represented accurately with integer delays in the comb filters. Second, the phase extraction can only be performed with accuracy equal to the envelope sampling rate, since the vector of delays has the same sampling rate.

In tradeoff to this, using a fast sampling rate for the envelopes entails a lot of work in the comb filtering, since the number of multiplies in each comb filter varies proportionately to this rate. Empirical testing over a variety of musical examples suggests that the envelopes should be sampled at least 100 Hz or so for best performance.

Number of resonators per frequency channel

The amount of computation required to track and analyze the comb-filter resonators varies directly with their number. If too few resonators are used, however, a problem develops with sampling the tempo spectrum too sparsely. That is, since each resonator is attempting to phase-lock to one particular frequency (not to a range of frequencies), if there is no resonator tuned close to the tempo of a particular signal, that signal cannot be accurately tracked.

Also affecting this sparsity consideration is the range of resonator frequencies to be tracked. The wider the range of tempi to track, the sparser a fixed number of resonators will spread over that range.

Good results have been generated using a bank of 150 resonators for each channel, covering a logarithmically-spaced range of frequencies from 60 bpm (1 Hz) to 240 bpm (3 Hz).

Analysis frame rate

In this particular implementation, a higher-level averaging scheme is used to decide where (at what times) to deduce beats in the input signal. That is, for each analysis frame, the phases of the resonators are examined; the evidence here suggests future beat locations. These suggestions are combined over multiple analysis frames; when several frames in a row point to the same future beat location, evidence accumulates for that time, and a beat is actually assigned there.

Thus, the frequency with which the procedure of examining and summing the outputs and internal states of the resonators is executed has a strong effect upon the performance and speed of the program. Good results can be obtained if the analysis frame rate is at least 15 Hz.

Real-time performance cannot be obtained with the parameter values shown above; on a 330 MHz Pentium II using unoptimized filtering and analysis code, with an envelope rate of 100 Hz, 75 resonators per subband, and frames of beat predictions at analyzed every 15 Hz, the required performance for real-time operation on 22 kHz input is reached. This real-time performance includes reading the sound file from disk and playing it back with short noise bursts added to highlight the beats. At this level of accuracy, the algorithm still performs acceptably well on some, but not all, musical examples.

4.3.2. Behavior tuning

The behavior of the algorithm can be tuned with the α parameters in the comb filters. These parameters can be viewed as controlling whether to value old information (the beat signal extracted so far) or new information (the incoming envelopes) more highly. Thus, if α is large (close to unity), the algorithm tends to "lock on" to a beat, and follow that tempo regardless of the new envelope information. On the other hand, if α is small, the beat-track can be easily perturbed by changes in the periodicity of the incoming signal. Manipulating these parameters for the comb filter structure is computationally similar to manipulating the windowing function of a narrowed autocorrelation (Brown and Puckette, 1989).

Higher-level or domain-specific knowledge could be used to set this parameter based on prior information. For example, in rock or pop music, the beat is usually quite steady, so a high value for α would be appropriate; while for classical music, particularly styles including many tempo changes, a smaller value would be more better.

4.4. Validation

It is difficult to evaluate the performance of an ecological beat-tracking model, for there are few results in the literature dealing with listeners' tempo responses to actual musical excerpts. Most psychophysical research has dealt primarily with special cases consisting of simple tones in unusual temporal relationships, which will typically be more difficult to track than "real music" for a listener. In contrast, most computerized beat-tracking systems have been evaluated informally, by using a small number of test cases (whether acoustic or MIDI-based) and checking that the algorithm "works right".

In this section, the performance of the algorithm is evaluated in both qualitative and quantitative manners. I provide results to demonstrate the qualitative performance for 60 ecological music excerpts, with sound examples publicly available for listening. I also present results from a short validation pilot experiment that was conducted to investigate the

degree to which the performance of the algorithm is like the performance of human listeners on a rhythmic-tapping task.

4.4.1. Qualitative performance

Examples of many different types of music have been tested with the implemented algorithm, using a short application that reads a sound sample off of disk, causally beat-tracks it, and writes a new sound file with clicks (short noise bursts) added to the signal where beats are predicted to occur. I have made these sound files available for listening via the World Wide Web ("Beat-tracking results" page), and summarized the results below.

The wide set of input data contains 60 examples, each 15 seconds long, of a number of different musical genres. Rock, jazz, funk, reggae, classical, "easy-listening," dance, and various non-Western musics are represented in the data set and can be tracked properly. Some of the examples have drums, some do not; some have vocals, some do not. Five of the examples would likely be judged by human listeners to have no "beat." For these cases, the algorithm would be performing unlike human listeners if it gave "correct" results. I recorded the input data from a radio tuner in the San Francisco area during the summer of 1997; they are the same examples used to test the speech/music discriminator that I reported previously (Scheirer and Slaney, 1997).

In Table 4-1, I summarize the results by musical genre; some qualitative descriptions of typical results are provided below.

I qualitatively classify 41 of 60 samples (68%) as tracked accurately, and another 11 (18%) as being tracked somewhat accurately. This accuracy percentage is not directly comparable to that reported for other systems, because the data set used here is more difficult. All of the "easy" cases of rock-and-roll with drums keeping a straightforward beat were tracked correctly; and 5 of the 8 examples not tracked accurately were those that would probably not be judged by human listeners to have any beat at all. It is premature to interpret these results as indicative of consistent genre-to-genre differences in accuracy; there are too few examples and the within-genre differences in accuracy too great.

Genre	# cases	Correct	Partial	Wrong
Rock	7	3	3	1
Country	3	3	0	0
Urban	9	7	1	1
Latin	5	3	2	0
Classical	9	4	4	1
Jazz	8	3	1	4
Quiet	3	2	0	1
Reggae	2	2	0	0
Non-Western	4	4	0	0
Total	60	41	11	8

Table 4-1: Performance of the beat-tracking algorithm, summarized by musical genre. Results were auditioned and classified into groups by qualitative success level. "Correct" indicates that the algorithm found a steady beat in the signal and locked onto it. For several of the examples, human listeners do not find such a beat, and so the "Correct" answer does not reflect human behavior. "Urban" styles include rap, funk, and R & B music; "Quiet" includes muzak and an "easy-listening" example. All sounds are available via the WWW for independent evaluation ("Beat-tracking results" page).

For the examples that the algorithm tracks consistently, there is a startup period between 2 to 8 seconds long, in which the resonant filters have not yet built up an accurate picture of the signal. After this period, for most signals, the algorithm has settled down and begun to track the signal accurately, placing the clicks in the same locations a human listener would. Examining some of the other, incorrectly-tracked examples, is instructive and highlights some of the deficiencies of this method.

Examples #1, #2, and #57 are all up-tempo jazz cases in which human listeners do perceive a strong beat, but no beat is ever extracted by the system. In these three cases, the beat is described by syncopated instrumental lines and complex drum patterns. That is, there is not actually very much energy modulating at the frequency that is the perceptual beat tempo for humans. Human listeners have a great ability to induce "apparent" frequencies from complicated modulation sequences. For these examples, the algorithm is not able to find a pulse frequency, and so the beat output is more-or-less random.

The same in apparent in example #37, which is a pop tune that has a mixed or *clave* beat—the beat is not even, but subdivided into oddly-spaced groups. Each two measures, containing 16 eighth notes between them, are divided into a 3-3-3-3-2-2 pattern. A human listener has no trouble understanding the relationship between this pattern and a more common 4-4-4-4 pattern, but the algorithm assumes that the groups of three are the basic beat, and then gets confused when the pattern doesn't come out right.

Among the examples judged as partly correct, the most common problem is phase shifting. For example, in example #16, a jazz piano trio, the algorithm estimates the tempo correctly, but switches back and forth between assigning beats to the upbeat and the downbeat Although this behavior is not unlike some human jazz listeners, a human would likely be more consistent, by deciding where to place the beat and then sticking to it. This behavior could probably be corrected by adding a small amount of high-level knowledge to the beat-tracking system.

Similar to this, in example #7, a rhythm and blues tune, the algorithm is uncertain about assigning the beat to the quarter-note pulse or to the eighth-note pulse, and so switches back and forth between them. A human listener might also suffer from similar confusion, but would likely make an arbitrary decision and then stay with it unless the music changed radically.

Other than these two sorts of confusions for certain rhythmically complex musics, the algorithm seems to perform quite successfully at tracking the musical beats.

Tempo modulation

Todd (1994) observes that to be an accurate model of human rhythm perception (and, of course, to be maximally useful as a music analysis tool), a beat-tracking system must be robust under expressive tempo modulation. The algorithm described here is able to follow many types of tempo modulations; this is effected in the signal processing network by simply examining, over time, the resonator producing the most energetic output. That is, when the tempo of a signal modulates, the response of the resonator corresponding to the old tempo will die away, and that of the resonator corresponding to the new tempo will gain.

Figure 4-9 shows "tempo curves" for two expressively modulated performances of a piece of music (Keith Jarrett and Andras Schiff performance, of the beginning of the G-minor fugue from book I of Bach's *Well-Tempered Clavier* [sound example 4-3]). Each curve shows the change in the resonant frequency that best matches the input signal over time. The algorithm is quite sensitive to the variations in tempo over time. Although the notion of tempo curves as a model of the perception of tempo has been criticized (Desain and Honing, 1992b), this example shows that this method for analyzing the tempo of real musical examples at least is sensitive to changes in tempo over time.

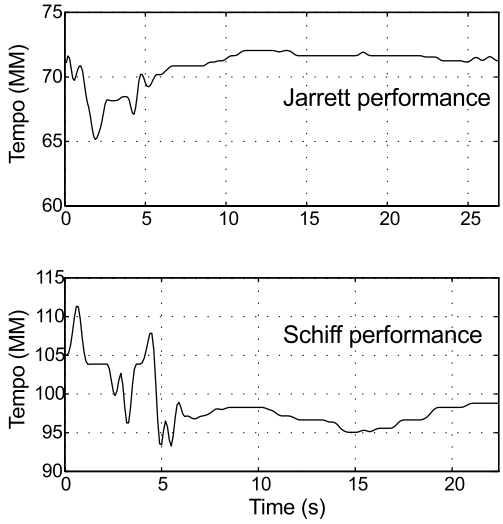


Figure 4-9: "Tempo curve" for two performances of the same piece of music. Each tempo-track has a short startup period during which the tempo estimation is unstable; after that there are clear differences in the two performances. The time-scales are slightly different to make the performance scales align (the same musical excerpt is used in both cases).

Of course, whether or not these tempo curves are "right"—that is, whether they match the perception of time-varying tempo by a particular listener or set of listeners—is a very difficult question to address. This is the topic, although considered only for simpler cases, of the next section.

4.4.2. Validation Experiment

I conducted a short validation experiment to confirm the qualitative results given in the previous section. I do not intend this experiment to highlight important psychoacoustic effects in beat perception, but only as an investigation into whether or not the beat-tracking algorithm performs generally like a human listener.

Subjects

Five adult listeners, all graduate students and staff members at the MIT Media Laboratory, participated in the experiment. All were experienced musicians with normal hearing.

Overview of procedure

Subjects listened to seven musical examples, drawn from different musical genres, through headphones. They indicated their understanding of the beat in the music by tapping along with the music on a computer keyboard.

Materials

Seven musical excerpts from the above set were used. Each was digitally sampled from an FM radio tuner to produce a monophonic 22KHz sound file, 15 seconds long. A computer interface was created on a DEC Alpha workstation with which the musical excerpts were presented to subjects at a comfortable listening level over AKG-K240M headphones.

The musical excerpts were as follows: a Latin-pop song at moderately-fast tempo (#10), a jazz piano trio at fast tempo (#17), a "classic rock" song at moderately-slow tempo (#20), an

excerpt from a Mozart symphony at moderate tempo (#40), an "alternative rock" song at moderately-slow tempo (#45), and a piano etude with varying tempo (#56).

A click track "step function" was also created for the experiment, in which 10-ms white noise bursts were presented at a tempo of 120 bpm (interonset time of 500 ms) for 6 sec, then at a tempo of 144 bpm (interonset time of 417 ms) for 4.6 sec, then again at 120 bpm for 6 more seconds. This stimulus is used to evaluate the response of human listeners and the beat-tracking algorithm to sudden changes in tempo.

I assigned ground-truth beat times to each excerpt by listening repeatedly and placing "click" sounds in the perceptually appropriate positions[13]. This task was different than the tapping task in which the subjects participated; I listened repeatedly to each stimulus with a waveform editor, placing beats, listening to results, and adjusting the beat positions as necessary. This method of finding the "correct" beat placement must be considered to be more accurate and robust than the real-time tapping task, although there is little literature on humans providing either sorts of judgment (see (Snyder and Krumhansl, 1999), (Drake, 1998) and (Parncutt, 1994b) for three other "tapping tasks"). The ground-truth labeling was conducted separately from the tapping experiment. I did not know the results of the experiment or the algorithm execution while labeling, and the subjects were not presented with the ground-truth data.

Detailed procedure

Subjects were seated in front of the computer terminal and instructed in the task: they were to listen to short musical examples and tap along with them using the space bar on the keyboard. They were instructed to tap at whatever tempo felt appropriate to the musical excerpt, and to attempt to tap in equal intervals (a pilot experiment revealed that some subjects like to "drum along" in rhythmic or even syncopated patterns with the music if they are not instructed otherwise). They listened to a 120 bpm click-track as a training sample to indicate they understood the procedure, and then proceeded with each of the seven experimental trials.

All seven trials were run in the same sequence for each listener, in a single block. The experiment was not randomly ordered, based on the assumption that there is little practice effect in this task. After each trial, the subject was instructed by the interface to press a key different than the space bar to continue to the next trial. The entire experiment took approximately 5 minutes per subject. The computer interface recorded the time of each tap, accurate to approximately 10 ms, and saved the times to a disk file for analysis.

Finally, the beat-tracking algorithm was executed on each of these seven stimuli to produce beat times as estimated by the model described in the previous sections. These beat times were saved to a disk file and analyzed for comparison with the human beat times. The algorithm parameters were adjusted to give optimum performance for this set of trials, but not changed from trial-to-trial (although see Section 5.5.4).

Dependent measures

The human and algorithmic beat-tracks were analyzed in two ways. First, the beat placements were compared to the ideal ground-truth placements; then, the regularity of tapping was assessed by examining the variance of interonset times.

To compare the beat placements, a matching comparison was conducted. Each beat placed by a human subject or by the beat-tracking model was matched with the closest (in time) comparison beat in the ground-truth beat-track. Initially, only the beats that I actually placed in constructing the ground-truth track were used, but since some subjects or the algorithm

[13] I am grateful to Dan Levitin of McGill University for the suggestion to use an expert analysis as the basis for quantitative evaluation.

tapped twice as fast as the ground truth on some examples, beats were also allowed to match the midpoint between ground-truth beats. The root-mean-square deviations of the subject's tap times from the nearest ground-truth tap times were collected for each subject and trial, averaging across taps within a trial.

This RMS deviation is a measure of how close the tapper came to the "ideal" beat locations. If it is very low, all of the tapper's placements were very close to ground-truth judgments; if high, the tapper's placements were randomly distributed compared to the ground-truth judgments.

This measure leaves open an important aspect of beat-tracking, which is regularity. As described in the qualitative results, the algorithm sometimes demonstrates unusual behavior by switching from one tempo to another, or from off-the-beat to on-the-beat, in the middle of a trial. To evaluate the regularity of tapping, the variance of interonset interval was calculated for each trial-by-subject, each trial by the model, and each trial in the ground-truth tracks. Note that, as described above, the human subjects were explicitly encouraged to tap regularly.

Again, the ground-truth behavior is taken as the standard of comparison; if the variance is larger for some subject than for the ground truth, it indicates that that subject's tapping was irregular relative to the ground truth. If the variance is smaller, it indicates that the tapping was more regular than the ground-truth. This does not necessarily mean "better"—in the case in which tempo is changing, to tap "correctly" subjects need to tap irregularly. However, most of the measured irregularity arose in this data from subjects leaving out beats. Each time a beat is left out, an inter-onset interval twice as large as the rest is added, increasing the variance.

Results and discussion

The beat-placement comparison is shown in Figure 4-10. Results indicate that the performance of the algorithm in placing beats in logical locations was at least comparable to the human subjects tested for all the musical cases; in four of the seven cases, the model was the most or second-most accurate tapper. This indicates that whenever a beat position was chosen by the algorithm, the position was very close to an ideal beat position as determined by the ground-truth judgment.

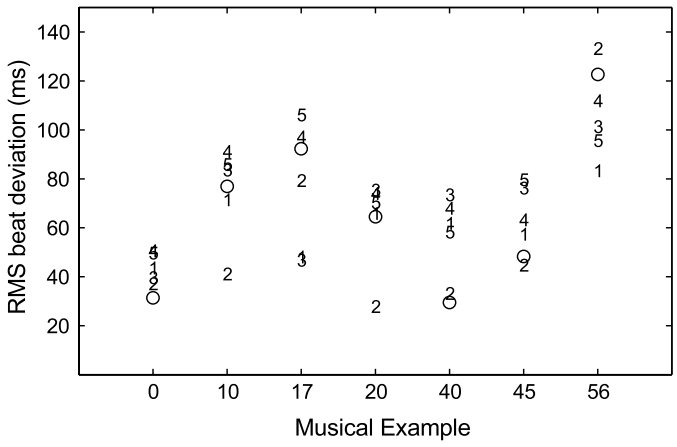


Figure 4-10: Scatter-plot of human (subj. number) and model (O) beat position accuracy for each of the seven experimental trials. Trial '0' corresponds to the click-track step function. Each point measures how accurate that subject was, relative to the ground truth (see text), in placing beats in time. The ground-truth judgments are at zero variance for each column. For each trial, the algorithm beat position was at least comparable to the performance of the human subjects Overall, the algorithm performance showed a highly significant correlation with the mean human performance ($r = .814$; $p(df=5) < 0.015$).

The regularity comparison is shown in Figure 4-11. Results here indicate that the algorithm was as regular as a human listener for five of the seven trials, and less consistent for two of the trials. In one case (#56), it and several of the human subjects were more consistent than the ground-truth indicated was appropriate. More post-hoc analysis is necessary to understand why the algorithm performance is irregular in these trials; preliminary results suggest that these two stimuli have relatively slow onsets carrying the beat (violins in one case, electronically gated drum sounds in the other). Note that the human listeners, evaluated as individuals, also show outlying behavior on certain trials. View the performance of subject #1 on musical example #20, or subject #3 on musical example #56.

These two results are consistent with the qualitative results described above. When the algorithm chooses to place a beat, it does so with great accuracy and musical relevance; however, for certain musical excerpts, it is somewhat inconsistent in its tapping regularity. That is, for these examples, it drops beats or shifts phase more often than a human listener. This is not a bad result, because it is exactly this inconsistency that could best be addressed by including high-level information in the model (such as simply including instructions to "try to tap regularly").

4.5. Discussion

In previous sections, the construction of a beat-tracking system has been approached from a largely empirical perspective. However, it is also valuable to compare the resulting algorithm to previous work on pulse perception in humans.

4.5.1. Processing level

Perhaps the most obvious difference between the method presented here and much of the previous work on beat-tracking is that this algorithm knows nothing about musical timbre, genres, or even notes or onsets. The beat-track that this model extracts is a continuous transformation from input to output, in keeping with the understanding-without-separation approach outlined in Chapter 3, Section 3.4. This approach to tempo analysis might also be

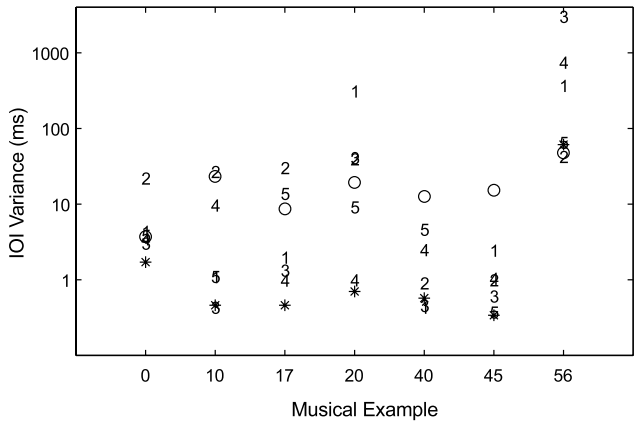


Figure 4-11: Scatter-plot of human (subj. number), model (O), and expert (*) IOI variances for each of the seven experimental trials. Trial '0' corresponds to the click-track step function. Each point shows the regularity of tapping of a subject for one trial; large values represent less regular tapping. For trials #40 and #56, the algorithm was not as consistent in tapping as a human listener. Overall, the algorithm performance showed a highly significant positive correlation with the mean human-subject performance, and both the algorithm and the human subjects showed highly significant positive correlations with the expert judgment ($r = 0.889$, $r=0.863$, $r=0.995$ respectively; $p(df=5) < 0.01$ in each case).

called a “perceptual model” of tempo, to contrast it with cognitive structuralist models.

That is to say, in models such as those developed by Povel and Essens (1985), Desain (1999) or Goto (1995; 1998), there are two stages of processing represented (the first is implicit in the Povel/Essen and Desain models). The first stage processes the acoustic stream, classifying the various pieces of sound into onsets and time-intervals, separating the streams of sound, and understanding the accent structure and timbre of various components. Then, the second stage places these events in relationship to each other in order to determine the tempo and beat phase of the signal.

In contrast to this, the model that I have presented agrees with the viewpoint of Todd (1994), in which tempo and rhythm are low-level “perceptual judgments” about sound, with little cognition or memory required for processing. This viewpoint is intuitively appealing for at least one major reason, which is that the tempo and beat of music can be processed in unattended auditory streams. Music listeners, even non-musicians, often have the experience of conducting a conversation and suddenly realizing that they have been tapping their foot to the background music. If the foot-tapping process required cognitive structuring of the input data, it seems likely that other cognitive hearing tasks such as speech-understanding would interfere with this ability.

Studies such as that of Povel and Essens (1985) have demonstrated convincingly that beat perception may be explained with a model in which a perceptual clock is aligned with the accent structure of the input. A clock model is fully compatible with the method proposed here; it seems natural and intuitive to posit such an internal clock. However, the Povel and Essens model of clock induction, and similarly the Parncutt model, relies heavily on structural qualities of the input, such as a sophisticated model of temporal accent, to function.

Todd has argued that such phenomena do not need to be modeled cognitively, but rather can be explained as natural emergent qualities of known psychoacoustic properties of masking and temporal integration. My model agrees here as well, and I have demonstrated empirically in the previous section that complex ecological musical signals can be accurately beat-tracked without any such factors explicitly taken into account. However, a more thorough evaluation of this model would include testing it on the unusual and difficult sequences tested in the course of developing accent models, to determine if changes to weighting factors or integration constants need to be made in order to replicate these psychophysical effects.

Of course it is the case that high-level knowledge, expertise, and musical preference have an effect on the human perception of beat. However, the results shown in the previous section demonstrate that quite a satisfactory level of explanatory performance can be produced with only a low-level, bottom-up model. Stated another way, the results here show at least that high-level, top-down information is not necessary to explain tempo and beat perception as well as this model does. The only remaining aspects of tempo perception for which high-level information or high-level models might be needed are those that cannot be explained with the present model.

4.5.2. Prediction and Retrospection

Desain's recent work on beat-tracking has included valuable discussion of the role of prediction and retrospection in rhythmic understanding. Clearly, prediction is a crucial factor in an accurate model of human rhythm perception, as simply to synchronize motor activity (like foot-tapping) with an auditory stream requires prediction. There is a pleasing symmetry between Desain's “complex expectancy” curves (Desain, 1995) and the phase-prediction vectors extracted here from the comb-filter delay lines (as in Figure 4-8).

Desain, citing Jones and Boltz (1989), draws attention to the utility of considering prediction and retrospection to be similar aspects of a single process. “Retrospection” refers to the manner in which new stimulus material affects the memory of previous events. Although

there is no retrospection included in the model—remembrance is an inherently cognitive process—the phase-prediction curves could be used as input for this process as well.

When evaluating any model of musical perception, it is important to keep in mind the complexity of introspection on musical phenomena. Although after-the-fact, a listener may have made a rhythmic model of the very beginning of a musical phrase, it is clear that this model must have arisen via retrospection, for there is not enough information in the signal alone to form it progressively. Simply because a listener feels that he understands the rhythm of the beginning of a musical segment does not mean that the beginning *itself* contains sufficient information to allow rhythmic understanding to occur. A natural extension to this model would be the ability to go back and reprocess previous moments in the signal in the light of later tempo analysis. Such reprocessing has much in common with the recent blackboard-systems approach to computational auditory scene analysis presented by Ellis (1996a) and Klassner (1996) and discussed in Section 3.4.1.

4.5.3. Tempo vs. Rhythm

This model cannot explain perceptual phenomena related to the grouping of periodic stimuli into a rhythmic hierarchy. There are many known effects in this area, ranging from the low-level, such as Povel and Okkerman's (1981) work on perceived accents in sequences of tones that are not physically accented, to very broad theories of generative rhythmic modeling such as the influential work of Lerdahl and Jackendoff (1983).

This model is compatible with and complementary to the bulk of this research, since most of these theories assume that a temporal framework has already been created. Synthesis of a model that operates from an acoustic source and one that includes musical assumptions and explanation should be possible. Such a joint model could represent a very robust theory of rhythmic understanding.

However, the model presented here should not be taken as attempting to explain rhythm perception as well as tempo; my viewpoint is rather that these processes are to some extent separable and may be addressed and modeled independently.

4.5.4. Comparison to other psychoacoustic models

The resemblance between this model of tempo, as shown in Figure 4-3, and modern models of pitch hearing, as discussed in Chapter 2, Section 2.1.1 is striking. Both models start with a filterbank; in the Meddis-Hewitt pitch model (Meddis and Hewitt, 1991) the filterbank is perceptually based (ERB gammatone filters), whereas in the vocoder pulse model it is not. The inner hair cells in the Meddis-Hewitt pitch model perform a similar function to the sequence of envelope extraction, differentiation, and rectification in the vocoder pulse model. The banks of comb filters are computationally similar to a narrowed autocorrelation (Brown and Puckette, 1989) for a infinitely-long exponentially decreasing window, although they preserve the phase of the input, whereas the autocorrelation is zero-phase. However, the energy calculation for the comb filter outputs also does not preserve phase, making the overall resonator-energy calculation process (considering tempo analysis only, leaving beat phase aside) nearly equivalent to the autocorrelation in the Meddis-Hewitt model. Finally, the cross-frequency summation is equivalent in the two models.

The obvious similarities between these two models naturally lead to the question of whether one model alone is adequate for explaining both processes[14]. I will examine this question in a bit more detail by implementing of a tempo-estimation algorithm as a quick-and-dirty variant on a pre-existing pitch-estimation procedure.

[14] I am grateful to Dan Ellis of the University of California at Berkeley for making this observation.

Algorithms from Slaney's Auditory Toolbox (Slaney, 1994) were used to implement the Meddis-Hewitt model of pitch perception as discussed in Section 2.1.1. The **MakeERBFilter** function was used to create a cochlear filterbank of 40 overlapping gammatone filters. A sound is processed with this filterbank and then passed to the **MeddisHairCell** function for IHC simulation. The IHC model output is smoothed and decimated by a factor of 20 to reduce the data size (see below). Finally, the **CorrelogramFrame** procedure is used to calculate the autocorrelogram of the signal based on this cochlear-model front end, and the rows of the correlogram summed to arrive at the summary autocorrelation.

Much longer lags must be used to calculate the correlogram for tempo analysis than for pitch analysis. The length of the input window for the correlogram must be as long as several periods of the input signal in order to show a strong peak in the summary autocorrelation. For pitch detection, periods are on the order of tens of milliseconds, but a pulse period might be as long as two seconds.

A great deal of memory and computation is therefore required to compute the "tempo periodogram" for a given signal. Ellis (1996a) suggested that a more perceptually-relevant correlogram might be computed with "log-lag" spacing on the lag axis, so that the logarithmic spacing of frequencies on the basilar membrane is echoed in the spacing of correlogram lag calculations. This approach is taken in the next chapter although it has the computational disadvantage of not being easily calculated using the FFT.

Figure 4-13 and Figure 4-12 show the outputs of various stages in the model. The input signal used was a "drum loop" taken from a compact disc used for popular music sampling and composition; the tempo is known (from the CD information) to be 100 beats per minute. The first four seconds of the signal, sampled at 11025 Hz, were used for analysis.

In both plots, commonality of pulse regions at lags corresponding to periodicities in the signal, in all frequency channels, can be clearly seen. Such similarity across frequency channels is not unexpected, since most of the energy in the input comes from broad-band drum sounds.

In the summary autocorrelation (Figure 4-12, bottom panel), we see a clear peak in the tempo periodogram at a lag of 0.6 seconds, corresponding to a tempo of 100 bpm. We also see peaks at fractions of this period, which correspond to the faster rhythmic divisions in the signal (eighth notes, sixteenth notes), as well as multiples of the fractions.

To examine this model for a case where the rhythm is not so clearly defined in the waveform, I conducted the same analysis on a five-second sample of a classical music track. This track contains strings, woodwinds, and French horns, but no drums, playing a piece of music that is immediately perceived to "have a beat". It is difficult to see this beat in the audio waveform (Figure 4-16); however, the excerpt was one of the ones tracked accurately in Section 4.4.1 (#40). I estimated the tempo (using a stopwatch and counting beats) as approximately 140 bpm, which corresponds to a lag of 0.42 seconds. Figure 4-16 and Figure 4-15 show the same computations on this signal as Figure 4-13 and Figure 4-12 showed for the drum signal. In this case, we see very little tempo evidence in the summary autocorrelation.

However, the correlogram (and thus, correlogram sum) is dominated by the high-energy frequency bands around channels 25-30 (near the 'cy' in 'Frequency Channel' on the y-axis of the plot). If we normalize the correlogram channels by energy, as shown in Figure 4-14, then the summary autocorrelation does show the tempo peaks, although still not as clearly as for the drum signal.

Thus, it seems that the Meddis-Hewitt model with normalized correlogram processing is adequate to track not only pitch of periodic signals with small period, but tempo in periodic signals with longer periods. This holds even for complex polyphonic music signals.

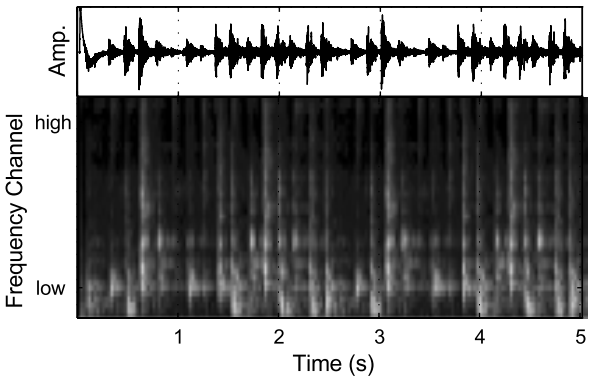


Figure 4-13: Cochlear model applied to a drum loop with known tempo of 100 beats per minute. The top panel shows the input signal waveform; it is aligned in time with the bottom panel, which shows the output of the inner-hair-cell model. The periodic structure of the input can be clearly seen in both cases; in addition, the lack of continuous energy in the hair-cell output shows the impulsive nature of the input.

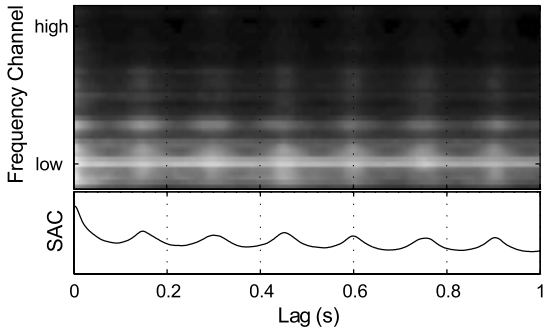


Figure 4-12: Autocorrelogram of the same signal shown in Figure 4-13. In the upper panel, each row is the autocorrelation of the corresponding row in the lower panel of Figure 4-13. In the lower panel, the summary autocorrelation or "tempo periodogram" is the vertical sum of the energy in each channel (compare to Figure 2-3, Chapter 2). There is a clear periodicity peak at lag 0.6 sec, corresponding to the tempo of 100 beats/minute.

The connection between pitch processing and one model of tempo processing, when described in this way, is clear. Both percepts may be explained with the same signal flow, in which a periodicity-detection algorithm is applied to the rectified outputs of a subband decomposition. If the model is run slowly, tuned to detect envelope fluctuations in the range 1-4 Hz, then tempo is detected. If the model is run quickly, tuned to detect envelop fluctuations in the range 50-2000 Hz, then pitch is detected.

These results are somewhat surprising—pitch and pulse seem intuitively like very different things—but naturally lead to the question of whether the auditory system might, in fact, be extracting them from a signal using the same methods. This question can not yet be answered, as there is not enough psychoacoustic and neurophysiological research into pulse perception to know exactly what the right features of a tempo extractor should be.

Put another way, in the search for computational models of pitch perception, we have been aided by a large number of pitch phenomena that must be reflected in the models: missing fundamental, harmonic shift, pitch of filtered noise, pitch of non-harmonic complex tones, and so forth. There are few if any analogous phenomena for tempo that are waiting to be explained, perhaps because tempo does not "feel" like an auditory process introspectively and has thus received little formal attention from psychoacousticians.

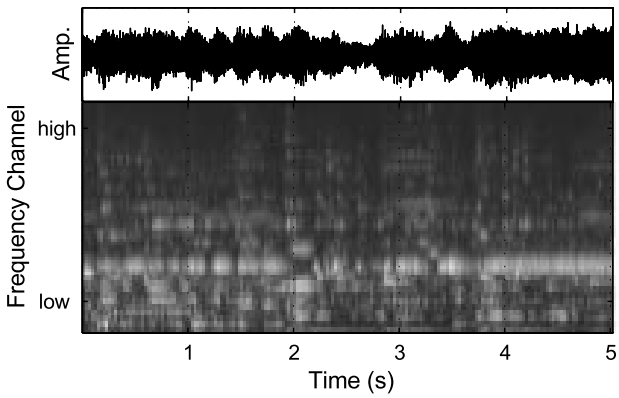


Figure 4-16**:** The input waveform and inner-hair-cell-model output for a classical-music track. Note that it is difficult to see the rhythm in the waveform or model output.

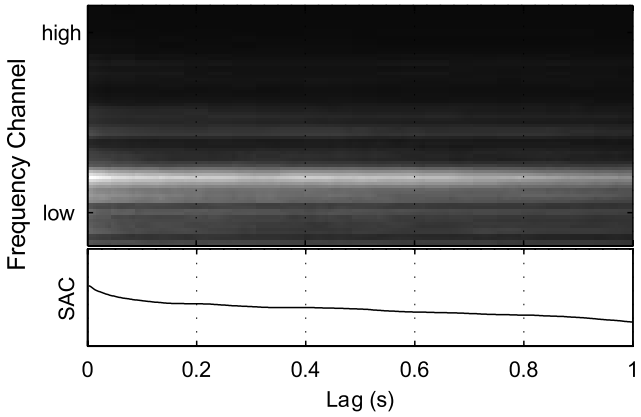


Figure 4-15**:** The correlogram and summary autocorrelation for the classical music sample. In this case, the tempo periodogram does not have clear peaks.

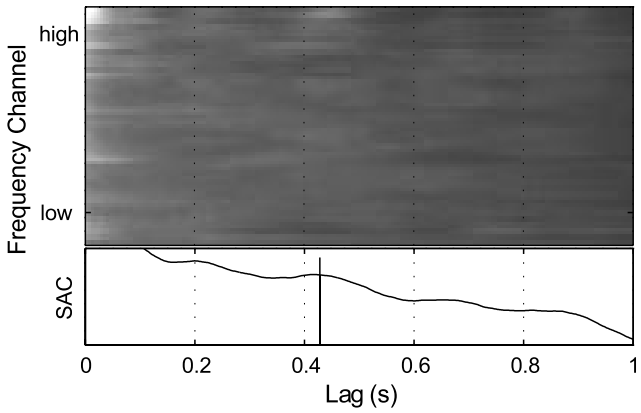


Figure 4-14: The normalized correlogram and summary autocorrelation. Now peaks are visible in the tempo periodogram. The hand-estimated tempo (see text) is marked with a line.

The study of envelope fluctuations and periodicity processing in the auditory system is still in its infancy (an excellent review of the neurophysiological data was provided by Langner (1992)). There is little concrete evidence for or against the idea that pitch and tempo perception are manifestations of a single periodicity-detection mechanism. But there is certainly an attractiveness about this possibility. If nothing else, it hints at an auditory brain that is unified around a single processing approach—namely, the discovery of coherent periodic envelope fluctuations in the multiple frequency subbands extracted by cochlear processing. As I will show in the next chapter, this principle also makes an excellent starting point for a theory of auditory-image formation in complex scenes.

4.6. Chapter summary

In this chapter, I have described an algorithm that can successfully beat-track digital audio representing music of many different types. The music does not have to contain drums or any other specific timbres, and it does not have to conform to any pre-determined set of musical templates. The beat-tracking procedure can be run in real-time on an advanced desktop workstation. This model of beat perception does not require that sound be "parsed" or otherwise converted into symbolic representations before being used to form a tempo percept; rather, the beat-track is produced via a continuous transformation from input to output.

I have demonstrated, both through qualitative evaluation on a wide range of ecological test examples, and via a formal (although small) psychoacoustic experiment, that the model performance is similar to that of human listeners. Finally, I have discussed the relationship between perceptual models of pitch and tempo, and shown that existing models of pitch perception are sufficient, if run very slowly, to explain the perception of tempo as well.

There are still aspects of the algorithm that are inadequately tested and understood. For example, would it be equally accurate but more efficient with a different filterbank, or could it be made more accurate in this way? What would be the implications of using a different temporal integration function, with different or more psychoacoustically accurate properties?

Errors made by the algorithm are typically due to the inability to understand beat relationships at various tempi; that is, a human listener intuitively understands the way eighth-note patterns group to form quarter-note and half-note patterns, and while some processing of this sort is done implicitly in the resonators due to phase-locking at harmonic ratios, it would clearly make the algorithm more robust to have an explicit model of this sort of rhythmic grouping.

Perhaps the way to build a system that can track complicated beat patterns is to construct it in two layers. The lower layer would be a simple perceptual beat extraction system as described here, which finds the level at which the pulse is evenly divided in time. Then, a higher-level grouping model selects and processes the beats to form an model of the rhythmic hierarchy present in the signal, based on pattern-recognition detection of accent structures and instrumental beat patterns. Building a system in this manner would allow us to leverage much of the existing work in cognitive rhythm models to apply to the analysis of digital audio as well as symbolically represented music.

In Chapter 6, I will revisit this model and present some simple continuous feature detectors based upon its outputs. These features will be related to high-level human semantic judgments about music in Chapter 7. But before that, I will continue the exploration of periodicity processing of subband representations, and show how such an approach may be used to explain the formation of auditory images in complex sound scenes.

CHAPTER 5 MUSICAL SCENE ANALYSIS

In this chapter, I present a new computational model for the processing of complex sound scenes by the auditory system. The model is based on a new principle of sound processing—dynamic detection of subband modulation within the autocorrelogram representation. It draws heavily on subband-periodicity models of auditory processing, and so readers with little previous exposure to this topic are encouraged to review Chapter 2, Section 2.1.1.

I will begin with an exploration of some under-appreciated aspects of the within-channel behavior of cochlear and subband-periodicity models. Then, I will present the new processing model. Following this, I will evaluate its performance when applied to several psychoacoustic test stimuli that are believed to reveal important phenomena of auditory grouping. A discussion of representational and processing aspects of the model in comparison with others in the literature concludes the chapter. Chapter 6 will discuss the application of this model, and the one I presented in Chapter 4, to the analysis of complex, real-world musical scenes.

5.1. The dynamics of subband periodicity

The basic subband-periodicity model was discussed in Chapter 2, Section 2.1.1 and was illustrated in schematic in Chapter 2, Figure 2-1. This model, and others like it, has mostly been applied in the past only to the analysis of static signals such as stimuli for psychoacoustic pitch tests and double-vowel experiments. The model has proven to be an excellent description of human pitch-analysis behavior. However, there has been little work on the application of this model to nonstationary stimuli—stimuli that change over time.

When a changing acoustic stimulus is processed by a subband-periodicity model, there are certain dynamic aspects of the autocorrelogram representation that are not often fully appreciated. These will form the basis of the processing model in Section 5.2, and so they are explained in more depth here. I wish to emphasize that the principles of this model are applicable to any of the variant models of subband periodicity discussed in Section 2.1.1. It operates on the changes in periodicity in the representation, not on the periodicity representation itself, and so any model that can dynamically represent changing periodicity in

the subband filter outputs could be used as a front-end instead. I use the autocorrelogram for simplicity (it is easier to analyze mathematically), and to build on previous work by Ellis (1996a) and Martin (1999) at the Media Laboratory (robust, efficient source code was already available). Models such as Patterson's Auditory Image model (1992) or De Cheveigne's subtraction model (1993) could equally well be used instead.

As a sound signal changes over time, the response of each channel of the autocorrelogram modulates in two ways. First, it undergoes *amplitude modulation* in response to changing level in the frequency subband of the signal associated with that channel. As the level in a subband increases, the energy output of the corresponding cochlear filter increases. Second, it undergoes *period modulation* in response to changes in the frequency of stimulation dominating that cochlear channel. In the rest of this section, I will explain what these terms mean, and the implications of this behavior for the construction of sound-processing systems.

The bandpass filters that comprise the cochlear filterbank are narrowly tuned. They can only produce output that is similar to a modulated sinusoid. In the output of each subband, the frequency of the sinusoidal carrier is near the center frequency of the filter and the modulation function is low-pass relative to the carrier. In the frequency region where phase-locking to the waveform occurs (all frequencies below about 2000 Hz), the half-wave rectification and smoothing processes shown in Figure 2-1 change the shape of the wave function, but not the underlying periodicity. Thus, for any input signal, the autocorrelation function in each of the low-to-mid subbands must be quasi-periodic. The period of the autocorrelation function in each of these channels must be near the period that corresponds to the center frequency of the subband.

As a time-varying signal evolves, the particular frequency dominating the filter channel—that is, acting as the carrier signal—changes. This change is reflected in a corresponding change in the periodicity of the autocorrelation function. The changes in the autocorrelation function is a type of modulation that I term the *period modulation* of that channel. A simple example is presented in Figure 5-1.

Period modulation of the autocorrelation signal in a subband does not only occur in response to frequency modulation of the input signal. It occurs any time that the frequency that is dominating a subband changes. This can happen due to amplitude modulation of the spectral components of the input as well as their frequency modulation. Amplitude modulation and period modulation are not independent; as the frequency dominating a channel changes, it moves closer to and farther from the center frequency of the channel, and thus leads to changing output response from the filter. Similarly, in most sound signals, multiple harmonic components are mixed together. As the relative levels of mixed components change, the one that dominates a particular channel may change. This leads to a *period* modulation in this channel as the reflection of the *amplitude* changes in the mixture.

Frequency modulation, both within the cochleagram and the autocorrelogram, is easiest to observe as a cross-band behavior. That is, in a frequency glide (Figure 5-2), the gliding sound can be observed as a diagonal line in the cochleagram (which naturally makes the observer think of frequency "components"), and as a horizontal stripe moving upwards and leftwards in the autocorrelogram when it is visualized as a moving image. In both of these cases, the observed elements of the scene move from one cochlear band to another over time. However, this visual effect does not necessarily have anything to do with the auditory *perception* of the sound—which is the goal of auditory modeling. It is essential that the visual appearance of a representation not be confused with the properties of the representation that pertain to its auditory properties.

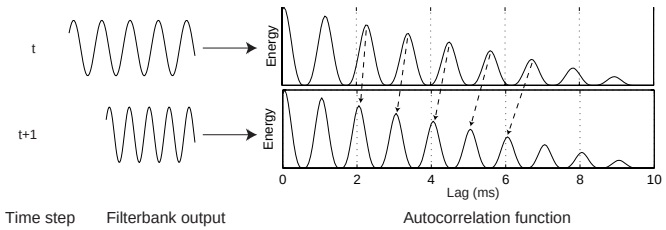


Figure 5-1**:** Autocorrelation of a windowed, modulated sinusoid. As the frequency of output from a particular filter channel changes, the period of the autocorrelation function of that output changes. The change in dominant frequency becomes a *period modulation* in the autocorrelation. The arrows indicate the motion of corresponding peaks in the autocorrelation from the first time frame to the second.

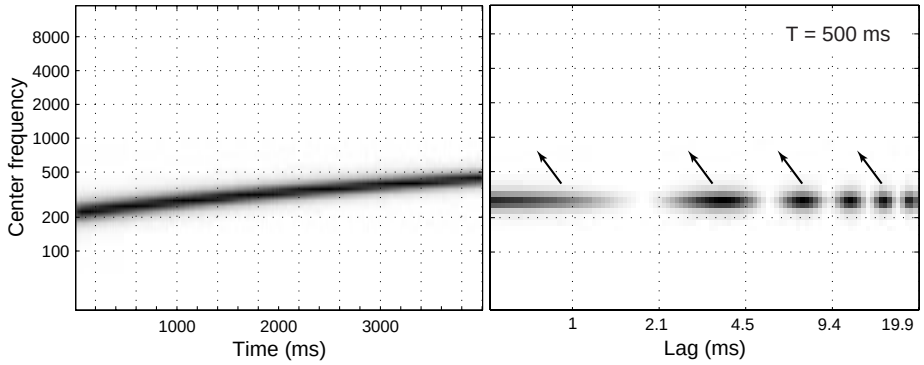


Figure 5-2: A frequency glide, as seen when it is processed in the cochleagram (left) and autocorrelogram (right) representations. The input stimulus is a sinusoid with instantaneous frequency starting at 220 Hz and moving to 440 Hz in 4 s. In the cochleagram, the glide appears as a diagonal line, oriented with positive slope. In the snapshot of the moving autocorrelogram, the glide appears as an upward and leftward shift of the periodicity peaks (the arrows indicate the direction of motion). Both of these visual motions are distracting and draw attention away from the within-band dynamics of the signal.

The viewpoint that I will present in Sections 5.2 and 5.3 is that, for the purposes of auditory grouping, the gliding percept is best understood as an example of coherent *within-band* behavior. As shown in Figure 5-3, at the beginning of the glide (Figure 5-3a), a cochlear channel centered on 220 Hz is responding strongly. In that channel, there is negative period modulation and downward amplitude modulation as the glide increases in frequency and moves away from the channel's center frequency. At the same time, the channel centered on 315 Hz is exhibiting upward amplitude modulation as well as negative period modulation. As the stimulus evolves, both the pattern of stimulation (the set of filters that are responding strongly) and the within-band dynamics in each channel change. Near the end of the glide stimulus (Figure 5-3c), the channels in all three regions negative modulation of both sorts.

I claim that *it is the negative period modulation in the three frequency bands that unifies these bands as part of the same auditory image*. The cross-band properties only affect the perceived qualities of the resulting image—the spectral center of the modulation shifts as the region of maximum stimulation moves higher.

The comparison that I am making between cross-channel and within-channel representations may seem unnecessarily elaborate at this point, but understanding this behavior is essential to understanding the model of dynamics presented in the next section. Part of my goal in this chapter is to demonstrate that a robust theory of auditory grouping can be based solely upon

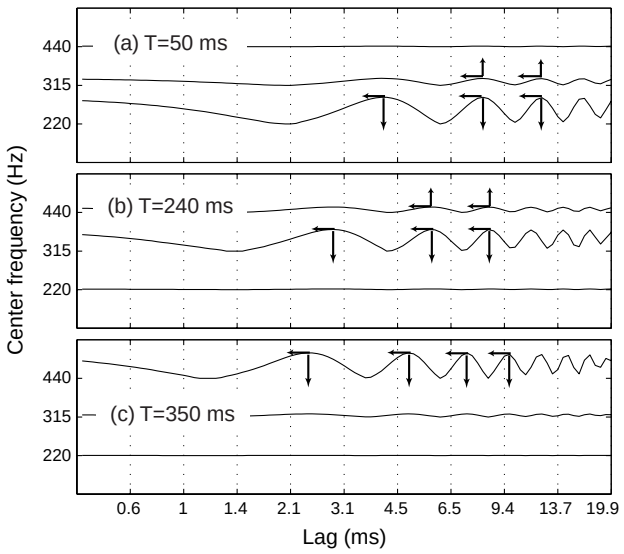


Figure 5-3: Three snapshots in time of the within-band autocorrelations of the same glide stimulus as in Figure 5-2, for three different cochlear channels. In this view, the within-band dynamics of the stimulus are more apparent; the superimposed arrows indicate the directions of motion over time. In (a), at the beginning of the stimulus, the lowest frequency channels are exhibiting negative amplitude and period modulation—that is, the amplitude is getting bigger and the period smaller, as shown by the arrows—while the middle frequency band is growing in amplitude. In (b), in the middle of the stimulus, the middle frequency band shows the strongest response; the middle band is shrinking and the high frequencies are growing in amplitude, while all three bands show common negative period modulation. By the time of panel (c), only three frequencies have positive amplitude modulation and continue to show negative period modulation. All channels exhibit negative period modulation at all times.

within-channel dynamics, where no components or other mid-level entities need to be estimated across channels.

5.2. Processing model

In this section, I will describe a computational model that is capable of allocating channels from a cochlear model into a partitioned representation suitable for further analysis. As such, it is only one constituent of a complete model of the perception of complex auditory scenes. I have not developed any innovations in the cochlear model or the periodicity analysis; therefore, I will only discuss these stages of the approach briefly. Further, although I believe incorporation of top-down information is necessary in order to build robust CASA systems, as discussed in Chapter 3, Section 3.4.1, it is not explicitly treated here.

A remarkable correspondence was reported some time ago (Duda *et al.*, 1990) between the perception of auditory scene analysis and the appearance of the temporal motion that is observed when the autocorrelogram is visualized. However, as yet there has been relatively little attempt to operationalize this discovery in a computational auditory scene analysis (CASA) system. The principle underlying the operation of the system I present is the same one articulated by Duda *et al*: parts of the sound scene that belong together (as part of the same auditory image) can be seen to undergo coherent modulation when the correlogram is visualized as a moving picture.

Slaney and Lyon (1991) produced an excellent "hearing demo reel" videotape that effectively illustrates this principle for many sorts of sounds, including multiple talkers, speech-in-noise,

and symphonic music. Their demonstrations suggest that the scene-partitioning problem might be solved by estimating modulation in the subbands of the autocorrelogram and using that information to group the cochlear channels.

The correlogram-comodulation analysis system (Figure 5-4) is divided into five rough stages that I will describe more fully in the subsequent sections. They are: (1) frequency analysis, rectification, and smoothing of sound through models of the cochlea and inner hair cells; (2) subband periodicity analysis with the autocorrelogram; (3) subband modulation detection; (4) clustering of the modulation data to discover comodulation patterns and to determine how many auditory images are present; and (5) assignment of each channel, over time, to the various auditory images. The last three steps make up the new approach, therefore my description is most detailed there.

5.2.1. Frequency analysis and hair-cell modeling

The front-end system that I use is very similar to others reported in the literature. This particular implementation was programmed by Martin (1999), following the work of Slaney (1994) and Ellis (1996a). The cochlear filterbank is modeled as a set of N=54 eighth-order *gammatone* filters; this model of the cochlea was introduced by Patterson *et al.* (1992). The phase-locking behavior of the inner hair cells in the cochlea is modeled with half-wave rectification and smoothing. The output of this stage of processing for a simple test sound was shown in Chapter 2, Figure 2-2.

In signal-processing terms, the input signal $x(t)$ is passed through each cochlear filter $\Gamma_i(t)$, $1 \le i \le N$, to produce an array of output signals

$$y_i(t) = x(t) * \Gamma_i(t) \qquad (5\text{-}1)$$

Martin (1999, pp. 70-75) presented more details about the exact properties of this filterbank.

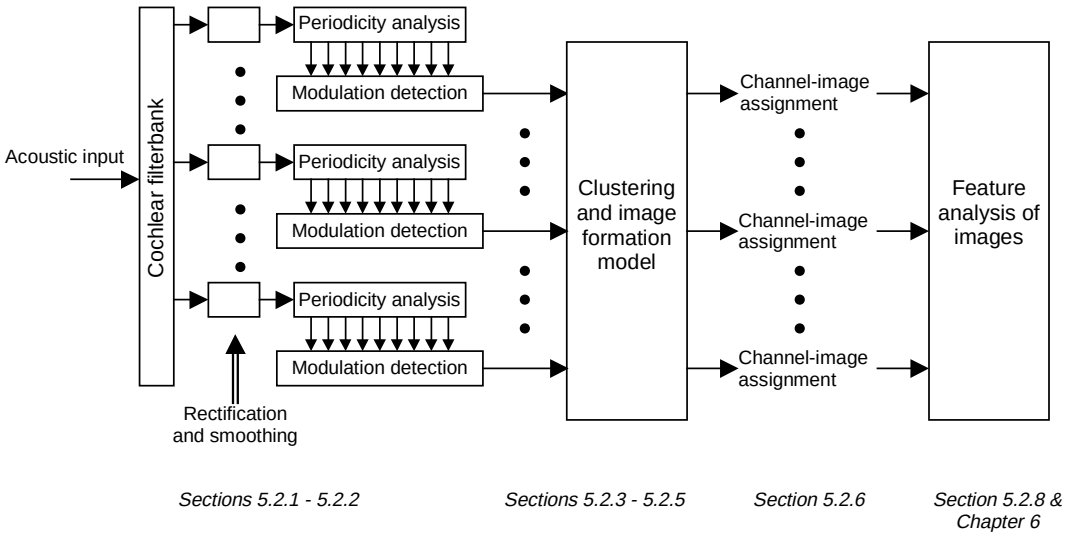


Figure 5-4: Overall schematic of the sound-analysis model presented here. Note the similarity of the initial stages to Figure 2-1 in Chapter 2. After filterbank decomposition and subband periodicity detection, the subband modulation is detected channel-by-channel and the result is passed to a clustering model, which determines the number of auditory images in the scene and the modulation properties of each. The output of the clustering model is an assignment of each channel to an auditory image at each time; the resulting channel-image assignments provide evidence for the analysis of auditory features of the images.

Each $y_i(t)$ is rectified and smoothed by convolution with a 0.25 ms raised-cosine window $W(t)$; this results in a set of cochlear channel signals

$$z_i(t) = \Re[y_i(t)] * W(t) \tag{5-2}$$

where $\Re[\cdot]$ is the half-wave rectification operator,

$$\Re[x(t)] = \begin{cases} x(t) & \text{if } x(t) > 0 \\ 0 & \text{if } x(t) \le 0 \end{cases} \tag{5-3}$$

Periodicity detection is performed using the running log-lag autocorrelogram. In earlier implementations of the autocorrelogram (Slaney, 1994), it was calculated on a frame-by-frame basis, by windowing the half-wave-rectified output signals of the cochlear filterbank and using the FFT to compute the autocorrelation. More recently, Ellis (1996a) and Martin (1999) have suggested using a running autocorrelation rather than a windowing operation. This eliminates edge effects (the decay shown in Figure 5-1) caused by windowing the signal before calculation, and it also makes it easier to sample the lag axis in different ways.

Ellis (1996a) observed that as human pitch perception is roughly linear with logarithmic frequency, when the autocorrelogram is used as a pitch model, it makes more sense to sample the lag axis with logarithmic spacing. He termed this the *log-lag autocorrelogram*. In Martin's implementation of the log-lag autocorrelogram, delay lines are used to calculate the continuous running autocorrelation without an analysis window (see Figure 23, p. 79, of Martin, 1999). The delay line outputs are computed using fractional-delay filters, and after multiplication with the undelayed signal, each lag signal is smoothed with a lowpass filter. This model is more computationally intensive than Slaney's FFT-based model, but has properties convenient to the detection of modulation, as will become clear in the next section. Three frames of the autocorrelogram of the synthetic test sound were shown in Chapter 2, Figure 2-3. For the analysis presented here, the autocorrelogram is sampled at a frame rate of 100 Hz.

For all time t and lag τ, the smoothed autocorrelation of the i-th cochlear channel is defined as

$$R_{zz}^{i}(t,\tau) = \int_{s=-\infty}^{\infty} w^2(s-t) z_i(t-s) z_i(t-\tau-s) ds \tag{5-4}$$

which is equivalent to

$$R_i(t,\tau) = [z_i(t) z_i(t-\tau)] * w^2(-t) \tag{5-5}$$

that is, delay, multiplication, and smoothing.

To form the correlogram *frame* at time t, the autocorrelation is sampled at $S_l = 150$ lags logarithmically spaced between 0.5 and 20 ms for each of the cochlear filters. That is, the correlogram frame at time t is a matrix $\mathbf{F}^t$, defined as

$$\mathbf{F}_{ij}^{t} = R_i(t,\tau_j),\ 1 \le i \le N,\ 1 \le j \le S_l \tag{5-6}$$

5.2.2. Modulation analysis

In this section, I present new techniques for analyzing the modulation behavior of channels of the autocorrelogram. The purpose of modulation analysis is to convert the dynamic motion of the autocorrelogram into static features that are suitable for inclusion in a pattern-analysis system.

On a linear lag axis, a simple frequency modulation such as vibrato corresponds to a period modulation that can be described as *stretching* and *squashing*. That is, when there is an increase in the frequency of the sound component that is stimulating a particular filter

channel, the output of the filter also increases in frequency. As a result, the x-axis of the autocorrelation function is compressed (squashed), with peaks closer together, as in Figure 5-1. As the signal frequency decreases, the output of the filter decreases in frequency and the spacing of the peaks of the autocorrelation function is stretched.

The utility of the log-lag autocorrelogram for detecting period modulation now becomes evident. When the lag axis is scaled logarithmically, the stretch-squash effect of period modulation becomes a simple shift to the right or left, which is easy to analyze. Cross-sections of the log-lag autocorrelogram for two cochlear channels (one steady, and one undergoing period modulation) for the synthetic sound used in the examples of the correlogram pitch model in Chapter 2 are shown in Figure 5-5. The period modulation is easily visualized in the cross-section of the modulating sound as parallel curves over time.

When two harmonics, or harmonic energy and noisy energy, collide in a single filterband, then the dynamics are more complicated. I have not conducted a full theoretical analysis of such situations; however, a few comments are pertinent. First, in many cases the autocorrelation function behaves as an either/or indicator of in-band signal strength. That is, when two harmonics collide, if one is stronger than the other, the period of the stronger harmonic will tend to dominate the autocorrelation function. Similarity, at reasonable signal-to-noise ratios, a tonal component embedded in a noisy background still gives rise to clear peaks in the autocorrelation. More importantly, since the basic processing step here is to detect the *dynamics* of the in-band behavior, the actual values of the periods and amplitudes in each band are not directly relevant. In complex scenes, these values will be sometimes changing in a smooth way for a particular channel, and sometimes in a discontinuous way. It is the pattern-recognition part of the model, presented in the next section, that sorts out the different ways that filter channels are changing.

In the log-lag domain, cross-correlation can be used to detect period modulation. At each time step, the autocorrelation function in each channel is cross-correlated with the autocorrelation function in the same channel from a previous time step. If the channel is undergoing period modulation, the peak of this cross-correlation function is off-center. I have found that peak-picking suffices to determine the period modulation, at least for the simple

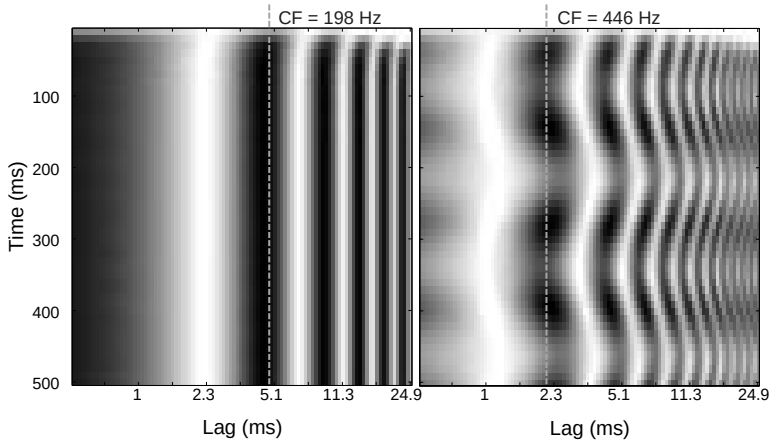


Figure 5-5**:** Two slices through the autocorrelogram of the McAdams oboe sound (which was introduced in Chapter 2, Section 2.1.1, Figure 2-2 and Figure 2-3). Each panel shows the autocorrelation response of a single filter changing over time. The left panel corresponds to a cochlear filter with center frequency 198 Hz; thus, this channel is dominated by the steady partial at 220 Hz in the sound. The right panel shows the response of a cochlear filter with center frequency 446 Hz; thus, this channel is dominated by the partial that frequency-modulates about 440 Hz. Since the lag axis is logarithmically scaled, the period modulation is reflected as linear shifting behavior over time, not stretching and squashing—all of the curves in the right panel are parallel.

examples in this chapter.

The finite length of the autocorrelation vector (that is, as discussed in the previous section, only the autocorrelation lags from 0.5 ms to 20 ms are used in analysis) provides an implicit windowing function in the cross-correlation. The running autocorrelation is only calculated over a finite set of lags, and so it can be viewed as the application of a rectangular (boxcar) window function to the true, infinitely-long, autocorrelation function. Since this windowing function has a triangular autocorrelation function, it biases the peak estimate in the cross-correlation towards the center. In order to provide accurate estimates, the cross-correlation is unbiased by multiplication with the inverse triangle window before peak-picking.

Given the definition of the $\mathbf{F}^t$ matrix as in (5-6), the column vectors

$$\mathbf{r}_{ti} = \left[\mathbf{F}^t_{i1} \quad \mathbf{F}^t_{i2} \quad \cdots \quad \mathbf{F}^t_{iS_l}\right]^T \tag{5-7}$$

hold the autocorrelation in each channel i at a particular time t. Each of these may be compared to the **r** vector in the same channel i at the previous time t-ΔT to calculate the period modulation. Period modulation is computed by finding the maximal point of the cross-correlation between these two vectors.

The lag axis of the autocorrelogram, and thus each **r** vector, is sampled discretely at a fairly coarse resolution. This introduces no artifacts since, as I discussed above, the output of each subband filter is narrowband. However, it is possible to gain more resolution in the peak estimates if the **r** vectors are upsampled before cross-correlation.

The cross-correlation of **r** in channel i at time t is defined as

$$K^{t,i}_{rr}(\lambda) = \sum_{l=1}^{QS_l} \hat{\mathbf{r}}_{t\,i\,l}\hat{\mathbf{r}}_{t-\Delta T,\,i,\,l-\lambda} \quad \text{for} -S_l < \lambda < S_l \tag{5-8}$$

where Q=10 is the upsampling factor, and $\hat{\mathbf{r}}$ denotes the **r** vectors after upsampling by Q. The summation is taken with appropriate zero-padding on the ends of the $\hat{\mathbf{r}}$ vectors. The maximum of this function after application of the unbiasing window $W(\lambda)$ is then

$$p_i(t) = \sup_{\lambda} K^{t,i}_{rr}(\lambda) W(\lambda) \tag{5-9}$$

This is the value termed the *period modulation* of channel i at time t.

The domain of period-modulation values is *lag scale per second*. At each time step, the autocorrelation function in each channel is scaled by some ratio that may be detected by looking for shifts in the log-lag autocorrelogram as just described. The primary region of interest in this domain is –2% to +2% lag scale per 10 ms frame. Very small modulations, between –0.2% and +0.2% per frame, are difficult to detect due to the lack of high-frequency information in the autocorrelation function.

I observe that, since the entire row of the autocorrelogram is used to estimate the period modulation, the resulting estimates are quite robust to noise. This would not necessarily be the case if, for example, I tried to determine the highest peak in each row, and then track the motions of the peaks over time.

The output of the period-modulation detection is shown in Figure 5-6. This figure, a *period modulogram*, shows the two-dimensional function mapping cochlear channel and time into the period-modulation estimation in that channel at that time.

The amplitude modulation in each channel is also measured. This is accomplished by dividing the zero-delay autocorrelation—the power—in each channel by the zero-delay autocorrelation from a previous frame, half-wave rectifying the result (so that decreases in amplitude are not salient) and applying a compressive nonlinearity (to boost the effect of

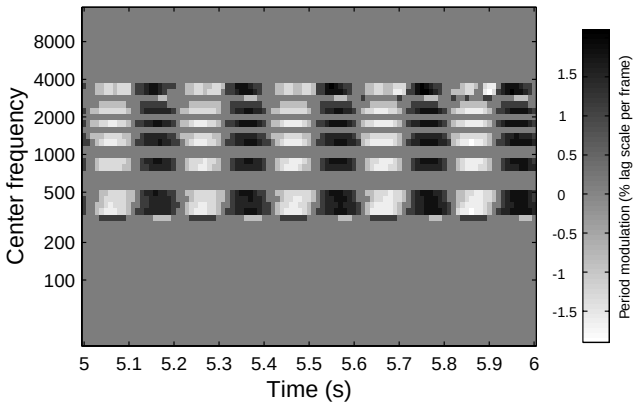


Figure 5-6: The period modulogram of the McAdams oboe, showing the period modulation of each channel at each time step. The period modulation is measured with cross-correlation as described in the text; the legend at right shows the correspondence between the gray level in the main panel and the degree of period modulation. Period modulation is measured in lag scale per time; a value of 1.5 % means that the lag axis in that channel at that time must be stretched by a factor of 1.5 % (that is, a multiplicative scaling of 1.015) in order to approximate the lag axis in the same channel at the next time step. Similarly, a value of -1.5 % corresponds to a multiplicative squashing factor of 0.985. Compared to Chapter 2, Figure 2-3, the channels responding to the vibrato are clearly visible.

small amplitude modulations relative to large ones). The power is just (5-4) evaluated at τ=0, which reduces to

$$P_i(t) = z_i^2(t) * w^2(-t) \tag{5-10}$$

The amplitude modulation of channel *i* at time *t* is then defined as

$$a_i(t) = \sqrt{\Re\left[\log \frac{P_i(t)}{P_i(t-\Delta T)}\right]} \tag{5-11}$$

with $\Re$ defined as in (5-3). The particular compressive nonlinearity used is the square-root function shown in (5-11).

The amplitude modulation is measured in *energy scale per frame*; it is greater than 0 dB when the channel is increasing in power, and less than 0 dB when the channel is decreasing in power. The output of the amplitude-modulation detection process—the *amplitude modulogram*—is shown in Figure 5-7. There is no explicit amplitude modulation in this signal after the onset, so all of the amplitude modulation arises from coupling to frequency modulation.

In the examples to be evaluated in Section 5.4, the time-delay for modulation analysis has been set to ΔT=40 ms. This is a value that I have found to give good empirical results on the evaluation tests. I believe that more information could be retrieved from the modulation patterns by using more than one previous frame—for example, by cross-correlating the autocorrelation function at time *t* in a channel with that of the same channel at 1 ms, 10 ms, and 100 ms previous. With appropriate smoothing, this sort of *multiscale* processing could give information on coherent motion at many time resolutions, from glottal jitter to syllabic or note-to-note transitions. I will not consider multiscale processing at all in the present model. It is left as a topic for future work.

Estimating the amplitude and period modulation of cochlear channels with very little sound energy is difficult since the signal-to-noise ratio may be low. The values are nonsensical if there is no energy at all in a particular channel. In the absence of a more principled approach,

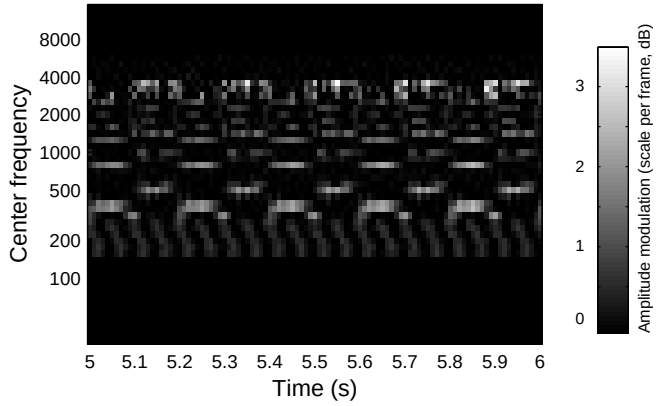


Figure 5-7: The amplitude modulogram, showing the amplitude modulation of each channel at each time step. Amplitude modulation is described as a scaling factor per frame; if the value is 2 dB, then the energy in that channel at that time step is 2 dB greater than it was at the previous time step. When the figure is compared to Figure 5-6, the correlation between period modulation and amplitude modulation is observed.

I have simply discarded channels with power –30dB compared the channel with the most power; they are not considered in the clustering and grouping stages of this model. (A more principled approach would be, at minimum, to consider loudness rather than power, and if possible to understand better any low-loudness circumstances that affect the segregation of the scene into auditory images. Regardless, within my present approach, where the overall goal is to analyze the features of the sound scene, the low-power channels are not important since they cannot affect the features very much unless entire auditory images are very quiet.)

5.2.3. Dynamic clustering analysis: goals

The modulation analysis described in the previous section converts the dynamic motion of the correlogram into static features, and the cross-channel concept of frequency modulation into the within-channel concept of period modulation. The next step in the comodulation analysis is to find groups of channels that are modulating in the same way. In this section, I will describe a clustering model that suggests one way to do this. Naturally, there are many other techniques that could also be examined. I will first describe the dynamic behavior of typical modulation patterns, to make clear the various behaviors that must be addressed, and then I will present one particular clustering model that can account for them.

There is an extensive literature on clustering and grouping feature data through pattern-recognition methods (Duda and Hart, 1973; Therrien, 1989; Bishop, 1995). By couching auditory-scene-analysis problems in a suitable analytic framework, techniques from this literature may be brought to bear, just as they are for problems in visual scene analysis.

In the present model, the feature space has only two dimensions; each channel at each time is mapped to an vector $\mathbf{x}_i(t) = [p_i(t)\ a_i(t)]^T$, where $p_i(t)$ and $a_i(t)$ are defined as in (5-9) and (5-11) respectively, as the current period and amplitude modulation. Within each frame, only those channels containing significant acoustic power (at least –30 dB compared to the channel with maximum power, see Chapter 2, Figure 2-3) are considered. Thus, at each time step there are at most 54 ordered pairs (one for each cochlear channel) but usually less, since not every channel has energy at every time. Figure 5-8 shows a scatterplot of the feature space at four points in time for a test sound.

The hypothesis on which this auditory grouping model is founded is that the perception of auditory images is due to coherence of multiple filter channels within this feature space. That is, whenever multiple points (meaning the modulation behaviors of filter channels) are nearby

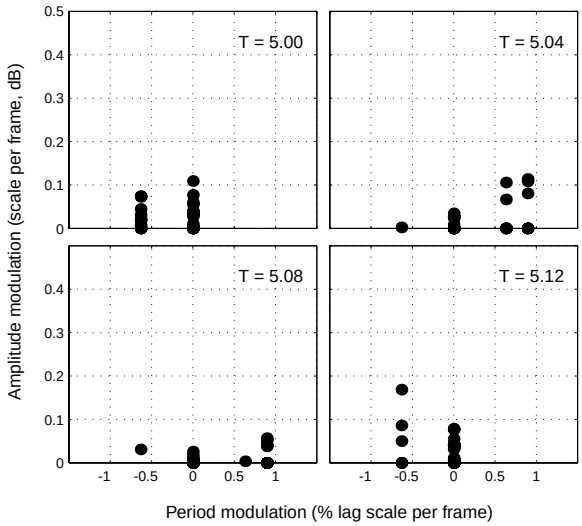


Figure 5-8: Scatterplots of the modulation data for the test sound at four different time steps. Each point corresponds to the behavior of one cochlear channel at one point in time. The period and amplitude modulation values for this sound were shown independently on separate graphs in Figure 5-6 and Figure 5-7.

each other in this space for a significant period of time, the channels are perceived as grouped together into a single auditory image.

This is the same hypothesis as the one suggested visually by Duda *et al*. (1990); namely, that the perception of the sound scene being divided into multiple auditory images is due to the coherent comodulation (in the autocorrelogram domain) of each image, and independent modulations of the different images.

The pattern-recognition problem that needs to be solved is as follows. In a certain time frame *t*, the 54 cochlear channels are arrayed as a distribution of points in the two-dimensional modulation feature space, as shown in Figure 5-8. We wish to find a set of *clusters* of points that accurately models (or "explains") this distribution. In the next time frame *t*+1, the distribution of points changes. We wish to find another set of clusters that models the new distribution; however, we do not want the cluster model at time *t* to be independent of that at time *t*+1. Rather, when possible we wish to explain the data as arising from clusters moving from one location in feature space from another, while keeping the groups of points assigned to each cluster as consistent as possible. As I will show, the fact that the frames of data are not independent and nonstationary leads to interesting analytic constraints on the clustering behavior.

Figure 5-9 shows a schematic of typical clustering behavior. It will be used as a reference for the discussion of the desired properties of the model. In this figure, some of the points have been labeled with their cochlear-filter-channel number for clarity. These labels emphasize the changing modulation of the filter channels over time.

The first panel ("T=1") of Figure 5-9 shows a configuration of points (cochlear channels) that is relatively simple to analyze. The data fall readily into two clusters, labeled A and B. Cluster A consists of several channels, including #1, #4, and #8, that are undergoing a large amplitude modulation and a slightly negative period modulation. Cluster B consists of several channels, including #2, #6, #5, #7, and #3, that are undergoing a small amplitude modulation and are relatively stable, on average, in period modulation.

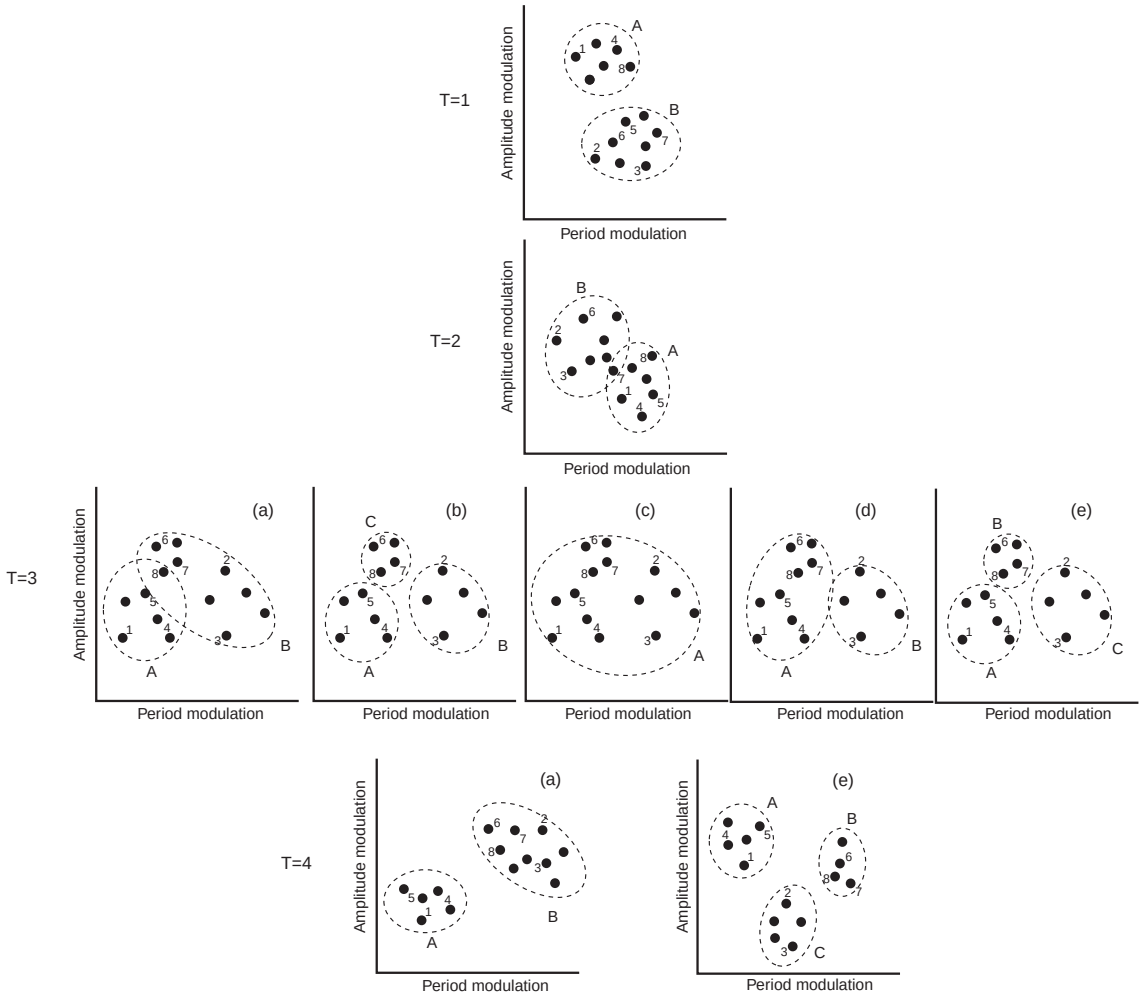


Figure 5-9: Schematic plot of the cluster analysis process. Each frame shows the configuration of cochlear channels in modulation space at a single point in time; each point corresponds to the instantaneous amplitude and frequency modulation in one cochlear channel. At T=1, there are clearly two auditory images (clusters of points) in the scene, as the points fall into two separable loci in the modulation space. At T=2, there are still two images hypothesized, but the clusters are closer to each other and it is ambiguous to which images several channels (for example, channel #7) should be assigned. At T=3, the situation is more difficult to resolve. Five different possibilities, given the same configuration of points, are shown; there is not sufficient evidence at a single point in time to choose one as clearly preferable. However, the data from a future point in time may help to disambiguate. For example, two possible configurations of cochlear-channel data are shown for T=4. If T=4(a) is the situation that is observed, then T=3(a) is preferable (since it preserves the maximum stability in each auditory image from one time step to the next). Alternatively, if T=4(e) is observed, then T=3(e) is preferable. As described in the text, the EM algorithm is used to hypothesize cluster arrangements in each time step, and the Viterbi algorithm is used to evaluate the temporal sequence of cluster arrangements over time and choose the most likely explanation for the observed data.

The second panel ("T=2") shows the next frame in time. Each point has moved to a new location in the feature space; that is, each filter channel is modulating in a different way compared to the way it was modulating in the first frame. As a result, the location and composition of the clusters changes. Channel #5 has switched and now belongs to Cluster A; it is indeterminate to which cluster channel #7 belongs. Notice that the two clusters have nearly switched places; this is a better description of the underlying data than an explanation in which the clusters stay in the same place, but all the channels switch clusters. In general,

explanations are preferred that keep as many channels in the same cluster as possible from one time step to the next.

For the third time step, a more ambiguous situation is presented. The five panels presented for this time step ("T=3 (a)-(e)") show several clustering explanations for the data at this moment. All five panels show the same data configuration; only the suggested clusters differ. Panels (a) and (d) show two different clustering models that each explain the data as arising from two clusters. In panel (a), Cluster B is large and awkward; while in (d) the clusters have a more coherent center, but require that channel #6 changes clusters. Panels (b) and (e) explain the data as three clusters, thus hypothesizing the emergence of a new auditory image represented by Cluster C. These two panels differ in the way they suggest the continuity of the previously existing images; in (b), channel #8 moves from Cluster A to the new Cluster C, while in (e), channel #8 switches to Cluster B. Thus, in (e), Cluster C is a pure subset of the channels in Cluster B at the previous time step (Cluster B has "split" into two clusters), while in (b), Cluster C is a combination of channels from the previous time step (Cluster C has "overtaken" channels from both A and B from the previous time step). Finally, in panel (c), the data are explained as arising from only one cluster; that is, the auditory images that were previously kept separate as A and B have merged together perceptually.

It is to be emphasized that there is no "correct" answer that can be chosen from these alternatives, at least if we only look at one instant in time. Even intuitively, the different cluster models for T=3 in Figure 5-9 all have different advantages and disadvantages. It may be the case that future data helps to disambiguate the choice of clustering. For example, two different frames for T=4 are shown in Figure 8; each of these represents a possible alternative for the continuing evolution of the scene. If the first ("T=4 (a)") is the continuation, then panel (a) would be preferred for T=3, since those channels that are tentatively grouped together in 3(a) are more definitely so grouped in 4(a). If the second—panel (e)—is actually the continuation at T=4, then panel (e) is preferred for T=3, for similar reasons.

Ultimately, of course, the desired clustering analysis is the one (or one of the ones) that reflects the human perception of the sound scene presented to the model. As will be discussed in Section 5.5.5, it is rather difficult to evaluate directly the moment-to-moment correctness of a proposed source-grouping theory, because for most complex sound scenes there is no experimental data available on the way or ways in which humans behave. However, Section 5.4 will demonstrate the performance of the model proposed here with regard to some well-known stimuli used for perceptual grouping experiments.

5.2.4. Dynamic clustering analysis: cluster model

In order to satisfy the desiderata outlined in the previous section, I have developed a three-stage clustering and grouping model. First, the Expectation-Maximization technique is used to estimate a Gaussian Mixture Model for the data in each time frame. Second, a Viterbi lattice technique is used to estimate the number and labeling of clusters in each frame. Finally, a second stage of Viterbi processing is used to determine the assignment of each cochlear channel to a cluster in each frame.

Gaussian Mixture Models (GMMs) are models of probabilistic data in which the probability of a random data point $\hat{\mathbf{x}}$ occupying a certain location in the feature space is given by a sum-of-Gaussians function. That is,

$$p_{\mathbf{x}}(\hat{\mathbf{x}}) = \sum_{i=1}^{G} \rho_i N(\hat{\mathbf{x}}; \boldsymbol{\mu}_i, \mathbf{K}_i) \tag{5-12}$$

where ρ_i are the prior probabilities of each of G clusters (thus $\Sigma\rho_i = 1$) and $N(\mathbf{x}; \mu, \mathbf{K})$ is the normal (Gaussian) distribution of the k-dimensional random vector $\mathbf{x}$ with mean μ and covariance matrix $\mathbf{K}$, that is

$$N(\hat{\mathbf{x}};\boldsymbol{\mu},\mathbf{K}) = \frac{1}{(2\pi)^{k/2}|\mathbf{K}|^{1/2}} \exp\left[-\frac{1}{2}(\hat{\mathbf{x}}-\boldsymbol{\mu})^T \mathbf{K}^{-1}(\hat{\mathbf{x}}-\boldsymbol{\mu})\right] \tag{5-13}$$

Given a set of observations $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ of the random variable $\hat{\mathbf{x}}$ and a certain number of target clusters G, we wish to find the most probable μ_i and $\mathbf{K}_i$ given the observations $\mathbf{X}$. Using the well-known Bayes' Theorem transformation (1995), we know that

$$p(\boldsymbol{\mu}_i, \mathbf{K}_i \mid \mathbf{X}) = \frac{p(\mathbf{X} \mid \boldsymbol{\mu}_i, \mathbf{K}_i)\, p(\boldsymbol{\mu}_i, \mathbf{K}_i)}{p(\mathbf{X})} \tag{5-14}$$

and so maximizing $p(\mu_i, \mathbf{K}_i \mid \mathbf{X})$ is equivalent to maximizing $p(\mathbf{X} \mid \mu_i, \mathbf{K}_i)$ if the prior probabilities $p(\mu_i, \mathbf{K}_i)$ and prior point distribution $p(\mathbf{X})$ are taken as uniform. The expression $p(\mathbf{X} \mid \mu_i, \mathbf{K}_i)$ (termed the *likelihood function*) is easier to maximize since there is a convenient expression, namely equation (5-12), for it.

Assuming the observations $\mathbf{X}$ are independent (not strictly true in this problem space), the final expression to be maximized is

$$\boldsymbol{\mu}_i; \mathbf{K}_i = \sup_{\tilde{\mu}_i; \tilde{\mathbf{K}}_i} \prod_{j=1}^{n} \sum_{i=1}^{G} \rho_i N\left(\mathbf{x}_j; \tilde{\mu}_i, \tilde{\mathbf{K}}_i\right) \tag{5-15}$$

There is no general analytic solution for this problem, but the Expectation-Maximum (EM) algorithm (Dempster *et al.*, 1977) provides an iterative method for finding locally-optimal parameters. A complete description of the functioning of the EM algorithm would take us too far afield from auditory matters; for my purposes, the EM algorithm is treated as a black-box technique for estimating the μ_i and $\mathbf{K}_i$ given the $\mathbf{x}_j$. In this model, the EM algorithm is initialized with random cluster parameters. It rapidly converges to a "locally optimal" solution in the parameter space.

As in many implementations of the EM algorithm, the determinants of the covariance matrices $\mathbf{K}_i$ are constrained by bounding the eigenvalues of each. If any of the eigenvalues gets too small during convergence, it is artificially set to be exactly the minimum, denoted λ_m. This is necessary because there are often pathological local optima in the parameter space where the covariance matrices of one or more clusters become singular.

The psychoacoustic problem at hand maps into this framework in the following way: the $\mathbf{x}_j$ are the observed two-dimensional feature vectors for each of the cochlear channels. Since a 54-channel cochlear filterbank is used, there at most 54 such points. The G clusters of points, by hypothesis, correspond to the auditory images. The means μ_i and covariance matrices $\mathbf{K}_i$ correspond to the location and shape of each of the clusters within the feature space. The ρ_i—the prior probabilities that any channel belongs to image i—are all fixed and equal to $1/G$ in the assumption that, *a priori*, there is no reason to prefer one of the clusters (images) over another. k in Eq. (5-13) is always equal to 2 since the feature space is always two-dimensional. The minimum length of an eigenvector λ_m can be interpreted as the perceptual resolution of modulation; clusters in modulation space cannot be distinguished more finely than this.

A well-known issue with models of this sort (see, for example, Duda and Hart, 1973, pp. 241-243) is that there is no principled way to determine how many clusters there should be. In the clustering framework that I have set up, the time-series (nonstationary) aspects of the clustering model provide a way to deal with this problem. The EM algorithm is executed numerous times with different settings of G. Each of these settings, for $G = 1, 2, \ldots G_{\max}$, results in a different configuration of clusters in feature space—for example, the choices shown in Figure 5-9 Time 3(a)-(c). The different configurations will be compared, and one selected, in the next step of the model, described in the next section.

The way in which the EM algorithm estimates the parameters of the Gaussian clusters depends on the distance function used. That is, given two points in the modulation space, the distance from one to the other may be measured in a variety of ways. In principle, it should be possible to use data from perceptual experiments to determine the human metric for distance in modulation space. To achieve the results demonstrated in Section 5.4, a simple spatial distance metric (the 2-norm $\| \mathbf{x}_i - \mathbf{x}_j \|$) was used.

5.2.5. Dynamic clustering analysis: time-series labeling

The first stage of the clustering analysis was described in the previous section. It gives, for each time frame, several hypotheses regarding the number and configuration of clusters in feature space at that time. The second stage of cluster analysis is to determine the optimal evolution of the cluster parameters over time. To illustrate, referring again to the schematic in Figure 5-9, the EM algorithm might provide the configuration shown for T=2, choices (a), (b), and (e) for T=3, and either configuration (a) or configuration (e) for T=4. Based on the configurations in T=2 and T=4, the second stage must select the optimum choice for T=3. The selection of the optimized sequence of cluster arrangement is a type of time-series analysis.

The time-series analysis required in this problem involves two aspects. First, the EM algorithm cannot give consistent labels to the clusters from one time step to the next. That is, even for an easy case such as T=1 in Figure 5-9, the EM algorithm is equally likely to produce two solutions. In the first, which is shown, class A contains channels #1, #4, and #8, with class B containing channels #2, #3, and #7 (among others). In the second, class B contains channels #1, #4, and #8, and class A contains channels #2, #3, and #7. The two solutions are identical except for the labeling of the classes, which is reversed. In general, if there are more than two classes, the labels are arbitrarily permuted.

In a single time frame, the labels are unimportant since there is no need to have a particular image bear a particular label. But in a sequence of time frames, it is important that the labels from one time step to the next agree with each other—this is known as a *correspondence problem*. One way to solve it might be to use the cluster arrangement at one time step as the *a priori* most likely arrangement at the next; this sort of heuristic may be implemented in the EM framework by using a non-uniform function for $p(\mu_i, \mathbf{K}_i)$. This approach is not taken here, because a different method allows the labeling of the classes to be corrected in conjunction with the second aspect of time-series analysis, which is the discovery of the correct number of auditory images in the scene.

It is important that a processing model that purports to explain the human perception of auditory scenes be able to dynamically detect the number of auditory images. In real-world situations, the human listener does not know the number of auditory images in a scene beforehand. In most cases, this number changes over time as new sound sources appear and old ones depart. An important feature of the model I have developed is that it is capable of dynamically determining the time-varying number of images in the auditory scene.

Since the clustering model in the previous section is couched in a Bayesian framework, there are principled ways to analyze the "fit" of the model to the data in each time frame. Further, it is relatively clear how to study the evolution of the model over time. There are two principles in conflict with each other:

1. We wish the fit of the model at each point in time to be as good as possible.
2. We wish the parameters of the model (particularly the number of clusters) to evolve slowly over time.

For example, given some data, we might find two different explanations, each corresponding to a hypothesis about the perception of the auditory scene. In the first, there are three images at all times T=1, T=2, T=3,... except at T=12, at which there are only two images. This

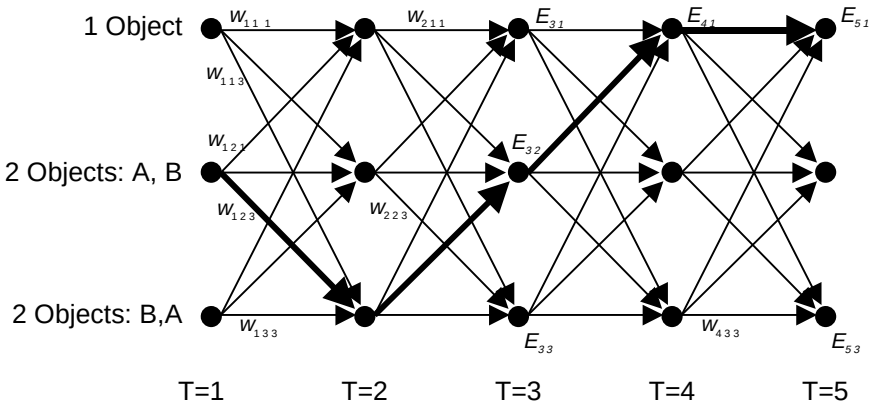


Figure 5-10**:** The Markov lattice used to optimize the number of images and their labels over time. Each node in the diagram represents a hypothesis about the number of images and the permutation labels in one time frame. In this schematic, the number of images is restricted to be either one or two at every time (in general, the number of images is not so restricted as shown here and so there are many more states). Each node (or *state*) is weighted by an error term E_{ti}, describing the goodness of fit of that clustering hypothesis with respect to the configuration of data at that time. At all time, $E_{t\,2} = E_{t\,3}$ since these configurations are equivalent except for labeling. Each arc is weighted by a transition weight w_{tij}, describing the likelihood of transitioning from state *i* at time *t* to state *j* at time *t*+1. (Not all nodes and arcs are shown with labels here, in order to reduce clutter in the diagram). The Viterbi algorithm is used to calculate the most probable sequence of states (shown with the bold arrows) given these weighting factors. In this schematic, the optimum solution is one in which there are two images in time T=1, T=2, and T=3, and one image at time T=4 and T=5.

hypothesis is equivalent to an auditory scene in which one image briefly disappears, and then reappears a short time later. A second explanation would be one in which there are three images at all times, including T=12, but the model fit at T=12 is not as good as that at other times[15]. This hypothesis is equivalent to an auditory scene in which there are three images at all times, and one of them is masked or otherwise obscured for one instant. Evaluation of these competing hypotheses depends on our knowledge about the world—how likely it is that there are two or three images, how likely it is that an image disappears and reappears, how poor the fit of the cluster model must be before it is considered unacceptable.

A Markov-chain model is suitable for addressing this problem. A schematic is shown in Figure 5-10. At each point in time, as described in the previous section, the EM algorithm generates several clustering hypotheses. For each of these, there are several possible labeling orders as described above. Each clustering-labeling hypothesis at each point in time is represented by a node on a directed graph. The progression from one time step to the next is represented by an arc connecting a node at T=*t* to a node at T=*t*+1. The graph is forward-fully-connected in the sense that there is an arc connecting each node at time *t* to each node at time *t*+1.

The number of columns in the lattice is equal to the number of time steps to be processed by the model. The number of rows, or *states*, depends on the number of hypotheses that will be considered at each time. As discussed above, in each time frame, the EM algorithm proposes hypotheses for several numbers of clusters: $G = 1, 2, \ldots G_{max}$. For a hypothesis with G

[15] "Model fit" here includes some sort of minimum-description-length criterion—of course it is the case that three clusters can always give a greater likelihood than two. But if the underlying density has two clusters, the three-cluster model is the weaker one.

clusters, there are $G!$ possible labeling orders (for $G = 3$, these are ABC, ACB, BAC, BCA, CAB, and CBA). Thus, given a value of G_{max}, at each time step there are

$$K = \sum_{G=1}^{G_{max}} G! \tag{5-16}$$

rows to consider, and K^2 arcs from one time step to the next. (These numbers grow very quickly due to combinatorial explosion. If G_{max} is 5, then K is 153 and so there are nearly 25 000 arcs to consider. If G_{max} is 6, then K is 873 and there are more than 750 000 arcs at *each* time step. Possible improvements to efficiency are discussed in Section 5.5.1).

Each node is weighted by an error function derived from the EM estimate of cluster likelihood. In particular, given the clustering hypothesis $G;\mu_i;\mathbf{K}_i$ that uses G clusters to explain a set of points $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, the error function

$$E = P(G)\prod_{j=1}^{n}\sum_{i=1}^{G}\frac{1}{G}N(\mathbf{x}_j;\boldsymbol{\mu}_i,\mathbf{K}_i) \tag{5-17}$$

with $N(\bullet)$ defined as in (5-13), is a simple criterion. This function is just the joint likelihood of all the points in $\mathbf{X}$ (again, with the assumption of independence) within the given cluster model, weighted by the number of clusters. I will use E_{tk} to denote the error function assigned to the node at time t in clustering hypothesis k, where $1 \le t \le t_f$ and $1 \le k \le K$. Each of the E_{tk} for hypotheses with the same number of clusters are equal, since they are the same except for renaming. Thus, in each time step, there are only G_{max} unique values for E_{tk}.

$P(G)$ in (5-17) denotes the prior probability that there are G images in the auditory scene. In principle, some kind of high-level or contextual knowledge could be used to create this function. For the examples computed here, a simple weighting by the number of images is used, so that

$$P(G) \propto \frac{1}{G} \tag{5-18}$$

This is a sort of minimum-description-length criterion.

Each arc is weighted by a transition weight that gives the likelihood of a transition from the start node of the arc to the end node of the arc. For example, in Figure 5-10, $w_{1\,1\,3}$ is the likelihood of making a transition from state 1 at time T=1 to state 3 at T = 2. The weights w_{tij} are computed as the product of two factors, that is

$$w_{tij} = q_{ij}\, m_{tij} \tag{5-19}$$

The q_{ij} are time-independent factors that describe the probability of going from a configuration in which there are G images to one in which there are G' images. This implies that $q_{ij} = q_{i'j'}$ whenever the number of images in state i is equal to that in state i' and likewise for j and j'. This factor, like $P(G)$, could be set contextually by some sort of top-down mechanism, but herein values are set through trial-and-error to give good performance on the examples.

The m_{tij} describe the match between the clustering hypothesis at time t and that at time t+1. In particular, let $c(\hat{\mathbf{x}}; k, t)$ be the likelihood of observing a value of random variable $\hat{\mathbf{x}}$ in state k at time t, that is

$$c(\hat{\mathbf{x}};k,t) = N(\hat{\mathbf{x}};\boldsymbol{\mu}_k,\mathbf{K}_k) \tag{5-20}$$

where μ_k and $\mathbf{K}_k$ are the cluster parameters for state k at time t. Then define

$$m_{tij} = \begin{cases} \sum_{l=1}^{n}\sum_{k=1}^{G_i} c(\mathbf{x}_{tl};k,t)c(\mathbf{x}_{t+1,l};k,t+1) & \text{when } G_i = G_j \\ \sum_{l=1}^{n}\sum_{k=1}^{Gj}\left[c(\mathbf{x}_{tl};k,t) + \sum_{r=G_j+1}^{G_i} c(\mathbf{x}_{tl};r,t)\right]c(\mathbf{x}_{t+1,l};k,t+1) & \text{when } G_i > G_j \\ \sum_{l=1}^{n}\sum_{k=1}^{G_i} c(\mathbf{x}_{tl};k,t)\left[c(\mathbf{x}_{t+1,l};k,t+1) + \sum_{r=G_i+1}^{G_j} c(\mathbf{x}_{t+1,l};r,t+1)\right] & \text{when } G_i < G_j \end{cases} \quad (5\text{-}21)$$

This messy expression can be interpreted as follows. The first clause, for $G_i = G_j$, operates when the number of clusters in state *i*, G_i, and that in state *j*, G_j, are equal. It says that the likelihood of a transition is given by the assumption that all channels stay in the same cluster from one time step to the next. That is, if channel #1 is in cluster A in time *t*, then it is in cluster A at time *t+1*, while if it is in cluster B at time *t*, then it also is in cluster B at time *t*+1, and so forth. Similarly, if channel #2 is in cluster A in time *t*, then it is in cluster A at time *t+1*, and so forth for all channels.

The second clause, for $G_i > G_j$, is the *death clause*, since if $G_i > G_j$, it means that one or more images have vanished between time *t* and time *t*+1. It says that the likelihood of a transition is given by the assumption that all channels stay in the same cluster if the cluster is present in both time frames. If the cluster has died, then it doesn't matter where the channels belonging to it go in the next time frame

The third clause, for $G_i < G_j$, is the *birth clause*, since if $G_i < G_j$, it means that one or more images have appeared between time *t* and time *t*+1. It says that the likelihood of a transition is given by the assumption that all channels stay in the same cluster if the cluster is present in both time frames, but channels may freely switch from any cluster to one of the newly appeared clusters.

Through this definition, m_{tij} formalizes the notion that we wish the clusters to be positioned such that it makes the assignment of channels to images as stable as possible.

Using the E_{tk} and the m_{tij} values, the Markov model for time-series analysis can now be presented. We wish to find a sequence of states $A = a_1, a_2, ..., a_{tf}$ through the Markov lattice (as shown in Figure 5-10) that minimizes the total sequence of errors and transition weights; that is, that minimizes

$$E_{1,a_1} m_{1,a_1,a_2} E_{2,a_2} m_{2,a_2,a_3} \cdots m_{t_f-1,a_{t_f-1},a_{t_f}} E_{t_f,a_{t_f}} = E_{1,a_1} \prod_{t=2}^{t_f} m_{t-1,a_{t-1},a_t} E_{t,a_t} \quad (5\text{-}22)$$

The obvious way to find this minimizing sequence is to evaluate (5-22) for each of the possible sequences *A*; unfortunately, there are far too many of them (K^{t_f} in all, with *K* defined as in (5-16)).

The Viterbi algorithm (Therrien, 1989, pp. 204-211) is a dynamic-programming technique that shows how to find the minimizing sequence for (5-22) in time proportional to the size of the whole lattice, that is $K\,t_f$, which is quite an improvement on K^{t_f}. Using the Viterbi algorithm, the best sequence of states given the weighting terms E_{tk} and m_{tij} can be computed—such a sequence might be the one shown with dark lines on Figure 5-10. This sequence is a joint estimate of the number of images at each point in time, and the correspondence of labels from one time frame to the next.

5.2.6. Dynamic cluster analysis: channel-image assignment

The stages of processing described in the previous two sections estimate the number and positions of images in modulation space at each step in time. The last stage of processing is to assign each cochlear channel to a cluster at each time. It is through this assignment that the

cochlear channels provide evidence about the features of the various auditory images in the scene.

For the purposes of simplifying the presentation, I will only consider hard assignment, in which each channel is assigned to only one image at a time (*exclusive allocation* in the terminology of Bregman (1990), also called *disjoint assignment* by Summerfield et al. (1990)). This clustering framework would easily allow soft assignment (for example, to say that at time T=4, channel #26 is 80% part of image A, and 20% part of image B). However, it is not clear that position in the modulation space is the best basis on which to hypothesize conjoint channel assignment. It would also make feature extraction (as in Chapter 6) more complicated. Nonetheless, the clustering and auditory-image-segregation parts of the model makes no assumptions about whether or not channels are exclusively assigned to auditory images.

In order to estimate the assignment of channels to images, the Viterbi algorithm is used again. In this instance, the transition lattice takes the form of a Hidden Markov Model (HMM) (Therrien, 1989, pp. 189-194), as shown in Figure 5-11.

At each point in time, a cochlear channel occupies a given point in modulation space; this is the two-dimensional feature vector **x** that was calculated in Section 5.2.2. Each image at each point in time corresponds to one of the Gaussian clusters that was computed in the previous stages.

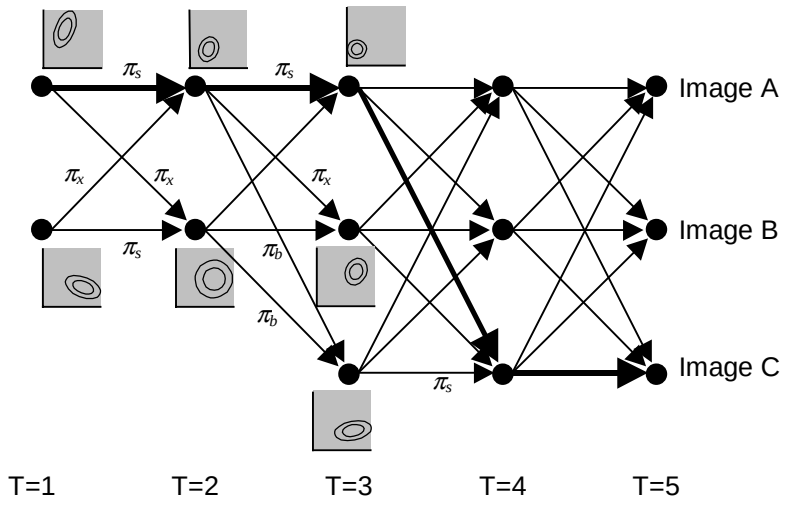


Figure 5-11**:** The Hidden Markov Model lattice used to assign cochlear channels to images. Each node represents the probability density function (PDF) associated with a particular auditory image at one point in time. Each arc represents the *a priori* probability that an channel stays in the same image, compared to that of moving from one image to another, at a certain time. The weights π_s, π_x, π_b, and π_d represent the weight assigned to a channel staying in a single image, moving to another image that was present at the previous time, moving to an image that has just been born, and moving away from a channel that has died, respectively (the last is not shown in this diagram). In this example, at T=1 and T=2 there are two images in the scene, and at T=3, T=4, and T=5 there are three auditory images. Based on the cluster model developed in the previous sections, each image is assigned a Gaussian cluster in modulation space at each time (shown as small two-dimensional plots next to some of the nodes). This PDF determines the probability that a cochlear channel that is part of that image will be modulating in a certain way. Based on this lattice, the Viterbi algorithm is executed 54 times, once for each cochlear channel. For each channel, the optimal path through the sequence of clusters is computed. For example, the path for one channel is shown in bold; this channel is a member of image A at time T=1 through T=3, and a member of image C at time T=4 and T=5.

Thus, the instantaneous probability that a channel is a member of a particular cluster at a particular time is given by

$$p(c_i \mid \mathbf{x}) = \frac{\rho_i N(\mathbf{x}; \boldsymbol{\mu}_i, \mathbf{K}_i)}{\sum_{j=1}^{G} \rho_j N(\mathbf{x}; \boldsymbol{\mu}_j, \mathbf{K}_j)} \tag{5-23}$$

where c_i indicates membership in image i, $\mathbf{x}$ is the feature vector measured from the channel, ρ_i is the prior probability of membership in class i (the ρ_i are taken as uniform herein), $N(\bullet)$ is defined as in (5-13) given that μ_i and $\mathbf{K}_i$ are the parameters of the cluster corresponding to image i, and G is the number of images in the scene at that time.

Since the denominator of (5-23) doesn't vary with class, and the ρ_i are taken as uniform, the error weight for each node of the HMM lattice i relative to cochlear channel #k is given by

$$E_{ikt} = N(x_k; \mu_i K_i) \tag{5-24}$$

where μ_i and $\mathbf{K}_i$ are given relative to the particular time step.

Weights on the arcs in the lattice are given as free parameters π_s, π_x, π_b, and π_d. These parameters are the *a priori* probabilities respectively that: a cochlear channel stays in the same image from one time frame to the next; a cochlear channel moves from an image in one time frame to a different image in the next when both images are present in both time steps; a cochlear channel moves from an image in one time frame to a new image that has just been born; a channel moves from an image that has just died to another image. Each weight w_{ij} going from node i at time t to node j at time t+1 is assigned one of these values in the obvious way. The arc weights are independent of the cochlear channel and of time, by hypothesis.

In the preceding section, the goal was to derive the sequence of clustering-labeling solutions that minimized the error function in (5-22). The goal of this second stage is to derive for each channel k the sequence of states $B_k = b_{k1}, b_{k2}, \ldots b_{ktf}$ that minimizes the overall channel-image assignment error; that is, that minimizes

$$E_{b_{k1}k1} w_{b_{k1}b_{k2}} E_{b_{k2}k2} w_{b_{k2}b_{k3}} \cdots w_{b_{kt_f-1}b_{kt_f}} E_{b_{kt_f}kt_f} = E_{b_{k1}k1} \prod_{i=2}^{t_f} w_{b_{ki-1}b_{ki}} E_{b_{ki}ki} \tag{5-25}$$

As with the optimization of (5-22), the minimal such sequence can be computed efficiently with the Viterbi algorithm. The resulting sequence B_k is the class membership of each cochlear channel at each time. The ensemble of such sequences over all cochlear channels may be considered a partition of the auditory scene. The function

$$I_{ikt} = \begin{cases} 1 & \text{if } B_{kt} = i \\ 0 & \text{if } B_{kt} \neq i \end{cases} \tag{5-26}$$

which indicates whether channel k is assigned to a particular auditory image i at time t, is termed the *channel-image assignment function*.

Section 5.4 of the present paper considers the degree to which the partition calculated in this manner can correctly predict the response of human listeners on some psychoacoustic tasks.

5.2.7. Limitations of this clustering model

It is important to recognize the limitations of this approach in estimating the number of images and their positions in modulation space. Some of them are:

- The overall sequence of states given is not globally optimal, for two reasons. First, the EM algorithm only gives a locally-optimal solution, not a globally-optimal solution, and so it might be the case that a different starting condition for EM would give a better fit to

the data for some number of clusters in some frames. For example, it is known that using the ISODATA algorithm (Therrien, 1989) to produce an initial estimate gives good results. Second, when using the Viterbi algorithm to determine the optimal time-series of clusters, there is only one hypothesis considered for each number of clusters G. Referring to Figure 5-9 at T=3, for G=2, either (a) or (d) will be the hypothesis, but they are not compared to each other. (For G=3, (b) and (e) *are* compared to each other since they are only different in labeling). Thus, it might be the case that (a) is the instantaneously optimal clustering pattern, and is thus selected in the EM step as the hypothesis for G=2, but that (d) is actually a better fit in the overall time series.

In order to achieve this sort of optimization would require joint processing of the EM algorithm and the Viterbi algorithm, so that the entire continuum of error functions possible from various clustering solutions is available to the Viterbi process. This seems to be a computationally intractable method, and so the tradeoff presented here optimizes the instantaneous solution before the time-series solution. It is unknown how much better the time-series solution might be if it were jointly optimized with the clustering model.

A possible intermediate improvement would be to generate multiple hypotheses for each number of clusters G, by using different EM starting states. Informal investigation of the solution space indicates that most configurations of data points in modulation space lead to three or four general classes of EM solutions when the EM algorithm is run multiple times. Therefore, at the cost of a perhaps an order of magnitude in execution time, it would be possible to consider multiple (but still not all) clustering hypotheses for a particular data-point configuration.

- Similarly, the constraint that leads to equation (5-21) is only a heuristic. In time frames where cochlear channels actually do move from one image to another without a birth or death occurring, the constraint is violated and it is possible that this constraint gives the wrong answer. For example, occasionally the Viterbi model seems to switch the labels of two images, and the resulting claim made by the model is that all of the channels belonging to image A now belong to image B and vice versa.

  In order to achieve a more optimal solution here would require joint processing of the time-series cluster analysis and the image-channel assignment. As with the previous point, it would be computationally very expensive to do this, and so treating these stages of the model as separable, while not strictly true, seems to be a good tradeoff between accuracy and efficiency.

- There are some prior constraints on cluster behavior that might be desirable but that are not directly achievable through the Markov model of cluster movement. For example, it might be desirable to assert that the birth and death of a single auditory image cannot occur within 50 ms of each other (that is, that the perceptual duration of an auditory image can be no less than 50 ms). This sort of constraint cannot be implemented directly within a Markov framework (although the results could be "cleaned up" afterward by additional post-hoc processing).

- The model of image birth/death is not very sophisticated. For example, if a new image arrives in the scene at the same time as a pre-existing one departs, then the number of clusters will be unchanged. The present model will interpret this by trying to find an explanation that provides continuity between the images over time, even though this is not a veridical interpretation of the data. Further, when an image birth or death occurs, there is no part of the present model that can describe clearly what happened—for example "image A has split into new images A and C." An extension to the concept of cluster labeling, with more-sophisticated continuity metrics, could be used to attack this problem.

- The entire time-series is processed at once. This has two disadvantages. First, it means that sound events at one time can affect the perception of sound events arbitrarily far away in time. Second, it means that the model is not really a *process model* in which data are processed in real-time as they come in. In the present model, all of the perception is done in retrospect, after all sound has arrived. Both of these problems would be solved by making the Viterbi stages only operate on short temporal windows, which could be interpreted as echoic memory. So long as the features that are derived from sound events are still within the extent of the window, they are able to affect and be affected by the probabilistic reasoning. As old sound events exit the window, their percepts would become fixed and unable to change.

5.2.8. Feature analysis

The final step in the model is to use the channel-image assignments calculated from the clustering stages to determine the perceptual features of the auditory images. The proper way to do this is still a topic for future work, although some basic feature-detection experiments have been performed. Results from these will be shown in Section 5.4 and Chapter 6.

An important principle is that *the channel-image assignments are not themselves the auditory images*. Rather, the channel-image assignment function determines the channels that provide evidence that can be used to make judgments by a particular image at a particular time. To actually compute the features is a process of integrating evidence when it is available, and making guesses when it is not.

It is clear from perceptual data, such as that on phonemic restoration (Warren, 1970), that sophisticated high-level models are used in the human auditory system to help make such guesses about missing data. Ellis (1996a) has outlined a computational theory and argued that it is sufficient for making such judgments, and Martin (1999) has shown how it is possible to build sophisticated structural sound models from the data available in the autocorrelogram representation. There is an extensive niche in the artificial-intelligence literature that discusses the problem of making inferences from incomplete data.

To follow such a hypothesis, the auditory image in the mind would be a sound model such as those described by Martin (1999). Such a model—say, for a perceived violin—has parameters that maintain the perceived pitch, loudness, playing style, etc. of the perceived object. Based on the channel-image assignment, evidence from the acoustic signal is integrated within the model-formation process to fill in the parameters of the perceptual model. The perceptual features are not calculated directly from the acoustic signal, but are rather induced through the model. A theory like this one has the advantage that when gaps arise in the evidence stream due to occlusion of the signal or a shift in attention, it is not the case that these gaps necessarily become part of the perceived model's parameter settings. If they do not, then the evidentiary gaps do not correspond to perceived gaps, and the features perceived in the sound are interpolated based on the previous parameters of the model.

Another way to say the same thing is that *a gap in evidence is not evidence of a gap*. In a complex sound scene, the auditory system must be prepared to deal with occlusions and attentional shifts that cause direct evidence for the features—or even for the existence itself—of a particular auditory image to be lost momentarily. This happens more or less continuously in the real world of hearing.

A simple way to use the channel-image assignments that does not take advantage of such sophisticated mechanisms is to treat them as masks for the cochlear filterbank data. That is, to hypothesize that an auditory image at some time consists simply of the output of the cochlear filterbank for the set of channels assigned to the object. Based on this model of the auditory image, standard perceptual feature models can be applied directly to it. For example, the autocorrelogram model of pitch perception, as discussed in Chapter 2, Section 2.1.1, can be applied to the outputs of the filter channels that are selected at a given time. This is the

model of the features of auditory images suggested by Meddis and Hewitt (1992), although of course their image-formation model is quite different than the one suggested here.

5.3. Model implementation

In this section, computational implementation issues are briefly described. Source code for the various stages of the model is available from me via electronic mail.

5.3.1. Implementation details

The model is implemented in two parts. The correlogram analysis stage was implemented by Martin (1999) in C++, and therefore runs relatively quickly, although not in real-time. The input to this stage is the sound to be analyzed, stored in an uncompressed audio file format such as AIFF. On a desktop PC (450 MHz Pentium-III) it takes approximately two minutes to compute the autocorrelogram for ten seconds of sound. The exact computation time depends on the number of filters used in the gammatone filterbank and the time resolution used to compute the autocorrelogram. The output of this stage is an analysis file that contains the entire autocorrelogram in a binary data representation, represented as a series of frames. This analysis file is many times larger than the original sound file.

The modulation analysis and grouping stages are implemented in Matlab. The input to these stages is the autocorrelogram analysis produced in the previous stage. For each frame of autocorrelogram data, the modulation parameters in each channel are computed, and several clustering hypotheses are generated with the EM algorithm. After the modulation analysis and grouping for each frame is complete, the two Viterbi passes through the grouping hypotheses are used to determine the sequence of images and to assign channels to auditory images. This stage runs very slowly due to the Matlab implementation; on a desktop PC it takes approximately three hours to compute the image grouping for ten seconds of sound, with G_{max}=3. This part of the model could be optimized by reimplementing it in C++ with more attention to the computational complexity of the particular techniques used. The output of this stage is the image-channel assignment, which associates the evidence present in each cochlear channel with a particular auditory image at each point in time.

More radical optimization will be needed if the model is to be practically useful for investigating more complex scenes. As the calculations based on Eq. (5-16) show, thousands or even millions of arcs per time frame would need to be considered if more than four auditory images were allowed to occur. This is computationally intractable on any computer system today using the method described. The most immediate improvement would be to find a way to arrive at the labeling order of the cluster hypotheses directly, rather than including this as part of the Viterbi optimization stage. The largest combinatorial-explosion problem comes from having to consider each order of cluster labels as a separate clustering hypothesis. On the other hand, it is unknown whether the human listener can actually perceive more than three or four auditory objects in a scene at once (see Section 5.5.4).

5.3.2. Summary of free parameters

As seen in passing in previous sections, there are a number of free parameters in the model, the values of which can be used to tune the behavior of the model. In principle, the model parameters should be tuned through reference to quantitative psychophysical data (see also Section 5.5.4). In the model is successful, once the parameters are tuned with respect to one experiment in this way, the model should be able to quantitatively predict the results of other experiments without altering the settings. However, in the present paper, the evaluation of the model is not so ambitious—only qualitative matching to experimental results is shown. On

the other hand, all of the results shown in Section 5.4 were achieved with a single set of model parameters. These are the settings that are shown as "default" settings below.

Free parameters in autocorrelogram analysis

Symbol	Meaning	Default value
S_a	Audio sampling rate	24000 Hz
S_c	Autocorrelogram frame rate	100 Hz
k_t	Time constant of one-pole smoothing applied to correlator output	25 ms
N_o	Number of octaves covered by cochlear filterbank	9
N_s	Number of filters per octave in cochlear filterbank	6
$l_{min} - l_{max}$	Range of lag axis maintained in one autocorrelogram frame	0.5 – 20 ms
S_l	Sample spacing of lag axis (logarithmically spaced samples)	150 samples/frame

Free parameters in modulation detection

Symbol	Meaning	Default value
ΔT	Spacing between correlogram frames for modulation analysis	40 ms
P_m	Minimum power allowing a cochlear channel to be included in analysis during one frame (relative to most powerful channel in that frame); channels with less power are discarded for that frame	-25 dB
Q	Interpolation factor used for resampling autocorrelation signal before cross-correlation. Controls quantization of cross-correlation results.	Factor of 10
D_{max}	Maximum detectable cross-correlation between autocorrelation signals	40 samples of resampled lag axis

Free parameters in EM clustering

Symbol	Meaning	Default value
G_{max}	Maximum number of clusters	2
$\hat{\boldsymbol{\mu}}_i$	Initial estimates of cluster means	Set randomly for each frame and clustering hypothesis within parameter space

$\hat{\mathbf{K}}_i$	Initial estimates of cluster covariances	Fixed at $\begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}$
λ_m	Minimum covariance eigenvalue	0.03
E	Convergence threshold (difference in likelihood from one iteration to the next that terminates EM iteration)	10^{-3}

Free parameters in transition model

Symbol	Meaning	Default value
q_{ij}	Prior likelihood of changing number of clusters	10^{-4} if $i > j$ (birth), 1 otherwise
π_s	Prior likelihood of a channel staying in the same cluster from one time frame to the next	1
π_x	Prior likelihood of a channel switching clusters	10^{-18}
π_b	Prior likelihood of a channel joining a just-born cluster	10^{-18}
π_d	Prior likelihood of a channel moving from a dead cluster to different cluster	10^{-18}

5.4. Psychoacoustic tests

In this section, I will evaluate the behavior of the auditory-segregation model on several simple psychoacoustic test stimuli. The goal is only to show qualitative agreement with human behavior on simple examples; more complex musical sounds will be considered in Chapter 6. For each of these sorts of stimuli, a great deal more is known about human perception that can be treated here; it is surely not possible to treat the entire psychoacoustic literature on the perception of auditory scenes within a single chapter. More extensive testing, and investigation of the use of the model to make quantitative predictions, are discussed in Section 5.5, but left mainly as a topic for future research.

5.4.1. Grouping by common frequency modulation

The so-called *McAdams oboe* was first used as an experimental stimulus by McAdams in his dissertation (McAdams, 1984). It was created to investigate coherent frequency modulation as a cue to the formation of auditory images, and may be generated with additive synthesis. For the stimulus used here (slightly different that the ones that McAdams used), a ten-harmonic complex tone was synthesized for 10 s. After the first 2 s, a coherent vibrato (frequency modulation) at 5 Hz was applied to the even harmonics only. The modulation depth was ramped from 0% to 4% (that is, until $\Delta f / f = 0.04$) in 4 s, and then maintained at 4% for the final 4 s.

The percept for most listeners when they hear this sound is that of a single complex tone splitting into two sounds. One of the two sounds is clarinet-like (due to the unmodulated odd harmonics), and one of which is somewhat soprano-like and one octave higher (due to the

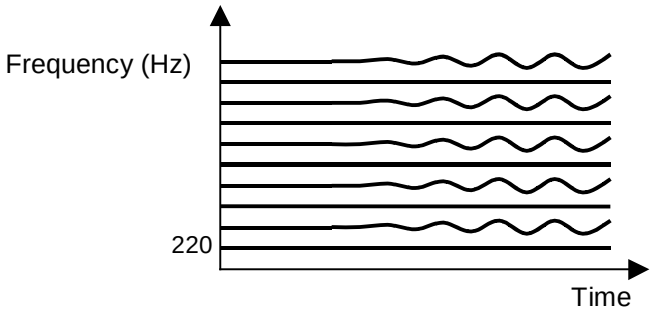


Figure 5-12: A schematic of the McAdams oboe sound. The stimulus begins with 10 harmonics of a 220-Hz fundamental, equally weighted in sine phase. After 2 s, a coherent vibrato at 5 Hz is applied to the even harmonics only. The modulation depth of the vibrato is ramped from 0% to 4% in 4 s, and then maintained at 4% for another 4s. The percept changes over time; at the beginning, a single oboe-like tone is heard, but at some point the percept switches to two sounds, a clarinet-like sound consisting of the unmodulated odd harmonics, and a soprano-like sound with pitch one octave high consisting of the even harmonics.

modulated even harmonics). The stimulus is shown in schematic in Figure 5-12. I graphically presented the output of the early stages of processing for a slightly different version in Chapter 2, Figure 2-2 and Figure 2-3.

A comparison sound, in which the maximum modulation was only 0.4%, was also created. The two stimuli can be heard as Sound Example 5-1 on my web page. The sounds were synthesized at a sampling rate of 24000 Hz with a digital additive synthesis procedure written in the digital synthesis language SAOL (Scheirer and Vercoe, 1999). They were written to computer files, and then processed by the model described in Section 5.2. The SAOL code for the synthesis procedure is given in Appendix B.

The auditory channel-image assignment plots for the two stimuli (the standard McAdams oboe, and the comparison sound) are shown in Figure 5-13. As can be seen in this figure, at the outset of the standard stimulus, there is only one auditory image present in the scene. At approximately 5.5 s into the sound, the percept splits into two auditory images. The energy in the filterbank near the odd harmonics continues as part of the previous image, while the energy in the filterbank near the even harmonics is grouped as evidence for a second auditory image. In the comparison stimulus, all of the channels are grouped into a single image throughout the stimulus; no perceptual segregation is predicted.

Thus, the model successfully predicts three important aspects of the perception of these stimuli. First, it predicts that some threshold of frequency modulation in the harmonics must be present for perceptual segregation to occur (although no attempt has been made to quantitatively model the particular threshold). Second, it predicts that when there is sufficient frequency modulation, the percept of the signal changes over time, from one image to two images. The model uses no prior knowledge of the signal to make this prediction, but produces it dynamically through signal analysis. Third, the model predicts that the odd harmonics and the even harmonics are assigned to different and coherent groups of auditory images.

Based on the evidence found in these partitioned regions of the time-frequency space, the features of the resulting images may be analyzed. For example, Figure 5-14 shows pitch estimates of the auditory images in the scene. These are calculated by applying the Meddis-Hewitt (1991) model to only those channels that have been assigned to each image (this procedure will be discussed in more detail in Section 6.3.3). As seen in this figure, there is a good correspondence between the pitches that can be automatically extracted from the auditory images and the known human perception for a signal of this sort.

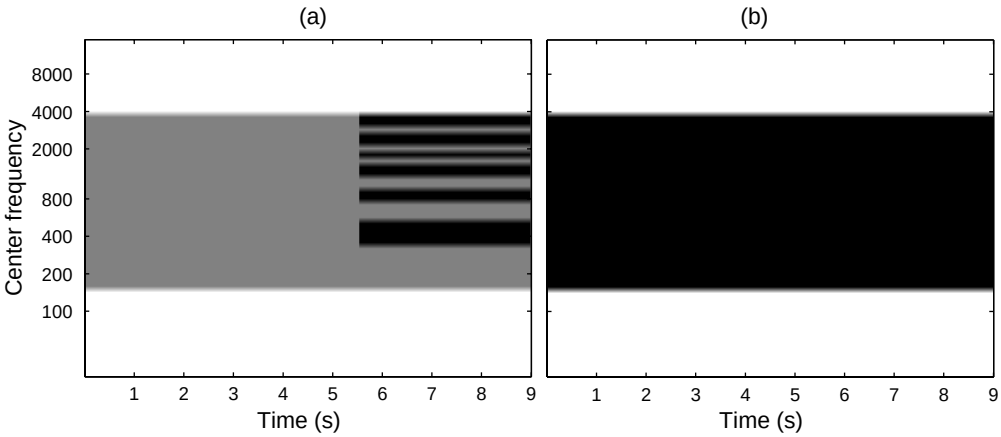


Figure 5-13**:** Model output for the two McAdams oboe stimuli. Each plot shows the assignment of time-frequency energy to the auditory images detected by the model. The color of each time-frequency cell indicates the channel-image assignment. Time-frequency cells that are not colored are not assigned to any image; these are cells with little sound energy, as described in Section 5.2.2. In (a), the standard stimulus, in which 4% maximum frequency modulation is applied to the even harmonics, is shown. In this stimulus, the model detects two auditory images, one (the gray image) that is assigned all of the sound at the beginning and the energy corresponding to the odd harmonics at the end, and one (the black image) that is assigned the energy corresponding to the even harmonics beginning approximately 5.5 s into the sound. In (b), a comparison stimulus, with 0.4% frequency modulation applied to the even harmonics, is shown. In this stimulus, no image segregation occurs, and all of the time-frequency energy is assigned to a single image throughout the sound.

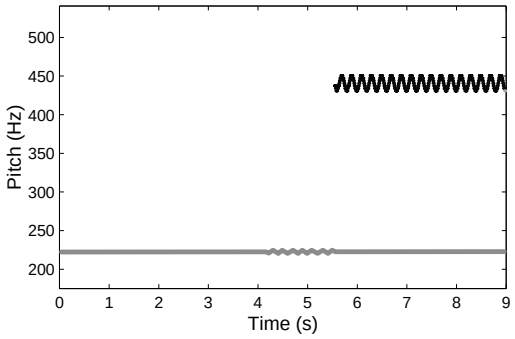


Figure 5-14: Pitch analysis of the auditory images suggested by the channel allocation shown in Figure 5-13(a). At the beginning of the stimulus, there is only one auditory image, with pitch (shown in gray) flat at 220 Hz. After the modulation of the even partials begins, the single image splits into two images, with one pitch remaining at 220 Hz, and the other (shown in black) modulating about 440 Hz. This agrees with human perception of this auditory stimulus.

Discussion

McAdams' work with stimuli of this sort focused on the role of different sorts of vibrato in promoting the fusion and segmentation of auditory images. Mellinger's dissertation (1991) was the first computational project to demonstrate successful perceptual-based segmentation of this sort of stimulus. His system operated by filtering the time-frequency plot with "modulation kernels"—time-frequency packets that could be used to select common FM components of the signal. This approach is somewhat similar to the more general time-frequency basis decomposition recently proposed by Shamma and colleagues (Wang and Shamma, 1995; Shamma, 1996; Versnel and Shamma, 1998). Early CASA models by Ellis (1994) and Brown and Cooke (1994b) were also targeted toward stimuli that could be easily

created through additive synthesis; that is, voiced speech and musical signals composed only of harmonic sounds.

At one time, there was general agreement that the auditory grouping for these sort of stimuli was governed by coherent frequency modulation. This is the explanation promoted by McAdams in his presentation of experimental results using these stimuli. However, this agreement no longer maintains; in particular, Carlyon (1991; 1994) has argued on the basis of more extensive psychophysical testing that the actual basis of auditory grouping in these stimuli is the harmonicity of the signal. That is, as the even harmonics move away from exact harmonicity with the odd harmonics, a pitch-based grouping mechanism selects them as part of a different auditory group.

The present model does adhere to the viewpoint that grouping is based on common modulation in these stimuli. Future work should examine more closely the stimuli developed by Carlyon and others to distinguish the harmonicity hypothesis from the common-modulation hypothesis. A general discussion of pitch-driven segregation models in comparison with this model is presented in Section 5.5.2.

5.4.2. The temporal coherence boundary

Van Noorden's (1977) early work on what are now called *auditory streams* was among the first study of this subject. He developed stimuli composed of a series of repeated tones to investigate the way in which sequential integration is performed in the auditory system. His "galloping" stimuli are shown in schematic in Figure 5-15.

Van Noorden's stimuli consisted of a series of tone pips, each 40 ms in duration. Each stimulus consists of interleaved sequences of fixed-frequency tones (denoted *F*) and variable-frequency tones (denoted *V*). Each *F* tone is at 1000 Hz; the frequencies of the *V* tones are 1000-Δf Hz, where Δf denotes the difference in frequency between the two tone streams. In Van Noorden's own experiments Δf varied from 2 semitones (109 Hz) to 14 semitones (555 Hz). The presentation rate of the tones is denoted by Δt and is fixed for each trial. Van Noorden investigated presentation rates from 60 ms to 150 ms.

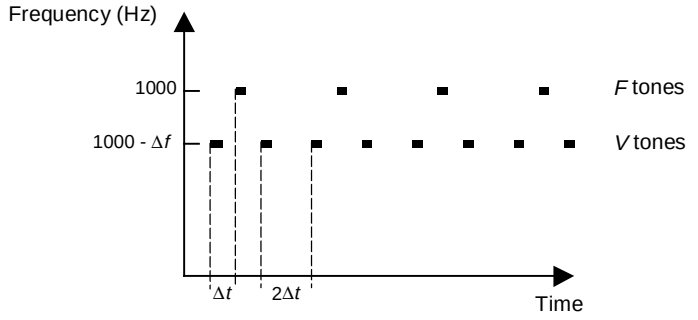


Figure 5-15**:** Stimuli used by Van Noorden (1977) to study temporal coherence in auditory stream perception. Each tone pip is 40 ms long; the onsets of tones are separated by Δt, which differs in different trials. The higher tone, called *F* for fixed frequency, is always at 1000 Hz. The lower tone, called *V* for variable frequency, is at some lower frequency 1000-Δf, which differs in different trials. Depending on the relative values of Δt and Δf, the percept is one of the following: (A) a galloping rhythm, with all tones perceptually connected (streamed) together, or (B) two separate isochronous streams, with the higher stream at 1/3 the rate of the lower, or (C) volitional or attention-based control of percepts (A) and (B), where the subject can choose or switch between them at will.

The subject's perception of the stimulus varies depending on the values of Δf and Δt. For sufficiently low values of Δf and/or large values of Δt, the subject perceives a single stream with a galloping rhythm: VFV-VFV-VFV. For stimuli in which Δf is relatively large and Δt relatively small, the subject perceives two streams, V-V-V-V and F---F---F---. The lower-pitched stream is at twice the rate of the higher-pitched one. For stimuli in-between these extremes, the subject experiences volitional, attention-based control and is able to switch between these percepts at will.

The boundary in the Δf - Δt plane above which the subject is only able to experience a two-stream percept is called the *temporal coherence boundary*. The boundary below which only a one-stream percept is possible is the *temporal fission boundary*.

I synthesized five stimuli at a sampling rate of 24000 Hz using a digital synthesizer written in SAOL. For each, the tone pip duration was 40 ms, including 5-ms linear onset and offset ramps, and the frequency of the *F* tones was 1000 Hz. The value of Δf and Δt for each stimulus is shown in Table 5-1. SAOL code that generates the stimuli is shown in Appendix B. They may be heard as Sound Example 5-2 on my web page.

The output of the segmentation algorithm for this set of stimuli is shown in Figure 5-16. As can be seen there, the model predicts that temporal coherence depends on the time/frequency spacing of the tone pips. For stimuli S1 and S3, where the tones are closely spaced in frequency relative to time, only one auditory stream is perceived by the model. For stimuli S2 and S5, where the tones are widely spaced in frequency relative to time, two streams are perceived by the model. The percept in stimulus S4 is ambiguous; this may be considered a situation that lies between the temporal coherence and temporal fission boundaries for this model. As with the examples in Section 5.4.1, more work would be needed to arrive at quantitative, rather than as qualitative, predictions.

A short discussion of the model output for stimulus S5 is illuminative. From a first inspection of the model output in Figure 5-16, it may not be clear that stream segregation is actually occurring for this stimulus. This illustrates a point made in Section 5.2.8: the channel-image assignment function is not *itself* the auditory image, but is rather a way of assigning evidence from different parts of the sound energy to the auditory images so that feature detection may proceed. For example, implementing a simple pitch detector on the output of the model for stimulus S5 results in the pitch-tracks shown in Figure 5-17. In these pitch-tracks, it may be observed that the auditory images estimated from the channel-image assignments seem to correspond to the human perception. It is on this basis—the features perceived in the set of auditory images—that the performance of the algorithm is most readily evaluated (see also Section 5.5.5).

Stimulus	Δf	Δt
S1	50 Hz	50 ms
S2	500 Hz	50 ms
S3	50 Hz	100 ms
S4	500 Hz	100 ms
S5	800 Hz	100 ms

Table 5-1: Stimulus conditions (for Van Noorden-style streaming stimuli) used to test the model.

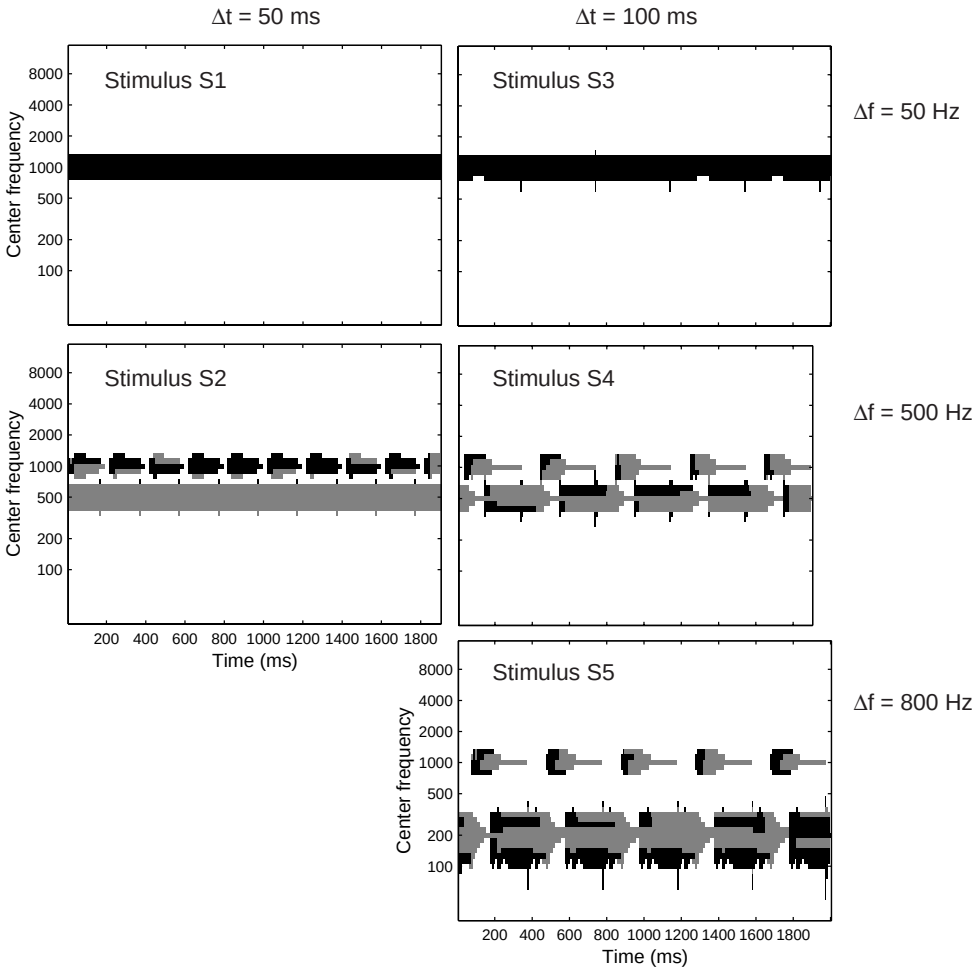


Figure 5-16: Model output for the five temporal coherence stimuli. Each plot shows the assignment of time-frequency energy to the auditory images detected by the model. The color of each point in time-frequency space indicates the channel-image assignment. For stimuli S1 and S3, one auditory image is perceived. For stimuli S2 and S5, two images are perceived. For stimulus S4, the output is ambiguous. See text and Figure 5-17 for discussion of stimulus S5 result.

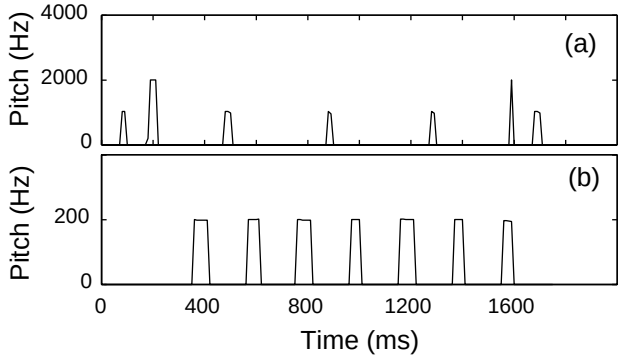


Figure 5-17**:** Pitch analysis of the auditory images that result from the channel-image assignment shown in Figure 5-16. Plot (a) is the pitch of the first image (the one in gray in panel S5 of Figure 5-16) whenever its power is above a threshold; plot (b) is similarly the pitch of the thresholded second image (the one in black in panel S5 of Figure 5-16). The pitches are computed for each image by applying the Meddis and Hewitt (1991) pitch model to the subset of channels assigned to that image. From these pitch-tracks, it may be observed that the correct percept is presented by the model output: one stream at 1000 Hz, and a second stream twice as fast at 200 Hz.

Discussion

The model predicts the human perceptions on these stimuli because sounds centered at a particular frequency bleed through to cochlear filters at nearby frequencies. That is, when a pure tone with frequency 1000 Hz stimulus the cochlear filterbank, not only the cochlear filter centered at 1000 Hz, but several nearby filters, will be excited. Moreover, the output of these filters is not at their own center frequency, but is an attenuated 1000 Hz tone. When the next tone begins at 950 Hz (for stimulus S1), there is a rapid and coherent frequency modulation among the outputs of this *entire group* of filters. The group stops responding at 1000 Hz and begins instead to respond at 950 Hz. This frequency modulation is observed as a coherent period modulation in the autocorrelogram, as discussed in Section 5.1.

For stimuli in which there is more frequency separation between the tones, such as S2, there is less cross-response between the set of filters that responds to one tone and the set that responds to the other. In this case, the amplitude-modulation differences at the tone onsets dominate—first one group of filters is stimulated, and then the other. Based on the alternating, incoherent, amplitude modulation, the sound is partitioned into two streams.

Thus, the model predicts that the temporal coherence boundary is a direct effect of the shape and time response of the cochlear filters. If this prediction is correct, changes to the bandwidth or adaptation properties of the cochlear filterbank, for example, cochlear hearing loss or large doses of aspirin, should alter the temporal coherence boundary. The small amount of experimental evidence, for example by Rose and Moore (1997), on the relationship between perceptual grouping and hearing loss seems to disconfirm this prediction, however.

Beauvois and Meddis (1996) have reported on the construction of an extensive low-level model that is also capable of predicting human perceptions on these stimuli, including the quantitative thresholds. The model that I have presented is generally compatible with theirs. Theirs includes more-sophisticated cochlear modeling, while mine uses more sophisticated feature-detection and pattern classification in the latter stages. It would likely be possible to replace wholesale the simple linear cochlea model I have used with theirs. This might enable the present model to make quantitative predictions of the Van Noorden results.

Recently, Vliegen and Oxenham (1999) presented experimental results showing that stream segregation, purportedly of the same sort, can be induced by alternating stimuli that occupy the same spectral region. In their stimuli, low-frequency (F_0 between 100-200 Hz) complex tones with alternating F_0 were high-pass filtered at 2000 Hz to remove any resolved harmonics. Subjects exhibited largely the same behavior on these stimuli as on the Van Noorden stimuli. Vliegen and Oxenham argue that the models of Beauvois and Meddis (1996) and McCabe and Denham (1997) are unable to account for these results.

The present model is also unable to account for these results, as the only sort of auditory streaming that it predicts is that occurring when the multiple streams occupy different spectral regions at the same time. However, the present model is sensitive to the periodicity cues that must be the basis for the segregation in the Vliegen and Oxenham stimuli.. Further, it extracts different features from the two sounds. Because of this, it is possible that a different sort of stream-formation mechanism, within the same basic framework, would account for their data. A follow-on study by Vliegen *et al.* (1999) found that the role of spectral cues outweighed the role of temporal cues when subjects were asked to integrate stimuli into a single stream rather than "hear out" one or the other sub-sequences (that is, in an attempt to elicit the fission boundary rather than the coherence boundary).

5.4.3. Alternating wideband and narrowband noise

Warren developed a stimulus that is now commonly used to illustrate what Bregman (1990) calls the "old-plus-new" principle in auditory grouping. It consists of the alternation of a wideband noise with a narrowband noise, as shown in Figure 5-18.

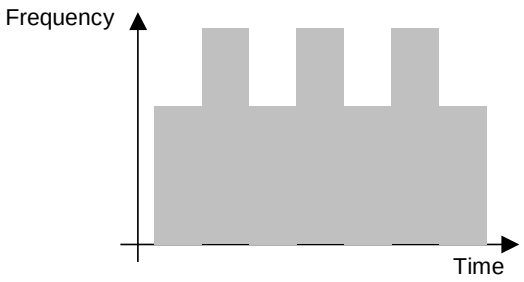


Figure 5-18**:** Alternating wideband and low-passed noise, commonly used to illustrate Bregman's "old-plus-new" grouping principle. The percept is of a continuous noise corresponding to the narrowband parts of the stimulus, with an additional high-frequency bandpass component periodically pulsing; that is, one continuous sound and one periodic sound.

This stimulus is perceived to contain two auditory images. The first is a low-passed noise corresponding to the low-frequency part of the signal that is always present. The second is a periodically-pulsed high-passed noise corresponding to the part of the signal that is only present during the wideband noise segments. The perceptual system seems to interpret the stimulus as consisting of an "old" ongoing, low-passed part to which is periodically added a "new" high-passed part.

The stimulus was synthesized with a digital synthesis program in SAOL at a sampling rate of 24000 Hz. Each wideband noise was created as uniform random noise, with sample values in the interval [-0.5, 0.5], where 1.0 is the maximum 32-bit sample value. Each low-passed noise was created by filtering a similar wideband noise with a 5^{th}-order digital elliptic filter. This filter, designed in Matlab, had a cutoff frequency of 2000 Hz and -40dB of rejection in the stopband. The overall stimulus was assembled from alternating 250 ms bursts of the wideband and low-passed noises, gated using a rectangular window. It can be heard as Sound Example 5-3 on my WWW page, and the SAOL code is given in Appendix B.

The channel-image assignment produced by the present model for this stimulus is shown in Figure 5-19. As can be seen in the figure, the perceived segmentation does not precisely match the human percept. Although immediately at the onset of the broadband sound, the high-frequency region is segregated as a separate image, once the amplitude modulation in this region ends and the 'steady-state' portion of the high-frequency range begins, there is no longer a basis in this theory for maintaining two images (since both images would be equivalently static).

Discussion

The segmentation difficulty here most likely stems from a mismatch between human time constancy for segmentation, and the model's time constancy. Imagine a stimulus of the same sort, except that it is much longer in duration. That is, rather than wideband noise and narrowband noise alternating every 250 ms, they alternate every 2 sec or 5 sec (Sound Example 5-4). In such a case, the human segmentation is much more like the segmentation shown in Figure 5-19.

Immediately after the wideband noise onset, the human listener hears a second auditory image in the sound. But after some small amount of time, since this second image is not doing anything to maintain its coherence as a separate object, it goes away and is forgotten. Only when the alternation is sufficiently quick does the *entire* segment of "extra" high-frequency noise become perceived as an image in its own right. (In fact, from listening to a series of these stimuli, it is clear that there is a sort of pulsation threshold governing the perceptual segregation of the high-frequency energy. I don't believe this has been formally studied.

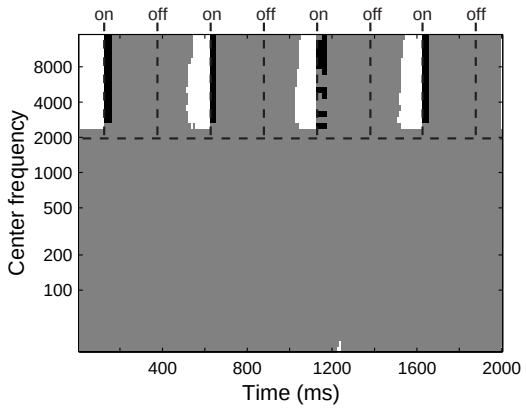


Figure 5-19**:** Model output for the alternating-noise stimulus. The 'on' and 'off' markers show the veridical onsets and offsets of the broadband noise. From 'on' to 'off', the input sound is broadband, and from 'off' to 'on' it is lowpass. However, there is still sound energy on the basilar membrane in the high-frequency regions during the lowpass intervals due to ringing of the auditory filters. The model segmentation is incomplete in this case; the second object (shown in black) lasts only a short time before vanishing.

Although it cannot be seen from this picture directly, the problem is in the second stage of the model, not the third stage. At the second stage, the model decides that only one image is needed to explain the modulation data at about 400 ms, about 900 ms, and about 1400 ms. The ease of explaining the modulation data with one image outweighs the cost of changing the number of images. Once the second-stage model makes this decision, then the third stage is trivial since there is only one image to which channels may be assigned.

If this explanation is correct, then the only change that would be needed in the model would be to reduce the cost of maintaining multiple objects over time even if they are not strictly needed to explain the modulation data, or to increase the cost of changing the number of images. Then the dynamic-programming optimization would be able to segment the sound properly.

Another simple modification to the model that might enable it to segment this sound in a way similar to the human listener would be the following. Presently, in the birth and death clauses in Eq. (5-21), there is no special consideration of the case in which a cochlear channel suddenly gains enough power to cross the low-power threshold, or falls below it and is removed from consideration. Intuitively, if a new cluster is to be born, it is better if the new cluster contains many channels that were previously unused—that is, that have just begun to contain energy. Similarly, when a old cluster dies, it is better if the channels that were in that cluster no longer contain any energy at all.

The exploration of such heuristics is left as a topic for future work.

5.4.4. Comodulation release from masking

The spectral-temporal phenomenon known as comodulation release from masking (CMR) was reviewed in Chapter 2, Section 2.1.3. I will describe the use of simple CMR stimuli to evaluate the grouping model.

Tone-in-noise stimuli were generated using the method that Hall *et. al.* (1984) described for their Experiment II. *Transposed coherent* (TC) noise was created by amplitude-modulating a six-tone set (1050-, 1150-, 1250-, 1350-, 1450-, and 1550-Hz tones) with a narrowband noise modulator, 100 Hz wide, centered at 350 Hz. This results in a signal with coherent 100-Hz-wide bands of noise centered every 100 Hz from 700 to 2000 Hz; that is, each band of noise has the same within-band envelope. Then, sound above 1350 Hz was filtered out to leave noise bands spanning the region from 650 to 1350 Hz.

Stimulus	Noise type	SNR
S1	TR	16 dB
S2	TC	16 dB
S3	TR	22 dB
S4	TC	22 dB
S5	TR	27 dB
S6	TC	27 dB

Table 5-2: Stimulus conditions (for comodulation-masking-release stimuli) used in testing the model

Transposed random (TR) noise was created using the same method, except that independent noise bands were used as the modulators for each tone. The only difference between the TC noise and the TR noise is that the six bands of noise are coherent in TC noise and incoherent in TR noise; comparisons of individual bands between the two signals are statistically indistinguishable. 1000 Hz tones were added at a variety of signal levels to noises at a fixed level. Hall *et al.* (1984) found a masking release of approximately 7 dB for the TC noise as compared to the TR noise.

All synthesis was performed digitally using a synthesizer written in SAOL. The filters for creating the 100-Hz-bandwidth modulators were each digital 6^{th}-order IIR elliptic filters designed in Matlab, one highpass with 300 Hz cutoff, and one lowpass with 400 Hz cutoff. These filters fall from the cutoff frequency to –50 dB rejection in approximately 120 Hz and are thus somewhat broader than the very steep analog filters used by Hall *et al*. (which had a transition bandwidth of only 14 Hz). However, the CMR effect is known to be quite robust to the effects of particular filtering methods and choices of bands. The initial synthesis was performed at a sampling rate of 8000 Hz (there is never spectral energy above 2200 Hz during the synthesis process, so there is no danger of aliasing), and then the resulting sounds were resampled to 24000 Hz for analysis. The SAOL code for the CMR stimuli appears in Appendix B, and the stimuli themselves can be heard on my WWW page as Sound Example 5-5.

Six stimuli were generated, as shown in Table 5-2. Each noise stimulus was 2.5 sec long, and the probe tone began after 1 s and lasted for 300 ms. Each noise was either TC or TR noise at a fixed in-band level of -36 dB relative to a full-power digital signal. The level of the probe tone was adjusted for the different stimuli, between 16 dB and 27 dB above the level of the noise.

The CMR stimuli were processed with the grouping model. The results are shown in Figure 5-20. As seen there, the model exhibits a CMR phenomenon similar to the human results on stimuli of this sort, although the overall masking threshold is higher for the model. In the model, randomly-modulated noise in the TR condition masks the tone at SNR 22 dB, but coherently-modulated noise does not. Both noises mask the tone at SNR 14 dB, and neither do at SNR 27 dB. Thus, the model predicts that coherently modulated noise is less able to mask the probe than is incoherently modulated noise. This prediction matches the experimental findings of Hall *et. al.* (1984) with regard to this simple set of stimuli. The model also predicts that there is a different quality to the noise in the TC and TR conditions (since the noise is grouped all into one image only in the TC condition). Experimental reports suggest that subjects anecdotally report such a difference in quality, although to my knowledge it has never been formally tested.

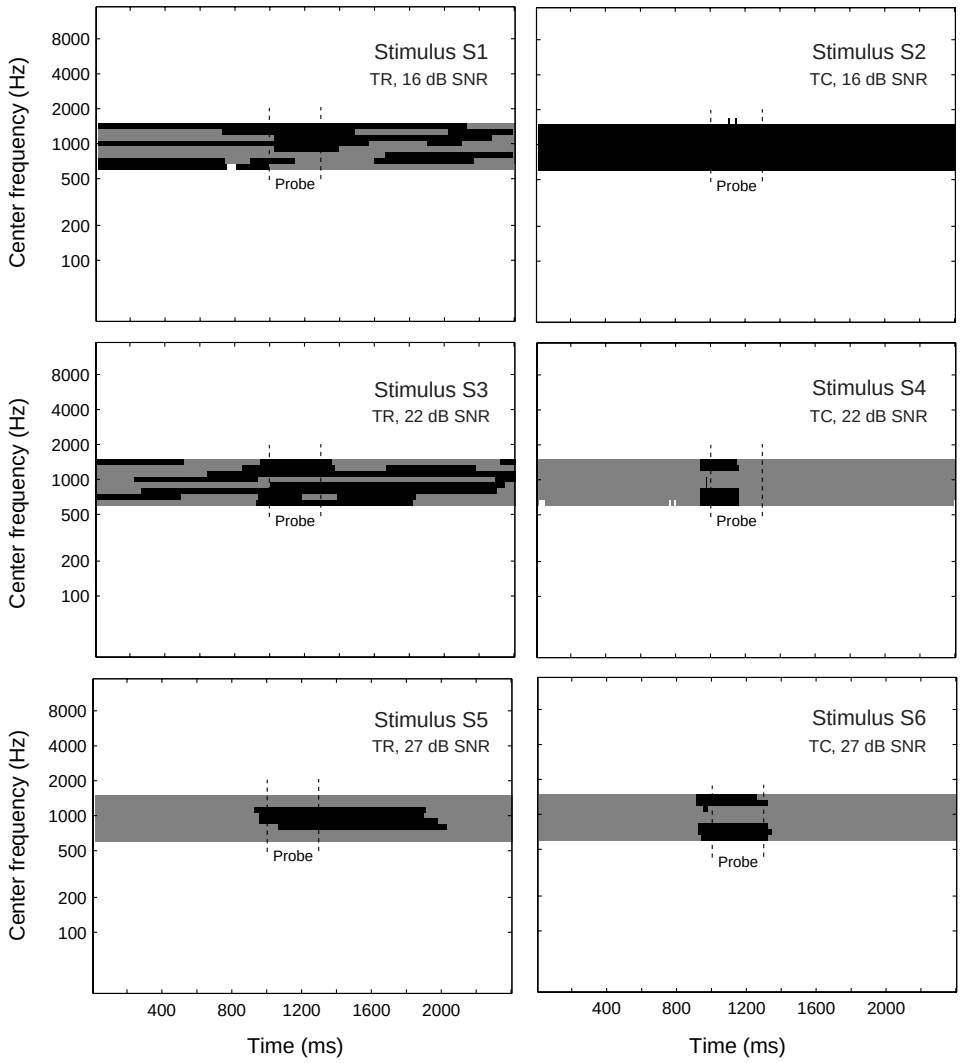


Figure 5-20: Model output for the comodulation-masking-release stimuli. For stimuli S1, S2, S3, and S5, no consistent second auditory image is formed. In stimulus S2, all channels are assigned to a single auditory image, which in stimuli S1, S3, and S5 the channels are randomly partitioned into two images that change over time. In stimuli S4, S5, and S6, auditory channels responding to the probe tone are coherently grouped into a different image than auditory channels that are not. The formation of a separate "tone" image occurs in the 22 dB SNR condition for TC noise, but only in the 27 dB SNR condition for TR noise. (It is only for the most-perceptible tone, in stimulus S6, that the duration of the tone is perceived veridically). The model predicts that coherently modulated noise is less able to mask the tone than is incoherently modulated noise, a prediction that matches the experiment findings of Hall *et. al.* (1984).

Discussion

We can look in more detail at the behavior of the model as it processes the CMR stimuli. This allows us to understand how the model exhibits CMR. The primary difference between TR and TC stimuli as viewed in the modulation space (as in Figure 5-8) is that the modulations of the various cochlear channels are coherent in the TC space and incoherent in the TR case. Most of the observed modulation is amplitude modulation; there is little period modulation in these stimuli since the sounds are noisy. That is, over time, all of the points in modulation space are moving up and down (along the amplitude modulation axis), *together* in the case of the TC noise, and *independently* in the case of the TR noise.

During the duration of the probe, it dominates some cochlear channels from time to time (depending on the moment-to-moment SNR in the channel). When this occurs, those channels exhibit neither period nor amplitude modulation, since in these regions of fixed-tone dominance the signal is stationary. In the case of TC noise, the channels that exhibit this lack of modulation are readily observed against the coherent background of comodulated noise. The nonmodulating channels are the only ones that are not doing the same thing as the larger group of background channels. But in the case of TR noise, it is more difficult to recognize that there is a group of fixed channels, since the background channels are modulating independently. Another way to put this is that the unmodulated tone is a modulation masker for some of the channels in the noise background, and that the detection of modulation masking is easier in the TC case than in the TR case.

As the SNR increases, the group of channels that is being dominated by the probe becomes larger (due to the spread of excitation of the probe) and the regions of dominance are more frequent. Thus, it becomes easier to see the static group of channels as a separate image from the background channels, whether the background channels are modulating coherently (TC noise) or independently (TR noise).

It is not the case, especially for the TC stimuli near the masking threshold, that the tone dominates the cochlear channels with CF near its frequency at each instant. The tone is only heard through the noise background at times of relatively low instantaneous noise energy (Buus (1985) called such a hypothesis a "dip-listening" model). The evidence integration in the clustering and sequence model allows the multiple glimpses of the tone to be gathered together as evidence for a single auditory image.

This explanation leads to a hypothesis about CMR; namely, that if the probe sound is amplitude-modulated in the same way as the masker, it should lead to a reduction in the masking release. That is, there should be less main effect of CMR with a modulated probe, since the modulated probe will not stand out against the background as does a static probe. There has been little study of CMR in the case of non-stationary probe signals.

It is to be noted that the probe levels used here are somewhat above threshold for a human listener. A human listener can easily hear the probe even in the highest SNR condition that I used. Hall and Grose (1984) found human thresholds to be approximately 13 dB SNR in the TC condition and 20 dB SNR in the TR condition. To better match the quantitative performance of humans on this task, the model would have to become more sensitive to the tone in the noise. This is a topic for future research; one immediate prospect is to increase the density of the cochlear filterbank, presently 6 per octave, to 12 or even 24 per octave. Oversampling the spectrum in this way would increase the amount of evidence that could be used to compare the modulating to non-modulating parts of the spectrum.

5.5. General discussion

By performing simple experiments with psychoacoustic stimuli, I demonstrated in the previous section that the model qualitatively performs as people do in a variety of tasks. This implies that the various perceptions of several different stimuli that are normally explained with different mechanisms can also be explained with a single modulation-detection mechanism. This result is attractive from the point of view of Occam's razor. If a single model can be used to explain multiple percepts, then this is a more parsimonious theory than one in which each sort of grouping percept requires a separate model. However, the predictions made by this model are not yet as accurate as those made by models developed for specific phenomena (Beauvois and Meddis, 1996, for example, for temporal-coherence stimuli). More research is needed to determine whether a model like this one can be used to produce quantitative predictions of human performance and thresholds in these tasks.

In this section, I will explore a variety of topics regarding implications of a model such as this for theoretical psychoacoustics. Particular attention is paid to the representational status of elements in this and other models in the literature; that is, what entities various models hypothesize as part of the perceptual representation involved in processing complex auditory scenes.

5.5.1. Complexity of the model

An immediate criticism that some might direct at the model in Section 5.2 is that it is very complicated. When the clustering model in Sections 5.2.4 and 5.2.5 is described mathematically, it seems extremely abstruse and unwieldy. This is particularly true in comparison to the simplicity and elegance of the pitch and loudness models that are available to present-day psychoacoustics.

There are two responses to this criticism. First, the complexity of the implementation of a model, or that of its mathematical description, should not be confused with the complexity of the conceptual underpinnings. The conceptual basis of the present model is quite simple. It is the same as that articulated by Duda *et al.* (1990): *auditory grouping is determined by common modulation in the autocorrelogram domain*. The method presented here for actually determining the common modulation patterns is rather messy, but it is possible that a simpler one could be found that abides by the same underlying principles. Even if this is not the case, there is a valuable distinction between the complexity of a particular theory and the complexity of a particular implementation of a model based on that theory that must be preserved (Marr, 1982).

Second, it is an invalid comparison to suggest that a model for auditory grouping and image formation should be as simple as those for pitch and loudness detection. From the range of experimental data on auditory scene analysis, it is clear that the auditory brain implements complicated processes. These give rise to a variety of complex behaviors. It is natural that complex problems with complex behaviors should require complex models to explain. The proper comparison of complexity is not models for other behaviors, or an abstract idea of how complex a model "should" be, but models that can produce equivalently satisfying results on the same range of stimuli. If two models can be shown to perform equivalently well (given the difficulties in evaluating performance of models of this sort, about which see below), then it is clear that the one that is simpler must be taken as embodying the better theory. But if there is only one model available, then there is no such comparison that can be made.

5.5.2. Comparison to other models

As discussed in Section 3.3.2, a number of previous approaches to the construction of computational auditory-scene analysis (CASA) systems have depended upon an initial stage in which the sound scene is analyzed in terms of sinusoidal *components.* In contrast, the present model assumes no components; rather, the basis for grouping is the direct output of the cochlear filters. As these outputs change, all of the energy in a filter channel is assigned first to one of the auditory images, and then to another. No intermediate representation is created, and thus, no intermediate representation need be defended.

A second important representational difference between the present model and others in the literature is that, in this model, pitch and harmonicity are not used as cues to grouping. In the component-based CASA systems just discussed, and in periodicity-based approaches to vowel separation, the harmonic relationships between components or subbands play a role of fundamental importance. This cue—called *harmonicity*—is most often incorporated using a residual-driven method: the strongest pitch in the stimulus is identified, and those components or subbands that correspond to it are removed from consideration. Then, considering only the residual, the strongest remaining pitch is identified, and so on.

Critical consideration of this model reveals several drawbacks. It is clear that a strict segregation by pitch does not always occur in listening to real-world signals. First, in nonattentive listening, varied groups of sources are fused into a single auditory image. Imagine an acoustic scene in which a listener sits in a room talking with friends while a radio (particularly a small one with a tinny and distorted loudspeaker) plays music. Even though the music consists of several instruments with multiple pitches, the sound coming from the radio is perceived only as a single image. The grouping cues given by the spatial location of the source and the transfer function of the loudspeaker override the pitch cues (if any).

Even in attentive contexts, such as music listening, there are many circumstances in which multiple-pitch stimuli are fused together. Notable here is the practice of homophonic writing in Western classical music, in which several instruments play sequences of notes with the same rhythm. The composer organizes the notes so as to encourage perceptual fusion of the chords. There are many sources, both in the music-pedagogy and music-perception (Sandell, 1995) literature that explore the roles of instrument timbre and harmonic relationships on the perceptual fusion effect. What is clear is that in many cases, chords are perceived as single elements rather than as multiple notes (Scheirer, 1996).

Some theorists argue that the percept in which multiple notes fuse into a single auditory image is itself best understood as a sort of pitch phenomenon. Terhardt (1982) extended the concept of "virtual pitch" (his term for the pitch of a complex tone with missing fundamental) to cover the perceived virtual root of a chord. Such a model can be used to make simple predictions about voice-leading and the role of chords within a harmonic context (Parncutt, 1997), but unfortunately makes other incorrect predictions about listening behavior in response to chords (Thomson, 1993). For example, this model predicts that in an operational pitch-matching task, listeners will match the whole chord to a "virtual subharmonic" that would be the common fundamental of all the constituent notes; however, experimental results do not bear this out.

The argument in this section is not intended to imply that pitch-based segregation of sounds never occurs; rather, the problem with existing models is that they do not include other cues, and so cannot make predictions about the cases in which other grouping cues are stronger. It is crucial that models of perceptual segregation predict human behavior in cases where segregation does not occur, as well as those in which it does. For example, a static double-vowel segregation model like de Cheveigné's (1998a; 1999)) seems to predict (although as far as I know this has not been tested) that multiple pitches should *always* be heard out from chords.

The strongest push towards residual-driven filterbank-allocation models has come from research into double-vowel stimuli. In these experiments, two synthesized vowels are mixed and a listener is asked to identify each. Independent variables that have been tested include differences in F_0, spectral shape, loudness, onset asynchrony, and many other features. It is clear that listeners have some ability to identify both vowels in such a mixture, but it is less clear that there are really two perceived constituents. An alternate hypothesis is that the vowel identification is actually revealing an ability to learn and report the combined, or fused, quality of the vowel pair. Under this hypothesis, no perceptual segregation actually takes place. Anecdotal evidence favors this view; listeners with no experience in double-vowel tasks generally cannot perform them "correctly." Rather, it is only after training, with feedback, that listeners develop the response characteristics usually reported for this task.

In an experiment that attempted to control for this possibility, Divenyi *et al.* (1997) found that dynamic cues (formant glides) increased segregation ability greatly. In fact, the results of Divenyi *et al.* indicated that little perceptual segregation occurred except in the case where dynamic cues were used. However, this experiment was only preliminary, and more research is needed on this topic.

The period-modulation cue used in the present model is not identical to the common-frequency-modulation approach to grouping suggested by McAdams (1984). Notably, period modulation as defined here is exhibited in signals for which no frequency modulation is present in the standard usage of that term. An example of this is the stimuli used to investigate the temporal coherence threshold (Section 5.4.2). The onsets in the sound, which have alternating pitch, give rise to period modulation in the output of the cochlear filters. This is true even though the sound has no components constructed via frequency modulation.

5.5.3. Comparison to auditory physiology

An important criterion for evaluation of a computational model of hearing is its connection to auditory physiology. In particular, it is crucial that models do not postulate computational elements that are not possible to implement as neural processing in the auditory pathway. That said, compared to the kinds of auditory processing that the human listener performs to understand complex auditory scenes, our present-day understanding of the neurological operation of the hearing system is very rudimentary. Because of this, there are few computational elements that can be ruled out as impossible. Further, as discussed by Marr (1982), there is much to learn from building computational models without necessarily drawing immediate connection to the perceptual physiology.

The connection of the first two stages of processing (cochlear filtering and hair-cell modeling) to the present understanding of the auditory physiology has been treated extensively in previous sources. Although the cochlear filterbank in reality has important non-linearities (in fact, a passive model may ultimately be unable to account for all the behavior of the cochlea), the results of many psychoacoustic experiments can be explained with a linear filterbank model such as the gammatone models developed by Patterson and his collaborators (1995). General references on auditory modeling such as Moore's book (1997, Ch. 3) explain the relationship between filterbank models and the present state of knowledge of the behavior of the cochlea.

A similar relationship holds between models of the inner-hair cells that are comprised only of a smoothing-and-rectification process (as used in Section 5.2.1), and the state of knowledge of the neuromechanical transduction process. Extensive stochastic models of the inner hair cells that take present understanding of the physical mechanics into better account are available in the literature. The relationship between the behavior of these sorts of models and the simple sort used here is well-understood.

My hypothesis in this dissertation is that important features of the early stages of hearing, including simultaneous segregation and sequential integration of the auditory scene, may be explained with simplistic models of the cochlea. However, this is only a hypothesis, and there may well be experimental results available in the future that demonstrate the need for models to include more sophisticated cochlear and hair-cell models. It is likely that more sophisticated front-end modeling would be necessary if the model is to make quantitative predictions of responses to auditory stimuli.

The status of the periodicity-detection step is more complex and problematic. There is little direct physiological evidence to support the notion of autocorrelation or other such direct periodicity-detection as part of the auditory pathway. Recent research on neural firings by Cariani (1996) can be taken as indirect evidence for the availability of a form of periodicity analysis; the statistical distribution of inter-spike-intervals from an ensemble of neural fibers (in cat) has been shown similar to the autocorrelation function of the rectified signal for many stimuli.

Other researchers have put forth models of periodicity detection and argued that they are more physiologically plausible. Notable among these is the Auditory Image Model (AIM) of Patterson and his collaborators (1992), which postulates an integrate-and-fire mechanism that "stabilizes" the auditory signal for periodic stimuli. Irino and Patterson (1996) have also

presented experiment data that suggest that certain perceptual behaviors (those having to do with short-term asymmetry in perception) are more readily explained with such a model than with autocorrelation. Slaney (1997) has presented a comparison of the implementation methods and predictions made by various subband-periodicity models.

With respect to the subsequent stages of the model, which are presented for the first time here, it is possible to imagine that modulation detection implemented with a set of difference detectors operating on the output of a band of modulation filters within each cochlear channel. Langner (1992) and others have provided evidence that suggests an important role in sound perception for the analysis of envelope modulation; these neural mechanisms might form the basis for the sort of modulation detection needed for this model. However, this argument must be taken as speculative at this time since there is little direct evidence available. The final stages of processing—clustering and channel-image assignment—seem like the sorts of operations that would occur centrally, in the auditory cortex. Little concrete is known about the kinds of representations or processing that occur cortically in humans or other animals.

Perhaps the best that can be said is that overall, where the physiology is well-understood, the model presented here follows general principles of auditory processing as determined from physiological evidence. For the parts of the model that purportedly correspond to less-understood parts of the physiology (which are most of them) at least there is no direct evidence suggesting that the proposed mechanisms are unlikely. It is likely to remain thus for some time.

5.5.4. The role of attention

An important aspect of sound perception that is not taken into account in this model is the role of the attentional set and goal of the listener. This is known to have important influences on even seemingly low-level aspects of hearing such as masking thresholds and the perception of loudness. It is of fundamental importance in the processing of some auditory-scene-analysis stimuli such as the van Noorden temporal coherence stimuli. The percepts of these sounds are phenomenally different when the listener is trying to integrate the percept than when he is trying to "hear out" separate parts of the percept. This difference can be objectively measured as differences in threshold or other behavioral characteristics. In the case of more complex stimuli such as music, listeners report that they are able to focus attention on one part of the sound (for example, the bass part), and thereby perceive more detail in that part at the expense of the other parts. There has been little objective study of this ability—it is unknown what mechanisms are being used, and whether they involve control over the physical mechanism of the ear, the way the sound is processed by the perceptual system, or both.

It seems likely that the particular thresholds of performance achieved in detailed psychophysical tests (such as the CMR experiments discussed in Section 5.4.4) represent the situation of focused attention, where the subject is committed to detecting the presence or absence of a particular test sound as accurately as possible. It is unknown in general whether these thresholds maintain in ecological listening, where the nature of the possible sounds that might be heard is not known *a priori* to the listener (as it is in a psychoacoustic experiment).

It is difficult to include aspects of hearing that are under volitional control in computational models. This is due to the difficulty of developing mechanisms to model the volition itself, as well as whatever effects it has on hearing. Critics of the traditional approach to the construction of artificial-intelligence systems use the term *AI-complete problem* to refer to a problem that requires a complete artificial intelligence in order to solve. Aspects of volitional control of hearing seem to have this nature, insofar as they require incorporating motivations, goals, and plans in order to model properly.

Until we can begin to develop models of volition and the way it relates to listening, the experimental paradigms that give different results (even consistently so) depending on the attentional set of the listener pose a difficult methodological problem for computer-modeling

research. Typically, it is considered proper when evaluating computer models not to alter the settings of the free parameters between experimental trials. For example, for the stimuli that have been run through the model in Section 5.4, all of the model parameters have been set as shown in the tables in Section 5.3.2 and not changed from one stimulus to the next. This is important, because otherwise there is a risk of the experimenter contaminating modeling results by using his own knowledge about the stimuli to improve the performance of the model.

However, in a circumstance in which certain free parameters of a model can be plausibly interpreted as related to the role of attention or volition, it seems more appropriate to *vary* them in order to evaluate their effect on model performance. For example, the parameter λ_m in the present model, which directly controls the minimum size of the clusters in the EM estimation procedure, indirectly controls the degree to which two groups of cochlear channels that are slightly separated in distance in modulation space are heard as separate images. If λ_m is set to a relatively large value, then the groups are likely to be heard as fused together; if λ_m is relatively small, then the two groups form separate perceptual images. This is exactly the kind of effect that is demanded by the results of van Noorden.

Based on this argument, it seems methodologically acceptable to vary the setting of λ_m based on an *a priori* theory connecting the attentional set to the λ_m value. But we are still left with the question of exactly how to choose the value and how to evaluate the model performance given a certain setting. In more rigorous quantitative psychophysical modeling than is presented here, one approach for setting the values of free parameters is to use the behavior on one experimental task to set values, and a different experimental task to evaluate the model. If the optimal settings for the independent task also give good results on the dependent task, it is likely that some underlying truth is revealed by the model.

This approach requires the multiple independent (and quantifiable) experiments treating similar independent variables and model parameters. In the case of less-well-understood independent variables such as attentional set, such experimental data are not yet available. Thus it is difficult at the present time to rigorously evaluate the performance of models that include free parameters corresponding to attention, volition, goal, and the like. This is an unfortunate conclusion, as it seems quite likely that the human processing of most complex sounds depends heavily on the role of attention. An engaged focus on attentional aspects of hearing is one of the most pressing experimental problems facing modern psychoacoustics.

5.5.5. Evaluation of performance for complex sound scenes

Trying to scientifically assess the performance of scene-segregation models on complex sounds is problematic, because it is extremely difficult to collect human experimental data on these aspects of hearing. Ellis (1996a), to take one example, performed a scene-analysis experiment with environmental stimuli in order to evaluate his model. He used a few, very complex, sounds, such as a recording of a construction site, and allowed listeners to respond freely in reporting the images (hammering, yelling, buzz-saws) they heard. He showed that his model could extract images that bore general similarity to the constituents reported by human listeners.

However, the particular model presented by Ellis was complicated and had many stages of processing. Only the topmost level (the "final list of images") could be evaluated in this way. Scientifically speaking, the material of predominant interest in Ellis' model lies in his mid-level representations, and it is exactly this level that is most poorly evaluated by a high-level listening task. In models posing attempts to model preconscious aspects of hearing, such as Ellis' and the present one, the inductive leap required to evaluate the internal representations and processing methods based only upon the final result of the model and a high-level listening experiment is problematic. For a complex model, a convincing argument must be

made that simpler models (or at least no *obvious* simpler model) cannot achieve similar performance on the task at hand.

Until psychoacoustic methods for studying the formation of auditory images in complex sound scenes become more mature, there seems to be no way to directly evaluate the representational claims made by a model such as the present one. Indirect evaluation seems to be the only method available. In Chapters 6 and 7, the results of the present model will be evaluated by extracting higher-level features that can be used to make semantic judgments about the musical source. Again, this is not a rigorous direct evaluation of the representational and processing claims themselves. It can only can provide indirect evidence for or against the model.

This is a disadvantage of the understanding-without-separation approach proposed in Chapter 3, Section 3.4. In non-perceptual signal-processing models where the goal is sound separation, there are a variety of engineering techniques for evaluation. For example, two known signals may be added together, and the model asked to extract each constituent from the mixture. The success of the model is judged by the similarity (using signal-to-noise ratio or some other criterion) of the extracted sounds to the original source sounds. This method of evaluation is inappropriate for perceptual models, because it is not always the case that the human listener will perceive both sounds in the mixture. Proper evaluation for perceptual models must focus on the representation of the scene by the human perceptual system.

5.6. Chapter summary and conclusions

As a review of this chapter, the most difficult and detailed of the dissertation, I present a summary of the major findings.

(1) A theory of the processing of complex sound scenes by human listeners has been presented. According to this theory, the formation of perceptual auditory images is governed by the discovery of groups of cochlear channels that are exhibiting common modulation behavior in the autocorrelogram domain.

(2) A computational model based on the processing theory has been implemented. The model can be used to examine the predictions the theory makes about the perceived segmentation of various sound stimuli.

(3) Several well-known auditory grouping and segmentation phenomena can be qualitatively explained by this theory, as demonstrated by the behavior of the computer model when applied to the stimuli that give rise to these phenomena in human listeners. These phenomena include grouping by common frequency modulation, the temporal coherence boundary of alternating tone sequences, segmentation of alternating narrowband and wideband noise, and a simple form of comodulation release from masking.

(4) The fact that a single theory can qualitatively explain these diverse behaviors is indirect evidence that they are all reflections of a fundamental comodulation-processing mechanism in the early auditory system.

(5) The present computational model cannot make accurate quantitative predictions of the performance of human listeners on these tasks, such as the particular thresholds of loudness and modulation that give rise to perceptual segregation and fusion. A different model based on the same theory could conceivably make quantitative predictions of this sort if more attention was paid to details of implementation in the cochlear filterbank and other components. This is a topic for future work.

(6) The theory is different in a number of respects from other models for computational auditory scene analysis that have been previously presented. Notably, nearly all of the perceptual processing occurs within-band, there are no mid-level “components” to maintain

an intermediate description of the auditory scene, and pitch is not used as a cue to perceptual grouping. The success of the model in qualitatively explaining the psychoacoustic effects shown here is a sufficiency proof that a pitch cue is not necessary to qualitatively explain these effects.

(7) The evaluation of computational models of the perception of complex sounds is very difficult. This is partly due to the complexity of human behaviors (including volitional aspects of perception) exhibited in response to complex sounds, and partly due to the lack of psychophysical experimental data on the perception of such sounds.

In the next chapter, I will show how this model can be applied to the analysis of real musical sounds, not only the sorts of simple psychoacoustic stimuli demonstrated here.

CHAPTER 6 MUSICAL FEATURES

In the previous chapter, I presented a model for the formation of auditory images based on the new principle of *autocorrelogram comodulation*. However, the model was only presented in the context of psychoacoustic test stimuli and was not evaluated on real musical examples. In this chapter, I will resume the discussion of real, ecological, musical sounds.

I will take as a starting point the models for tempo and image-formation presented in Chapters 4 and 5. First, I will show what happens when these models are applied to real music. Then I will discuss the difficulty of directly evaluating the results of doing so, and propose a different methodology for evaluation. Finally, I will present several simple features that can be extracted with a little post-processing on the outputs of the auditory-image model and the tempo-perception model. The features will not be used directly in this chapter, but will form the basis of larger models of music perception in Chapter 7.

6.1. Signal representations of real music

The model of the formation of auditory images was couched in Chapter 5 only in terms of its application to auditory test stimuli. These stimuli are important to discuss in the context of evaluating the model as a psychoacoustic theory, since most of our experimental data deal with them. The results in Chapter 5 are a preliminary indication that the principle of correlogram comodulation is a useful basis for a theory and model of auditory image formation.

However, as I discussed in Section 3.1.5, I am primarily interested in examining perceptions of real music, and in understanding the sensory principles on which these perceptions rest. It is essential that I examine the performance of these models on real music as well as on test signals.

Applying the models of Chapters 4 and 5 to real signals produces rich representations of the musical information that are difficult to interpret. Examples are shown in three figures. Figure 6-1 shows the simulated cochleagram—that is, the rectified outputs of the gammatone filters that stand in for the cochlear frequency decomposition in the early stages of the auditory model. As seen in this figure, real musical signals are noisy and full of complex

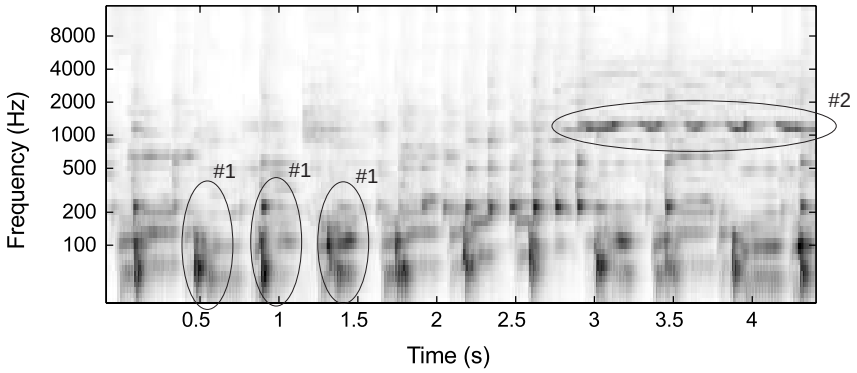


Figure 6-1: Cochleagram of 5 sec of an ecological music example (#24-1). Gray levels correspond to the amount of energy output in each filterband over time. It is difficult to see exactly what parts of the time-frequency energy correspond to perceived objects in the sound. Superimposed circles seem to correspond to at least part of the bass guitar sound (#1) and the harmonica sound (#2).

structure. It is difficult to reconcile the sounds that we hear[16] in the music with the visual information that we see in this representation.

A few things may be noted from observing the cochleagram while listening to this sound. There are four images that are immediately: the bass guitar, the snare-drum backbeat, the vocal "Ahhhh" at about 1.5 sec into the clip, and the harmonica sound that enters at the end. (Of course, it took me several listenings to arrive at this list, which begs the question of what my *initial* perception really was). Of these, only the bass guitar and the harmonica are clearly observed in the cochleagram as well-delineated individual entities.

The basis of most previous research on sound processing, and nearly all music-signal-processing research, has been a representation like this one (notable exceptions are Weintraub (1985) studying mixtures of speech, Leman (1995) on harmonic analysis, Ellis (1996a) on acoustic scenes, and Martin (1999) on sound-source identification). The sound is transformed into some spectrogram-like representation, and then grouping heuristics are used to extract regions of the time-frequency plane that belong together.

It has proven very difficult to make this approach work robustly when applied to complex ecological sounds. Of the literature I know, only the recent work of Goto (1999) attempts to apply time-frequency analysis to sounds as complex as this one. There are, in particular, four sorts of complexity that make time-frequency analysis of this sort of sound very difficult:

1. Many of the interesting aspects of the sounds are noisy. For example, the broad snare sounds and reverberation wash a lot of the image in general noise, and so it is difficult to recognize the vocalization (which is whispered, but immediately salient to the listener) as a *different* sort of noise signal.
2. The attacks of the instruments are not sharp, and therefore not well-localized in time. If we wish to assign each time-frequency cell to one object, we must make difficult decisions about when slowly-attacking instruments (like the harmonica) "really" begin to make sound.
3. Time-frequency uncertainty makes it difficult to choose a window size. If the goal is harmonic analysis, the filterbank must be narrowband enough that we can locate individual harmonics (which is not the case in Figure 6-1). But using a narrowband filterbank will exacerbate point (2) by smearing onsets in time.

[16] This sound example (which is #24-1), and the others used in Chapters 6 and 7, may be heard on my website at http://sound.media.mit.edu/~eds/thesis/.

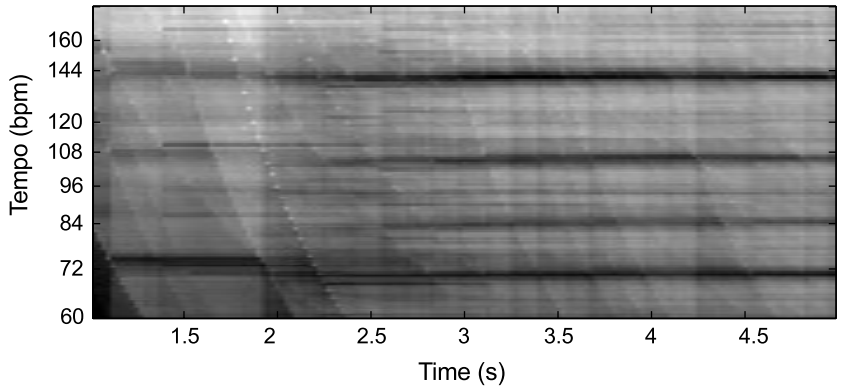


Figure 6-2: Tempo spectrogram of the same piece of music shown in the previous figure. The tempo spectrogram is one way of visualizing the output of the tempo model presented in Chapter 4. It is a sort of time-frequency plot, in which the graylevel at each tempo and time shows the amount of energy estimated at that tempo at that time. (The curves shown in Figure 4-7 are the instantaneous vertical cross-sections through a figure like this).

4. If the goal is to locate and pitch-track each individual instrument, harmonic relationships between the instruments make it difficult to decide which overtones correspond to which voice in the sound. When two harmonics (originating with the same or with different instruments) "collide" within a single filterband, the magnitude-filterbank representation has nonlinearities that are difficult to analyze (Mani, 1999 presents one attempt to do this).

The new processing models that I presented in Chapters 4 and 5 can also be used to produce visual representations of music for further analysis. For example, Figure 6-2 shows what might be termed a *tempo spectrogram* of the same piece of music.

The tempo spectrogram shows the amount of tempo energy[17] in the musical signal at each tempo at each time. Rather than visualizing the exact distribution of time-frequency energy as we can with the cochleagram (in which we try to see the visual correlates of auditory objects), the tempo spectrogram allows us to see the buildup of tempo over time in the signal. The different tempi might be interrelated to form a sense of rhythm in the signal. As discussed in Chapter 4, the phases that are not shown in this figure (which is only a *magnitude* tempo spectrogram) can be processed to find the beat in the signal as well as the tempo.

Similarly, Figure 6-3 shows the channel-image assignment produced by applying the Chapter 5 auditory-image-formation model to this sound. In this figure, each time-frequency cell receives a color indicating which auditory image it is assigned. (Although, as discussed in Section 5.2.8, I do not wish to consider these assignment functions *themselves* as the auditory images. The auditory images are better considered as underlying sound models, with parameters that are probabilistically updated based on evidence contained in the time-frequency energy assigned to them).

In this figure, we can observe that several of the local assignments seem to correspond to auditory images that we can hear in the sound. For example, the broadband snare drum sounds, the low-pass bass sounds, and the band-passed harmonica sound at the end all seem to have visible correlates in the image assignment function.

[17] Strictly speaking, *tempo energy* is a misnomer since tempo is a perceptual property while energy is a physical measurement. More properly, each cell in this plot shows the summed cross-band energy among the resonators tuned to a particular frequency at each time. The same point holds for terming the image as a whole a *tempo spectrogram.*

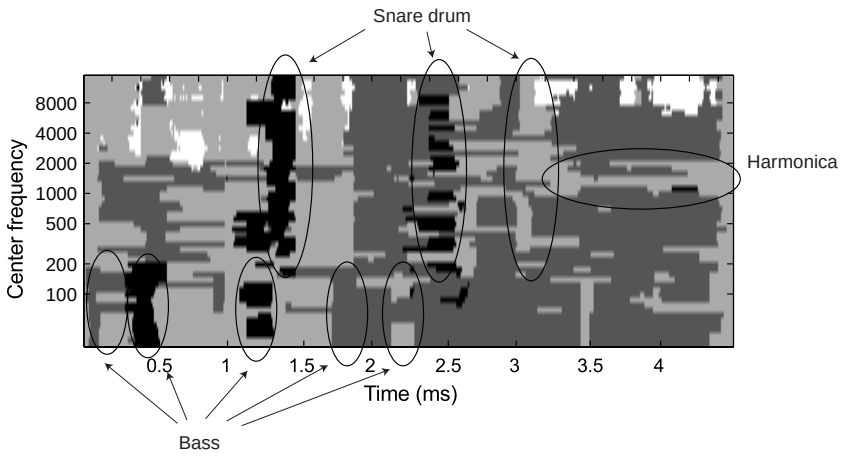


Figure 6-3: Channel-image assignment function produced as the output of the auditory-image-formation model in Chapter 5 to the same piece of music used in Figures 6-1 and 6-2. The graylevel indicates the auditory image that each cochlear channel is assigned to at each time. The set of channels sharing a color are grouped together and assigned to the same image. At the some times during the excerpt, the model uses three auditory images (colored dark gray, light gray, and black) to explain the modulation data, while at other times, it uses only two (colored dark gray and light gray). Time-frequency cells colored white have little energy relative to the other channels at that time and so are not assigned to any object. Some of the segmented time-frequency cells can be associated with sound events that are audible in the music.

We can also observe some of the difficulties of the present form of algorithm in analyzing complex sounds. Notably, there are sometimes still correspondence problems—at the beginning, the "background" auditory image is assigned the light gray time-frequency cells, but in the second half, there is a foreground-background switch, and now the *dark gray* time-frequency cells are assigned to the background. Just considering the first two low-frequency events, which likely correspond to bass notes in some way, the first is dark gray, and the second is black. Although the model recognizes these events in the sound, it cannot connect them as arising from the same source. In its present form, the model has no way of sequentially grouping objects together other than when this happens naturally from the dynamics of the cochlear filters, as in Section 5.4.2.

When we return to thinking about the perception of music as outlined in Chapter 3 and look at the visualizations in Figure 6-2 and Figure 6-3, we are confronted by an immediate question. Namely, are these plots correct or not? That is, since we are interested in the *perception* of sounds rather than the *separation* of sounds, is it really the case that a particular listener allocates the time-frequency cells to auditory images as shown in Figure 6-3? And is it really the case that a particular listener perceives different tempi as having different strengths as shown in Figure 6-2?

The analogous question is not typically raised for spectrogram-like representations as shown in Figure 6-1. This is because such a representation is easier to calculate (since it is just the magnitude output of a filterbank) and has a direct physical interpretation. Because of this, there is less concern about the nature of the processing algorithm itself. Further, there is good anatomical and neurophysiological justification for a filterbank-based representation. This is certainly not the case for the methods I have presented in Chapters 4 and 5. For these methods, it is a research question of some interest whether the representation is being calculated the right way, and whether they plausibly correspond to anatomical structures in the auditory system. However, once spectrogram data are allocated into a multiple-object segmentation as a perceptual model, we should ask these questions regarding the resulting auditory groups as well.

I must emphasize again that it is not the case that I am trying to build systems that can find all of the instruments in the musical mixture. It may be the case that a particular listener hears images corresponding to each of the instruments in the mixture, or it may not be the case. But at any rate, it is still an open question how to experimentally determine the actual set of images that a listener hears in a sound stimulus. And until we can determine the listener's perception of such sounds, we have no basis whatsoever for evaluating potential models of scene perception directly.

6.2. Feature-based models of musical perceptions

Many readers, at this point, may consider my argument against segmentation-based evaluation to be a sort of cop-out. That is, that the real reason I choose not to evaluate segmentation performance is that I have failed in an attempt to separate sounds. In this section, I will argue against this view, and instead present an alternative view of the way to construct and evaluate music-listening systems that is more in line with the overall approach that I presented in Chapter 3.

As discussed in Sections 2.5 and 3.4, most previous research, both in musical signal processing and in music perception, has taken a transcription-driven approach. That is (the implicit argument goes), first we must build automatic polyphonic transcription systems and then we will be able to use the results to solve music-analysis problems of interest. In the study of perception, the analogous argument is that first the *human listener* performs a signal separation that is something like a polyphonic transcription, and then music understanding proceeds based upon the structural relationships in the transcribed analysis.

But in neither of these cases is it the scene segmentation itself that is the primary goal of analysis. Rather, the segmentation or transcription is a subgoal that forms part of a larger goal—to analyze the musical scene—or a subtheory that forms part of a larger theory of music understanding by humans. The crucial point is this: *if useful analyses can be obtained, or coherent theories of understanding formulated, that do not depend on transcription or sound separation, then for many purposes there is no need to attempt separation at all*.

In fact, this point holds true for nearly all the practical problems that have been considered in the sound-analysis literate. The most important and interesting engineering problems, such as automatic classification and retrieval, performance analysis, human-computer musical interaction, soundtrack indexing, and intelligent composing assistants, do not require transcription except insofar as it would be a useful means to an end. Perhaps the only problem that depends on transcription in an interesting way is that of forming structural descriptions of sounds for low-bit-rate encoding (Vercoe *et al.*, 1998). The overwhelming focus on separation and transcription as an appropriate to solve has been counterproductive from the point of view of the field at large. It is essential that, instead, we think in an application-driven or theory-driven manner when determining the utility or success of a particular system.

From the scientific point of view, there is no particular evidence that a transcription-like representation is maintained in the auditory brain during musical hearing. This lack persists despite the general and largely implicit assumption, as I discussed in detail in Section 3.4, that the musical score is a appropriate representation for the development of perceptual and cognitive models.

The remainder of this chapter, and the whole of Chapter 7, will be concerned with explicating and exploring an alternative model of musical perception—one that follows the understanding-without-separation approach outlined in Section 3.4. A schematic of the model is shown in Figure 6-4.

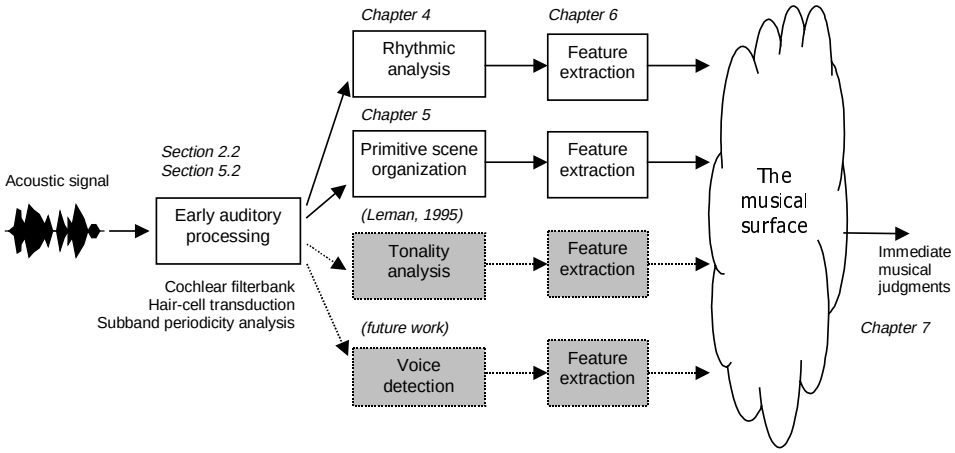


Figure 6-4: A model of musical perception built on the principle of understanding without separation. See text for details.

In this model, an acoustic signal is processed by the early auditory system as described in Sections 2.1 and 5.2. The neural sound representation is the basis of a rhythmic analysis subsystem (described in Chapter 4) and a subsystem that performs auditory-image formation and primitive scene segregation (Chapter 5). (Additional subsystems might perform tonality analysis (Leman, 1995) and detection of vocals, which are not discussed directly here). The outputs of the analysis subsystems are used as the basis of the extraction of musical features, as described in this chapter. The set of musical features extracted from the scene comprises the musical surface and is used by the listener to produce immediate judgments of the music, as will be discussed in Chapter 7.

This model interrelates all of the new results in the thesis, Chapters 4 through 7, and shows how to construct a computational model of the perception of music by human listeners that operates on ecological musical stimuli represented as sound signals. In this model, no notes or other reified entities of sound perception are used. Rather, the perceptions arise from the continuous transformation of the sound signal into surface features, and the surface features into judgments.

Evaluating this model allows indirect evaluation of the behavior of the tempo and auditory-image models. If the features that can be extracted from these models are sufficient to explain certain interesting musical judgments, then this model is consistent with the hypothesis that immediate musical judgments are made using such intermediate stages of processing. It is not the case, of course, that this demonstrates that these models are *necessary* to explain these judgments. This is a different question to which I will return in Section 7.2.3.

6.3. Feature extraction

To implement a feature-based model as shown in Figure 6-4, I have developed 16 features that can be extracted directly from the musical representations presented in Chapters 4 and 5. In this section, I will describe how each of them is extracted and show simple first-order statistics about their distribution. The statistics were collected from a musical test set containing 150 examples, each 5 sec long (more details on the collection of musical examples will be presented in Section 7.1.3, and Appendix A contains a complete list).

It is not the case that any of these features are supposed to directly correspond to interesting semantic judgments about real sounds. Rather, the features will be used in linear combination (and potentially, in the future, in nonlinear combination) as the basis for modeling the semantic judgments that humans can make. My goal in this section is simply to enumerate

many features that can be extracted with as little additional processing work as possible. For each feature, I have a brief story to explain why I think it might be a useful property of sound to measure, but the only real proof comes when we put these properties into action as part of a classification or regression model.

Further, it is not important that individual features be robust to noise or competing sounds. The only way that even very simple feature models, like pitch detectors, can be robust in this way is if we carefully bound and restrict the kinds of allowable noise and interference. To try to do this is pointless when we wish to consider analysis of every possible musical sound. A better and more general hypothesis, although this is only pursued in a simple fashion in Chapter 7, is that when one feature is unavailable or unreliable, others are used instead. Martin (1999) presented a more in-depth exploration of this idea and emphasized that deciding *whether* a feature is reliable for a particular sound is itself a very difficult problem.

Each feature is conceptually associated with a single point in time, and would vary over time in extended musical listening. For the example stimuli that I use here and the perceptual modeling in Chapter 7, the entire stimulus is taken as the analysis window for simplicity. To apply this approach to applications such as the automatic segmentation of music (see Section 7.7.2), more attention should be paid to windowing issues.

The features I will describe in this section fall into three categories: Features based on the auditory image configuration, features extracted from the tempo model, and features based on acoustic processing of individual auditory images.

6.3.1. Features based on auditory image configuration

Four features are extracted from the musical sound scene based on the configuration of the auditory images within the scene: the mean and variance of the number of auditory images, the average channel-by-channel modulation, and the spectral coherence of auditory images.

Mean number of auditory images

At each point in time, the auditory-image formation model of Chapter 5 determines how many auditory images best explain the modulation data at that time (for computational reasons, the model is presently restricted to using one, two, or three images). Thus, within any time window, it is possible to collect statistics on the distribution of the number of images found.

The *mean number of auditory images* is defined as

$$\mathrm{MEANIM} = \frac{\sum_{t=1}^{t_{max}} G_t}{t_{max}} \tag{6-1}$$

where t_{max} is the number of time frames analyzed, and G_t is the number of auditory images (clusters) determined by the image model at time t.

Figure 6-5(a) shows a histogram of this feature over the 150 sound examples in the test-stimulus database.

Variance of number of auditory images

The previous feature indicates how many images, on average, there are over time in the auditory scene. As well as the average, another useful feature is whether there are always the same number of images, or the number of images changes a lot over time. Thus, I calculate the *variance* of the number of auditory images. The variance is defined as the mean-square deviation from the mean of the number of images, calculated over all frames. That is,

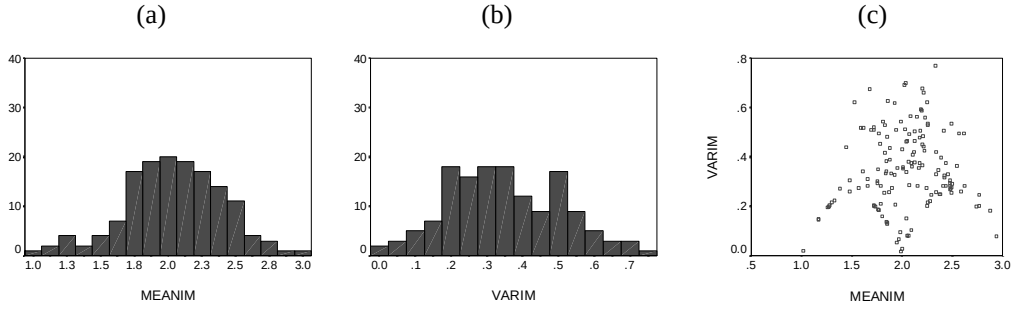


Figure 6-5: (a) Histogram of MEANIM, the mean number of auditory images extracted from a stimulus, over the test stimulus set. (b) Histogram of VARIM, the variance of the number of auditory images in a stimulus over time, for the test stimulus set. (c) Scatterplot of MEANIM vs. VARIM for the test stimulus set. The two features are uncorrelated with each another (r = .074, p = n.s.), although some nonlinear statistical dependency is evident.

$$\text{VARIM} = \frac{\sum_{t=1}^{t_{max}} (\text{MEANIM} - G_t)^2}{t_{max}} \tag{6-2}$$

Figure 6-5(b) shows a histogram of the distribution of this feature over the test-stimulus set. Further, Figure 6-5(c) shows a scatterplot of the mean number of images against the variance of the number of images for the examples in the test-stimulus set. As seen from the scatterplot, the two features are relatively independent of one another.

Mean channel modulation

In Section 5.2.2, modulation features were extracted from the sound stimulus as the initial stage of analysis. These features were used as the basis for the clustering model of auditory images in the sections that followed. However, the raw modulation features themselves are also potentially useful in support of semantic judgments.

Recall that for each cochlear channel at each time, the amplitude modulation and period modulation were extracted by analyzing the autocorrelogram. I define the *mean channel modulation* as

$$\text{MEANMOD} = \frac{\sum_{t=1}^{t_{\max}} \sum_{i=1}^{N} |p_i(t)| + |a_i(t)|}{t_{\max}} \tag{6-3}$$

where N is the number of cochlear channels, t_{max} is the total size of the stimulus (or the analysis window if these are not the same), and $p_i(t)$ and $a_i(t)$ are defined as in Eq. (5-9) and Eq. (5-11) respectively as the current period and amplitude modulation.

Other methods of extracting a similar feature might use the two modulation features jointly, for example, by taking the mean of the norms of the feature vectors $\mathbf{x}_i(t) = [p_i(t)\ a_i(t)]^T$. I haven't yet explored variants on most of the features presented in this chapter.

Figure 6-6(a) shows a histogram of this feature over the test set of stimuli.

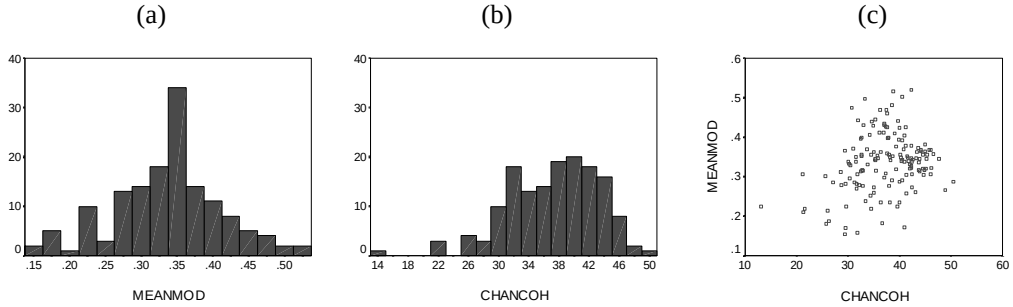


Figure 6-6: (a) Histogram of MEANMOD, the average modulation in each cochlear channel at each time, over the set of 150 musical examples. (b) Histogram of CHANCOH, the average number of channels assigned to the same auditory image as their lower neighbor. (c) Scatterplot of MEANMOD against CHANCOH for the test stimulus set. The two features are only weakly correlated (r = .270, p = .001).

Channel-allocation coherence

For some kinds of auditory scenes (particularly noisy ones), the auditory-image model of Chapter 5 partitions the spectrum into large subbands by placing all nearby channels in the same group. For other scenes (particularly ones with a simple harmonic structure), there is more overlap and alternation between channel allocation—that is, channels nearby in frequency are often assigned to different images. I call the degree to which nearby channels are assigned to the same image the coherence of channel allocation.

The *channel-allocation coherence* at each time t is defined as

$$C(t) = \sum_{i=1}^{N-1} \begin{cases} 0 & if\ B_{it} \neq B_{i+1t} \\ 1 & if\ B_{it} = B_{i+1t} \end{cases} \tag{6-4}$$

where B is the class-membership as defined in Eq. (5-25). The overall coherence feature CHANCOH is calculated as the mean of $C(t)$ over time.

A histogram of this feature is shown in Figure 6-6(b). Figure 6-6(c) shows a scatterplot of this feature against the previous one, mean channel modulation. As seen in the figure, the two features have only a slight correlation.

6.3.2. Tempo and beat features

Seven features are extracted from the tempo model that I presented in Chapter 4. Each of them is easy to compute as a single post-processing step on the output or intermediate processing of the tempo model. Only the final 2 sec of the signal is used for tempo feature analysis in each case, because the tempo model has a fairly slow startup time while the resonators lock into the various periods in the subbands. If the features were to be extracted from longer signals, all windows except the first could use the whole length of the signal within the window.

Best tempo

The best tempo is simply the one that is returned most often by the tempo-processing algorithm as the tempo of the signal. At each time frame, the algorithm produces an estimate of the energy at each tempo, as shown in Figure 4-7. Call this function $E_t(\tau)$, where t is the time frame and τ is taken from the range of tempi analyzed. At each time t, the instantaneous best tempo can be defined as

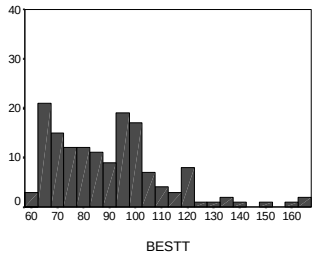


Figure 6-7: Histogram of BESTT, the tempo (in beats per minute) selected as best by the tempo-tracking algorithm in Chapter 4, for each of the stimuli in the test set.

$$\hat{\tau}_t = \sup_{\tau} E_t(\tau) \tag{6-5}$$

Then, over the range of *t* to be analyzed (the final two seconds of the signal, in this case), we select the $\hat{\tau}_t$ that occurs the most often as the *best tempo*. If more than one tempo occurs equally most often, we choose the best tempo as the one of the most-occurring tempi that occurred the most recently.

Figure 6-7 shows a histogram of the best-tempo values over the 150 test stimuli.

Mean tempo entropy

Within each frame of tempo estimates (as in Figure 4-7), we can consider the various amounts of energy assigned to each tempo[18] as a probability distribution governing the choice of a particular tempo. That is, the tempi with high energy are most likely to be chosen, and the tempi with least energy are least likely to be chosen. Then the entropy of each of these distributions tells us how much tempo information is present. If there are only one or two strong spikes (as in the top subplot of Figure 4-7), then little information is present and the tempo percept is very stable. If there is a broader distribution of tempo energy through the tempo spectrum, then the tempo percept is relatively unstable since there is a lot of information in the distribution.

The *mean tempo entropy* is defined as

$$\text{TEMPOENT} = \sum_{t=t_{\min}}^{t_{\max}} \sum_{\tau=\tau_{\min}}^{\tau_{\max}} -\hat{E}_t(\tau) \log \hat{E}_t(\tau) \tag{6-6}$$

where $\hat{E}_t(\tau)$ is the normalized tempo energy, that is

$$\hat{E}_t(\tau) = \frac{E_t(\tau)}{\sum_{\tau} E_t(\tau)} \tag{6-7}$$

Figure 6-8(a) shows a histogram of this feature over the 150 test examples in the stimulus database.

[18] See the footnote on page 153.

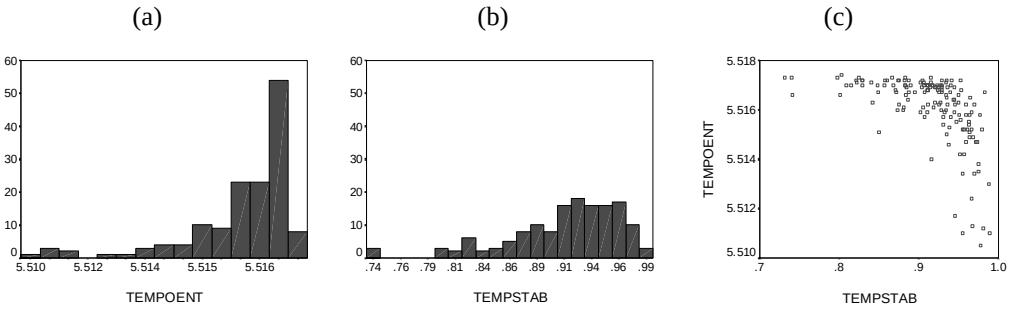


Figure 6-8: (a) Histogram of TEMPOENT, the mean entropy of the tempo energy distribution in each frame. (b) Histogram of TEMPSTAB, the correlation of tempo energy distribution at two points in time. (c) Scatterplot of TEMPOENT against TEMPSTAB for the 150 examples in the music database. There is a strong statistical relationship (r = -.527, p < .001) between these two features, showing that the more stable the estimate of tempo is, the more similar the tempo energy distributions at two different points in time are.

Tempo stability

For some genres and pieces of music (particularly rock and roll), the tempo changes very little over time. For other pieces (for example, *rubato* performances of Western classical music), the tempo may change a great deal in a short period of time. The tempo stability feature tries to measure the relative stability of the tempo energy over the fairly short duration of each of the stimuli. The normalized tempo energy spectrum $\hat{E}_t(\tau)$, defined in (6-7), is averaged within two time windows, each a half-second long. The first window begins 2.5 sec into the signal, and the second begins 4.5 sec into the signal. The correlation coefficient, a measure of vector similarity, is calculated between these two averaged normalized spectra.

The *tempo stability* of a stimulus is defined as

$$\frac{\sum\left[\hat{E}_1(\tau)-\overline{E}_1(\tau)\right]\left[\hat{E}_2(\tau)-\overline{E}_2(\tau)\right]}{\sqrt{\sum\left[\hat{E}_1(\tau)-\overline{E}_1(\tau)\right]^2\sum\left[\hat{E}_2(\tau)-\overline{E}_2(\tau)\right]^2}} \tag{6-8}$$

where

$$\overline{E}_1(\tau)=\frac{\sum_{t=t_1}^{t_2}\hat{E}_t(\tau)}{t_2-t_1},\ \overline{E}_2(\tau)=\frac{\sum_{t=t_3}^{t_4}\hat{E}_t(\tau)}{t_4-t_3} \tag{6-9}$$

with t_1, t_2, t_3, t_4 defined as given above.

Naturally, there are other methods available for computing various kinds of distances between the two averaged tempo energy spectra, for example simple Euclidean distance or Kullback-Leibler divergence (to continue treating the energy vectors as probability distributions). I haven't tested any of these methods yet.

Figure 6-8(b) shows a histogram of this feature over the database of 150 songs, and Figure 6-8(c) shows a scatterplot of the previous feature, tempo entropy, against this one. As can be seen from the scatterplot, there is a strong (although nonlinear) relationship between the two features.

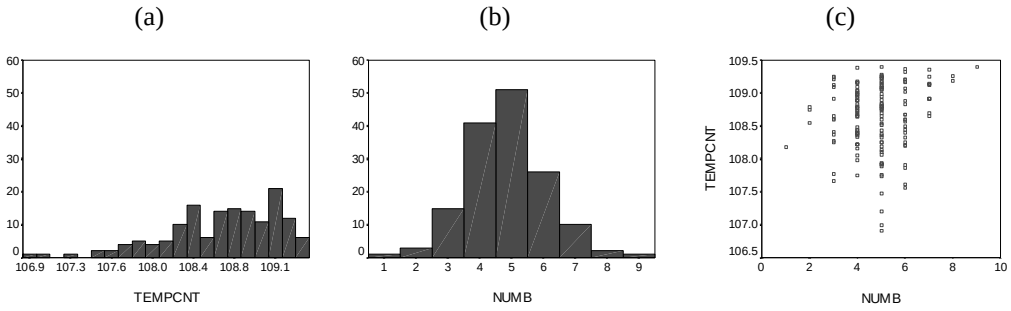


Figure 6-9: (a) Histogram of TEMPOCNT, the centroid of the tempo spectrum (in beats/minute), for the sample database of musical examples. (b) Histogram of NUMB, the number of beats produced by the beat-tracking algorithm for a 3 sec segment of the stimulus. (c) Scatterplot of TEMPOCNT against NUMB. NUMB is quantized since it can only take on integer values. There is no significant correlation between these two features (r = .143, p = n.s.).

Tempo centroid

As well as considering the distribution of energy to various tempi as a probability distribution, we can also consider it a sort of spectrum or weighting function. The centroid of this spectrum gives an estimate of the relative amount of periodic energy at high tempi in the signal compared to the amount at low tempi. This might correspond with a sense of "energeticness" for a musical signal.

The *tempo centroid* is defined as

$$\text{TEMPOCNT} = \sum_{\tau} T(\tau) E_2(\tau) \tag{6-10}$$

where $T(\tau)$ is the tempo that corresponds to each band of the periodicity-detection bank, and $E_2(\tau)$ is as given in (6-9).

Figure 6-9(a) shows a histogram of this feature over the test set of 150 musical examples. As seen there, even through the best-tempo estimates (Figure 6-7) vary widely, the tempo *centroids* are clustered in a narrow region.

Number of beats

As well as producing an estimate of the distribution of tempo energy in the signal, the beat-perception model in Chapter 4 produces a semi-rhythmic "tapping" output. These taps can be measured and transformed into features in their own right. The simplest of these is simply to count how many beats are produced by the algorithm. This is not simply the inverse of the best tempo, because for musical examples that have uncertain or changing tempo, the beats may be produced irregularly. Counting the number of beats also reflects the confidence of the algorithm in producing beats.

The *number of beats* for a stimulus is the number of beats generated by the beat-tracking algorithm after 2 sec of startup time until the end of the stimulus. A histogram of this feature is shown in Figure 6-9(b). A scatterplot of the number of beats against the tempo spectrum centroid can be seen in Figure 6-9(c), and shows that there is no linear correlation between the two features.

Mean interbeat interval

The features used in Section 4.4 to compare the beat-tracking model performance to the performance of human listeners can also be used as musical features of the sound. The first of these is the mean time between beats produced by the algorithm. This feature does not

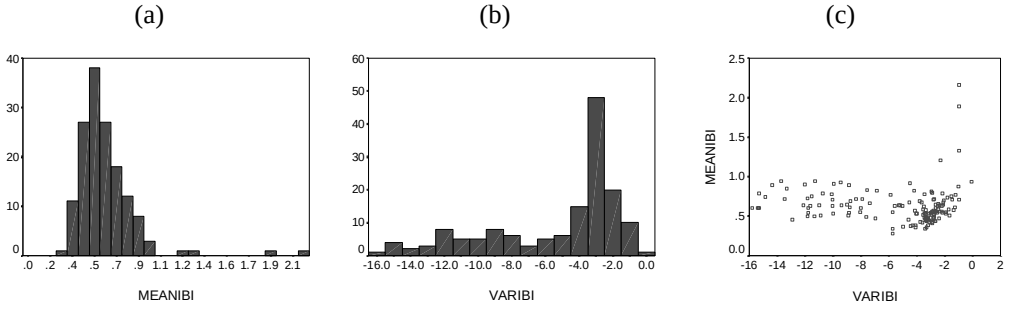


Figure 6-10: (a) Histogram of MEANIBI, the mean time (in sec) between beats. (b) Histogram of VARIBI, the log variance of the time (in sec) between beats. Very negative values (to the left) indicate that the beats were very regular; less negative values (to the right) indicate that the beats were very irregular. (c) Scatterplot of MEANIBI against VARIBI for the test set. There is a complex nonlinear relationship between the two variables, but no correlation (r = .094, p = n.s.).

exactly correspond the number of beats because of the possible edge effects of the limited window length. That is, suppose that on one stimulus the algorithm produces beats at t = [3.0, 3.5, 4.0]. There are three beats here with a mean interbeat interval of 0.5 sec. But if on another stimulus the algorithm produces beats at t = [2.75, 3.5, 4.25], there are still three beats, but with a mean interbeat interval of 0.75 sec.

Let β_i indicate the time of the i-th beat occurring after 2 sec of the signal. Then the *mean interbeat interval* is defined as

$$\text{MEANIBI} = \frac{\sum_{i=1}^{\text{NUMB}-1} \beta_{i+1} - \beta_i}{\text{NUMB}} \tag{6-11}$$

A histogram of the distribution of values of the mean interbeat interval for the set of 150 test stimuli is shown in Figure 6-10.

Variance of interbeat interval

The final feature that I extract from the beat-model output is the variance of the time between beats. In Section 4.4.2, this feature was used as one of the dependent variables comparing the performance of the algorithm to human performance in a tapping task. The interbeat interval variance is a measure of the regularity of tapping. If the beats are regularly spaced, the interbeat interval variance will be small; if they are irregularly spaced, it will be large. This feature is defined as

$$\text{VARIBI} = \log \frac{\sum_{i=1}^{\text{NUMB}-1} \left[(\beta_{i+1} - \beta_i) - \text{MEANIBI}\right]^2}{\text{NUMB} - 1} \tag{6-12}$$

The logarithmic transformation is used because the values can become very small if there are many very evenly-spaced beats. If there are not at least three beats for comparison, VARIBI is taken as missing data. A histogram of the distribution of VARIBI over the stimulus set is shown in Figure 6-10(b), and a scatterplot showing the values of VARIBI against the values of MEANIBI is in Figure 6-10(c). As seen in that figure, there is a complex relationship between these two features: at low values, there is a roughly linear correlation between the two variables, but at high values, the relationship is less clear.

6.3.3. Psychoacoustic features based on image segmentation

The previous two sections have presented features that can be easily extracted directly from the output of the image model and tempo model. In this section, I present five additional features that require slightly more post-processing. These features are based on looking at the features of the auditory images themselves, following the hypothesis that the allocation of time-frequency energy by the image model corresponds to the listener's sense (as discussed in Section 3.1.3) that there are multiple sources, each with its own properties, that can be used to explain the overall stimulus percept.

In a more sophisticated approach, the time-frequency energy allocated by the image model would be used as evidence for and against the existence of various sound-source models. Then, the properties of the source models would determine the perceptual features to be associated with each auditory image, as discussed in Section 5.2.8.

Here, I am not so sophisticated. I use two simple feature models to extract the psychoacoustic features of pitch and loudness from each of the auditory images according to the data allocated to it by the model in Chapter 5. Then I process these psychoacoustic features further to derive stimulus features like the other ones that I have already presented.

Pitch stability

The first stimulus feature extracted from psychoacoustic post-processing is the stability of pitches over time in each auditory object. This feature is partly a function of the nature of the stimulus, and partly a function of how well the auditory-image-segmentation model allocates data from different sources to different images.

The pitch of each auditory image is calculated according to the method of Meddis and Hewitt (1991), but at each time, only on those channels allocated to that image. That is, let B be the channel-image assignment defined in Eq. (5-25), and let $\mathbf{F}^t$ be the autocorrelogram frame at time t as defined in Eq. (5-6). Then let I_{kt} be the set of all cochlear channels at time t assigned to image k, that is

$$I_{kt} = \{i,\ 1 \le i \le N \mid B_{it} = k\}. \tag{6-13}$$

The *summary autocorrelation* of image k at time t is defined at each lag j as

$$\mathrm{SAC}_{kj}^{t} = \sum_{i \in I_{kt}} \mathbf{F}_{ij}^{t}, 1 \le j \le S_l \tag{6-14}$$

That is, the sum of the energy at that lag over each of the cochlear channels assigned to image k. The pitch is defined as the frequency corresponding to the lag at which the maximum of the summary autocorrelation is reached. The best lag is therefore defined as

$$\hat{j}_{kt} = \sup_{j} SAC_{kj}^{t} \tag{6-15}$$

and the pitch P_{kt} is the frequency (in Hz) corresponding to this lag. No further smoothing or post-processing is attempted. (This is the method that was used to estimate the pitch for the psychoacoustic tests in Chapter 5, Figure 5-14 and Figure 5-17.)

The stability of the pitch estimates over time is a measure of two things. First, if the sound stimulus is complex, then the pitch estimates will be messy as the image model allocates energy first to one source, then to another. Second, even the sound stimulus is relatively simple, then the pitch estimates will reflect the changes in pitch in the underlying sound sources, and will be as stable or instable as these source sounds were.

The stability of pitch estimates is defined simply as the average change in pitch estimate of each image from one frame to the next, averaged over both time and the number of images:

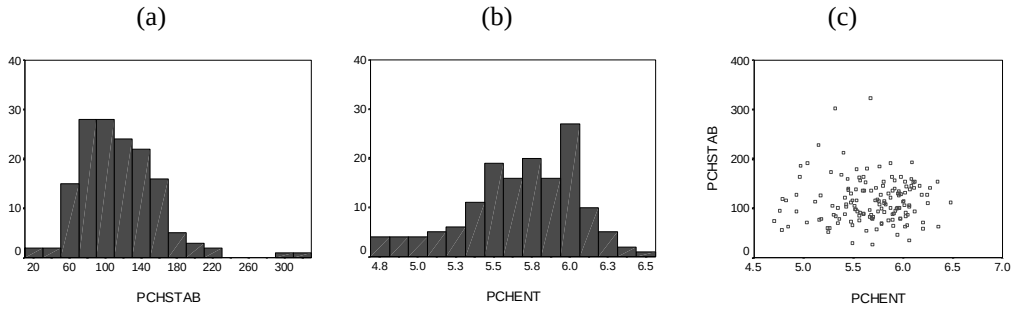


Figure 6-11: (a) Histogram of PCHSTAB, the average change (in Hz) in pitch of each auditory image in the sound scene from one time step to the next. (b) Histogram of PCHENT, the entropy of the distribution of pitch estimates over time and the various auditory images. (c) Scatterplot of PCHSTAB against PCHENT. There is no linear relationship between these features (r = .007, p = n.s.).

$$\text{PCHSTAB} = \frac{\sum_{t=1}^{t_{\max}} \frac{\sum_{k=1}^{G_t} \left| P_{kt} - P_{kt-1} \right|}{G_t}}{t_{\max}} \tag{6-16}$$

(Actually, since the number of images changes from time to time, the inner summation is only taken for those images that also existed at the previous time step).

A histogram showing the distribution of this feature over the set of 150 musical test stimuli is shown in Figure 6-11(a). As seen in this figure, the values are quite large—this likely indicates that the feature mostly reflects the changes in channel allocation and image selection from one moment to the next in the output of the image model.

Pitch entropy

Another metric that corresponds to the stability of pitch estimates over time is the entropy of the estimates. If all of the pitch estimates are very similar, then there will be low entropy in the overall estimation distribution. If they are very different, then the distribution has high entropy. In order to calculate this, we must consider the set of pitch estimates as a probability distribution, namely to compute

$$\hat{P}_{kt} = \frac{P_{kt}}{\sum_t \sum_k P_{kt}} \tag{6-17}$$

as the normalized pitch estimate.

The *entropy of pitch estimates* is defined as

$$\text{PCHENT} = -\sum_t \sum_k \hat{P}_{kt} \log \hat{P}_{kt} \tag{6-18}$$

A histogram of the entropies of pitch estimates for each of the 150 test stimuli is shown in Figure 6-11(b). Also, Figure 6-11(c) shows a scatterplot of this feature against the previous one. Interestingly, there is no correlation between these two features.

Loudness entropy

As well as estimating the pitch of the various auditory images as a basis for perceptual features, I estimate the loudness of the images. This is at once easier and more difficult than estimating pitch. It is more difficult because the only models that have been presented for

loudness analysis of full-bandwidth sounds are very complex, and we know little about how loudness is integrated across critical bands (Allen, 1999). As a result, I have used a simpler model that is really estimating cross-frequency power more than loudness—it doesn't incorporate known data on masking, spread of excitation, or temporal effects, for example. But this model is easy to calculate, and even given its limitations, I think that the features *derived* from it are similar to what they would be (in terms of inter- and intra-stimulus variances) if they were derived from a more accurate psychoacoustic loudness model instead.

Beginning with the same channel-image allocation I_{kt} as discussed above, I simply compute the total power in each spectral band assigned to each image. The total energy in all images is not simply the total energy in the signal because the auditory filterbank does not have the property called perfect reconstruction.

Let Q_{it} be the power in channel i at time t, as calculated by the zero-delay autocorrelation given by Eq. (5-5) where $\tau = 0$. Then the loudness of each auditory image k is defined as

$$\mathrm{LOUD}_{kt} = \sum_{i \in I_{kt}} 20 * \log Q_{it} \tag{6-19}$$

that is, the cross-band sum of the log power in each channel assigned to that image.

Based on the time-varying loudness of each image, we can compute various features of the stimulus. For example, as with pitch, I can compute the entropy of the distribution of loudness values over time. The *entropy of loudness estimates* is defined as

$$\mathrm{LOUDENT} = -\sum_t \sum_k \hat{L}_{kt} \log \hat{L}_{kt} \tag{6-20}$$

where $\hat{L}_{kt}$ is the normalized loudness, that is

$$\hat{L}_{kt} = \frac{\mathrm{LOUD}_{kt}}{\sum_t \sum_k \mathrm{LOUD}_{kt}} \tag{6-21}$$

A histogram of this feature for the 150 test stimuli is shown in Figure 6-12(a).

Dynamic range

Another feature that can be extracted from the loudness of the images is the overall dynamic range of the signal. This is a measure of whether or not the loudness changes suddenly over time. For each time t, the total loudness in all images is calculated:

$$\overline{L}_t = \sum_{k=1}^{G_t} \mathrm{LOUD}_{kt} \tag{6-22}$$

Then the *dynamic range* is defined as the greatest of the local differences in total loudness within short windows throughout the signal, thus

$$\mathrm{DYNRNG} = \max_t \left[\max_{\tau=t}^{t+20} \overline{L}_\tau - \min_{\tau=t}^{t+20} \overline{L}_\tau \right] \tag{6-23}$$

(The overall analysis, as discussed in Section 5.2, is performed at 100 frames/sec, so 20 frames corresponds to a 200 msec time window. I haven't tested any other frame sizes for calculating the dynamic range).

A histogram of this feature for the music in the test database is shown in Figure 6-12(b). Figure 6-12(c) shows a scatterplot of the dynamic range against the loudness entropy. As

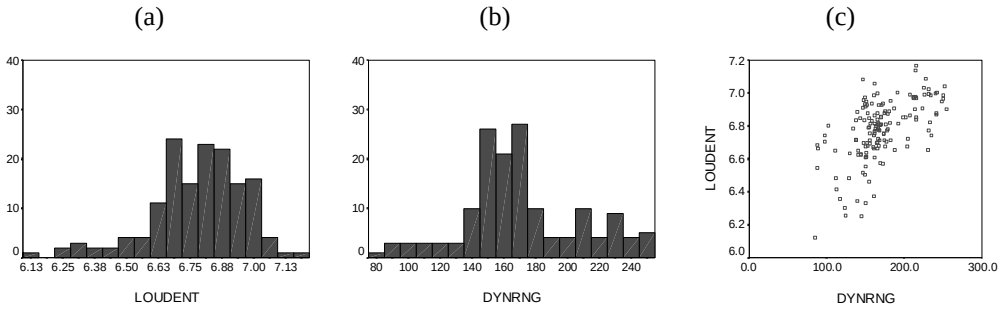


Figure 6-12: (a) Histogram of LOUDENT, the entropy of the distribution of loudness estimates over time and the various auditory images. (b) Histogram of DYNRNG, the dynamic range of the signal within local windows. (c) Scatterplot of LOUDENT against DYNRNG. There is a very strong linear relationship between these two features (r = .578, p < .001), showing that the more dynamic range a signal has, the more randomly-distributed are the loudness estimates from frame to frame in that signal.

seen in this figure, there is a strong linear relationship between these two features. Knowing the value of one feature explains 30% of the variance of the other, on average.

Loudest frame

The final feature that I extract from the loudness estimate, and in fact from the signal as a whole, is simply the loudness of the loudest frame in the signal according to the total loudness. Note that (as will be discussed in more depth in Section 7.1.3) a simple (although different) frame-by-frame psychoacoustic model is used to normalize all of the sounds to the same loudness, so in theory this measure should be the same for all pieces. The feature might actually be said to be measuring the differences between the two models of loudness.

The *loudest frame* is defined as

$$\text{LOUDEST} = \max_t \overline{L}_t \tag{6-24}$$

A histogram of this feature is shown in Figure 6-13

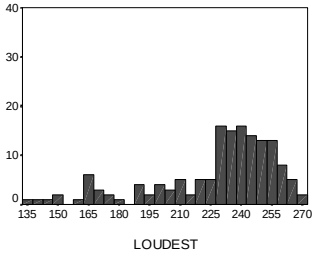


Figure 6-13: Histogram of LOUDEST, the individual loudest frame (according to a logarithmic scale) of each of the 150 test examples. The maximum possible loudness is 270, that is, 90 dB for each of three objects in a sound scene.

6.4. Feature interdependencies

As the various scatterplots of pairs of features have shown, the feature set described here is not completely orthogonal. That is, there are relationships, both linear and nonlinear, that make the various features statistically dependent on one another. This is both expected and natural, as there are only so many degrees of freedom in sound signals containing music.

It is possible to quantitatively analyze the feature set to determine the degree and potential effect of the statistical regularities. One method is by simple inspection of the interfeature correlation matrix, as shown in Table 6-1. As seen in this table, about half of the interfeature pairwise comparisons are statistically significant. This indicates that a few simple linear transforms are not sufficient to characterize the degrees of freedom in the feature space. Two of the interfeature pairs show very strong relationships. More than 95% of the variance of MEANIM can be explained by LOUDENT (or vice versa), and about 50% of the variance of TEMPOCNT can be explained by TEMPOENT, indicating that these pairs of features are essentially measuring the same thing.

Another test of interfeature correlation is to look for "island groups" of correlations, that is, subsets of features that have the property that each is statistically dependent on the other, and all are statistically independent from all of the features not in the group. There are no subsets of the features (other than the trivial one containing all of them) that have this property.

A more sophisticated way to look for dependencies among the features is to perform a principal-components analysis (also called a factor analysis) of their correlations. Doing this allows us to find a rotation of the feature space that still accounts for most of the variance among the features. If a low-dimensional subspace can be found that contains most of the interfeature variance, this is the same thing as showing that there are fewer degrees of freedom in the feature space than there are feature.

Figure 6-14 shows the fit obtained by basis rotations with subspaces using the best *k* eigenvectors. As seen in this figure, to explain most of the variance among the features requires a feature space with almost as many (at least half as many) dimensions as the original. This indicates that most features are bringing new information that cannot be explained as the linear combination of other features.

If the best *k* of the eigenvectors are chosen and used as a basis rotation for the overall feature space, the variance in the original features will be explained well for some features, and not as well for others. This is an indication of how well each feature is oriented with the principal eigenvectors (most important factors) in the basis rotation and is shown in Table 6-2.

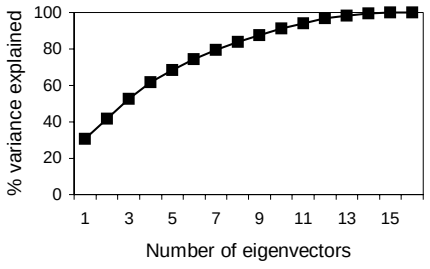


Figure 6-14: Amount of variance in the overall feature space explained by a rotated subspace with various numbers of basis vectors. If the features were mutually independent, the curve would be a straight diagonal line; if they were mutually perfectly correlated, it would be flat and horizontal. As seen in the plot, using only a few basic vectors fails to cover a lot of the variance in the space. Further, there is no clear knee in the curve that indicates a optimum tradeoff between the number of eigenvectors and the amount of variance explained.

	BESTT	DYNRNG	LOUDENT	LOUDEST	MEANIBI	MEANIM	MEANMOD	NUMB	PCHENT	PCHSTAB	CHANCOH	TEMPCNT	TEMPENT	TEMPSTB	VARIBI	VARIM
BESTT	1.000															
DYNRNG	.061	1.000														
LOUDENT	.133	**.578**	1.000													
LOUDEST	.013	**.440**	**.648**	1.000												
MEANIBI	-.090	.036	.061	.050	1.000											
MEANIM	.159	**.584**	**.986**	**.596**	.052	1.000										
MEANMOD	-.003	**.403**	**.636**	**.503**	.110	**.602**	1.000									
NUMB	**.237**	-.040	-.065	-.131	**-.464**	-.045	-.141	1.000								
PCHENT	-.020	.155	**.417**	.128	.062	**.426**	**.313**	-.063	1.000							
PCHSTAB	-.022	**.317**	**.342**	**.341**	-.025	**.336**	**.415**	.105	.007	1.000						
CHANCOH	-.046	**-.185**	**-.355**	-.064	.013	**-.413**	**.270**	-.137	**-.204**	.031	1.000					
TEMPCNT	.100	**-.267**	**-.380**	**-.341**	-.063	**-.414**	**-.350**	.143	.083	**-.229**	.195	1.000				
TEMPENT	**-.190**	**-.311**	**-.485**	**-.268**	.022	**-.550**	**-.267**	-.028	-.020	**-.217**	**.316**	**.760**	1.000			
TEMPSTB	.105	**.299**	**.376**	**.344**	-.014	**.394**	**.294**	-.014	.069	**.284**	-.060	**-.474**	**-.527**	1.000		
VARIBI	-.104	**-.278**	**-.298**	**-.208**	-.097	**-.322**	-.158	.138	-.087	**-.085**	.156	.241	**.409**	**-.251**	1.000	
VARIM	-.091	.099	.097	**.382**	.050	.074	-.116	**-.188**	-.079	-.129	-.147	-.360	-.249	.116	-.146	1.000

Table 6-1: Matrix of interfeature correlations. Each cell shows the Pearson's *r* correlation coefficient, calculated over the 150 test stimuli, between the two demarcated features. Each coefficient that is significant at $p < 0.05$ or less is marked in bold. That is, for the coefficients not marked in bold, there is a 1-in-20 chance or better that the given coefficient could simply arise through the random fluctuations of the features throughout their ranges. For the coefficients marked in bold, such an occurrence is unlikely, and so there is likely a linear relationship present between the features. 56 out of 105, or about 54%, of the pairwise interfeature correlations are significant.

Feature	Proportion of variance
MEANIM	.931
LOUDENT	.919
MEANMOD	.817
TEMPENT	.817
SPECSTB	.768
TEMPCNT	.750
NUMB	.741
VARIM	.698
LOUDEST	.655
MEANIBI	.642
PCHENT	.637
BESTT	.575
TEMPSTB	.543
PCHSTAB	.538
VARIBI	.471
DYNRNG	.457

Table 6-2: Feature extractions showing commonality between each feature and the 5-vector rotated basis determined by principle-components analysis (factor analysis) of the 16-dimensional feature space. Each row shows the proportion of variance in that feature that is explained by the lesser-dimensional space. The features that are best explained are the ones that are most closely aligned with the principle components extracted from the correlation data.

6.5. Chapter summary

This chapter has described a feature-extraction philosophy and methods that can be used to study real, ecological musical signals with models of psychoacoustics. I have mathematically derived 16 different features that can be easily extracted from the sound-perception models presented in previous chapters. Direct evaluation of sound-perception models is very difficult for such complex sounds, and so I argue that a better approach is to study them indirectly, by examining the utility of the features they support for building perceptual models of known judgments.

The next chapter therefore collects some data on human judgments in a new kind of music-psychology experiment, and demonstrates how the features presented here can be used to model the results.

CHAPTER 7 MUSICAL PERCEPTIONS

This is the final chapter that presents the results of new research. In it, I will incorporate results from the previous three chapters, by using them as the basis of a computational model of music perception that starts with acoustic signals. First, though, I will set the stage for the modeling research by presenting the results from two music-listening experiments with human subjects. These experiments treat the kinds of features of the musical surface that were discussed in Chapter 3, Section 3.2[19].

The organization of the chapter is as follows. In the first section, I will describe and present the results from an experiment that examines the abilities of human listeners to rapidly associate semantic judgments with short musical stimuli. I will discuss the results of the experiment in order to examine my definition of *musical surface* more critically. In the second section, I will show that a signal-processing model based on the three previous chapters can be used to predict the responses of listeners in this task.

In the third section, I will present the design and results of a second listening experiment; this experiment examines the perception of musical *similarity* by human listeners. The experiment uses the same stimuli as the one in Section 7.1, and an overlapping pool of listeners. I will then show that the results of the *first* experiment—and therefore also a signal-processing model based on the principles I have outlined previously—can be used to predict the results of the *second* experiment. That is, that a certain amount of the variance in the perception of musical similarity can be explained as the comparison of a small number of high-level semantic feature judgments operating on the musical surface.

Following the two main experiments, I will present the results of a short post-pilot third experiment investigating a methodological point. I will conclude the chapter with a discussion of the use of psychoacoustic and signal-processing models in the construction of multimedia-indexing systems and other practical applications.

[19] I am deeply indebted to Richard Watson, my undergraduate assistant, for his help building the computer interfaces, collecting and segmenting the musical stimuli, and running the subjects in these experiments.

7.1. Semantic features of short musical stimuli

This section describes an experiment that examines the ability of human listeners to make rapid judgments about brief musical stimuli. By hypothesis, the kinds of judgments that the listeners made in the experiment are those that they would naturally make in an ecological listening situation.

7.1.1. Overview of procedure

Thirty musically trained and untrained subjects listened to two five-second excerpts taken from each of 75 pieces of music. The subjects used a computer interface to listen to the stimuli and make judgments about them. The particular judgments elicited fell into two categories: *primary judgments*, which by hypothesis are direct reflections of immediate perception of the musical surface, and *secondary judgments*, which by hypothesis can be explained as stemming from some combination of the primary judgments. The primary judgments elicited were the degrees to which the musical stimulus was simple or complex, loud or soft, fast or slow, and soothing or harsh. The secondary judgments elicited were the degrees to which the musical stimulus was boring or interesting, and enjoyable or annoying.

7.1.2. Subjects

The subjects were drawn from the MIT community, recruited with posts on electronic and physical bulletin boards. Most (67%) were between 18 and 23 years of age, the rest ranged from 25 to 72 years. The median age was 21 years. Of the 30 subjects, 10 were male and 20 were female, although there were no gender-based differences hypothesized in this experiment. All but four subjects reported normal hearing. 22 reported that they were native speakers of English, and 6 reported that they were not.

9 subjects reported that they had absolute-pitch (AP) ability in response to the question "As far as you know, do you have perfect pitch?" No attempt was made to evaluate this ability, and it is not clear that all respondents understood the question. However, as reported below, there were small but significant differences on the experimental tasks between those who claimed AP and those who did not. The subjects had no consistent previous experience with musical or psychoacoustic listening tasks.

After completing the listening task, subjects were given a questionnaire regarding their musical background. No formal tests of audiology or musical competence were administered. The questionnaire allows classification of subjects based on their musical experience, as follows:

- Subjects that reported at least 10 years of private study on a musical instrument, ending no longer than 3 years ago, and at least two years of ear-training, music theory and/or composition study were classified as M2 subjects ($N = 3$).
- Subjects that were not classified as M2 subjects, but that reported at least 5 years of private study on an instrument, ending no longer than 5 years ago, or at least one year of ear-training and/or composition study were classified as M1 subjects ($N = 15$).
- Subjects not classified as either M1 or M2 subjects were classified as M0 subjects ($N = 12$).

Breakdowns of musical ability by age and by gender are shown in Table 7-1 and Table 7-2. Note that the experiment is not counterbalanced properly for the evaluation of consistent demographic differences.

	Male	Female
M0	1	11
M1	8	7
M2	1	2

Table 7-1: Cross-tabulation of subjects' musical ability by gender.

	18-25	25-30	30-40	40+
M0	9	0	2	1
M1	9	5	0	1
M2	2	0	1	0

Table 7-2: Cross-tabulation of subjects' musical ability by age.

7.1.3. Materials

The experimental stimuli were 5-second segments of music. Two non-overlapping segments were selected at random (based on their starting positions in time) from each of 75 musical compositions. The 75 source compositions were selected by randomly sampling the Internet music site MP3.com, which hosts a wide variety of musical performances in all musical styles by amateur and professional musicians. A complete list is given in Appendix A.

In their original form, some samples were recorded in mono and some in stereo, at a variety of sampling rates and peak power levels. To make the stimuli uniform, each was mixed down to mono by averaging the left and right channels, resampled to 24000 Hz, and amplitude-scaled such that the most powerful frame in the 5-second segment had power 10 dB below the full-power digital DC.

It is worthwhile to explore the implications of this method of selecting experimental materials. MP3.com is (as of this writing) the largest music web site on the Internet, containing about 400,000 freely-available songs by 30,000 different performing ensembles. Using of materials from such a site enables studies to more accurately reflect societal uses of music (by the segment of the population that listens to music on the Internet) than would selecting materials from my personal music collection. The materials are certainly more weighted toward rock-and-roll and less toward music in the "Western classical" style than is typical in music-psychology experiments. However, this weighting is only a reflection of the fact that the listening population is more interested in rock-and-roll than it is in "Western classical" music.

A second advantage of selecting music this way is that scientific principles may be used to choose the particular materials. In this case, since the set to be studied is a random sample of all the music on MP3.com, it follows from the sampling principle that the results I will show below are applicable to *all* of the music on MP3.com (within the limit of sampling variance, which is still large for such a small subset). This would not be the case if I simply selected pieces from a more limited collection to satisfy my own curiosity (or the demands of music theorists).

A third advantage is that the materials remain easily accessible to other researchers. Researchers who wish to replicate or extend my results can simply download the same compositions from the MP3.com site. This might not be so easy if I used materials from major-label records that, at some time in the future, went out of print. I have saved digital copies of the downloaded music files to guard against the event that they eventually become

unavailable from MP3.com (please contact me via electronic mail to receive digital copies of the songs excerpts for research purposes).

A fourth and final advantage is the relevance of the results to practical problems on the Internet. It is likely that as the research field of which this dissertation is a part begins to mature, the results will be most enthusiastically consumed by businesses trying to provide multimedia-indexing services to consumers or other businesses. Therefore, to use actual Internet music materials for study shows that that the systems I build are directly applicable to pressing problems in the real world today and in the future.

It is to be noted that many of the pieces of music selected (through random sampling) as stimulus materials are not "good." That is, in several cases, the performances or compositions that have been recorded, or the production quality of the recordings, do not meet the standard of quality that would be demanded by a major-label release. However, this does not diminish the quality of the *research results* I present. In fact, I believe quite the opposite: if we wish to build systems that can operate in the real musical world of the Internet (which is simply a microcosm of the musical world at large), they must be able to deal with "bad music" as well as "good music."

Real listeners are confronted with "bad music" (meaning simply music they don't like) every day. The aesthetic reactions thereby evoked are no less valid, or worthy of psychological study, simply because they are negative ones. It is a research question of the utmost interest to understand how it is that individual listeners decide for themselves what music is "good" and what is "bad," and what features they use to make this decision when they hear a new piece of music. Bad music is part of the world of music just as good music is; to understand the *complete* world of music, music researchers must be willing to move beyond the "classics" and examine a more complete spectrum of examples.

7.1.4. Detailed procedure

Subjects were seated in front of a computer terminal that presented the listening interface, as shown in Figure 7-1. The interface presented six sliders, each eliciting a different semantic judgment from the listener. The scales were labeled **simple—complex**, **slow—fast**, **loud—soft**, **interesting—boring**, and **enjoyable—annoying**. The subject was instructed that his task was to listen to short musical excerpts and report his judgments about them. It was emphasized to the subject that there are no correct answers on this task, and that the experiment was only designed to elicit his opinions. Three practice trials were used to familiarize the subject with the experimental procedure and to set the amplification at a comfortable listening level. The listening level was allowed to vary between subjects, but was held fixed for all experimental trials for a single subject.

Each of the 150 stimuli (75 musical excerpts x 2 stimuli/excerpt) were presented in a random order, different for each subject. When the subject clicked on the **Play** button, the current stimulus was presented. After the music completed, the subject moved the sliders as he felt appropriate to rate the qualities of the stimulus. The subject was allowed to freely replay the stimulus as many times as desired, and to make ratings in any order after any number of playings. When the subject felt that the current settings of the rating sliders reflected his perceptions accurately, he clicked the **Next** button to go on to the next trial. The sliders were recentered for each trial (Section 7.5 will present a brief experiment studying a possible artifact due to the slider-based interface).

The subjects were encouraged to proceed at their own pace, taking breaks whenever necessary. A typical subject took about 45 minutes to complete the listening task.

The six scales in the interface were positioned in random up-down orientations when the interface was designed. This is important because I hypothesized correlations between the scales—that is, a stimulus perceived to be "fast" seems more likely to be rated as "complex",

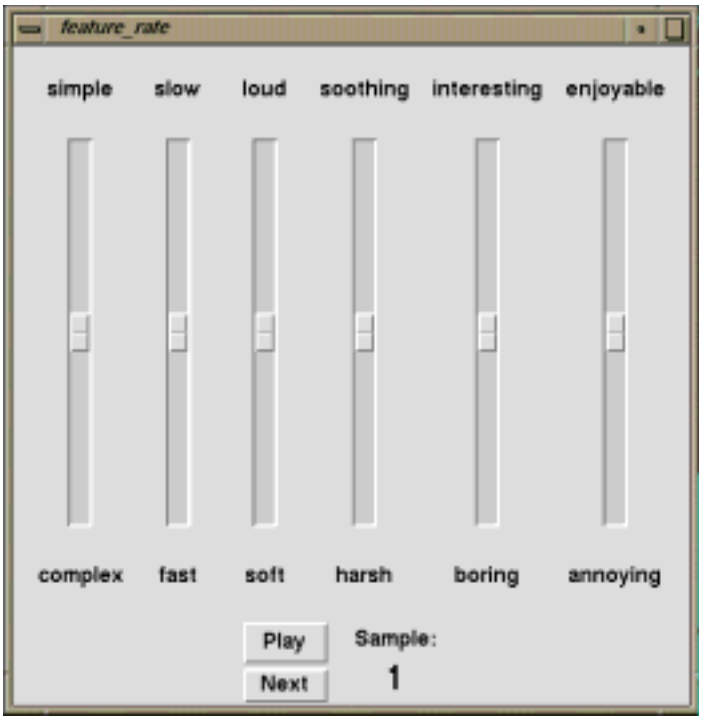


Figure 7-1: The computer interface for listening experiment I. Each subject rated each stimulus on 6 semantic scales.

"loud", and "interesting" than one rated "slow." If "fast", "complex", "loud", and "interesting" were all at the top of the scales, it might bias subjects' responses, since to move all the sliders in the same direction is a simpler response than moving the sliders in different directions. This bias would artificially enhance this correlation, with a danger of producing a false-positive result. By randomizing the orientation of the sliders, the danger of such a bias is eliminated.

7.1.5. Dependent measures

For each trial, the final settings of each slider were recorded to a computer file. The computer interface produced a value from 0 (the bottom of the slider) to 100 (the top) for each rating on each trial. For ease of description, I will refer to the dependent variables by their data-analysis names: SIMPLE, SLOW, LOUD, SOOTHING, INTEREST, and ENJOY. For each of these, a large value indicates that the subject set the slider towards the top of the corresponding scale, and a small value indicates that she set it towards the bottom. For example, a large value for SIMPLE indicates that the subject judged the stimulus to be simple, and a small value that she judged it to be complex. Any feature on which the slider was not moved at all (that is, for which the slider value was 50) was rejected and treated as missing data for that stimulus. This is to offset the response bias that originates in the initial setting of the slider (about which also see Experiment III, Section 7.5). Approximately 6.1% of the ratings were rejected on this basis.

As discussed below, each of the response variables were shifted to zero-mean and scaled by a cube-root function to improve the normality of distribution. After this transformation, the responses lie on a continuous scale in the interval [-3.68, +3.68]. Two additional dependent variables were derived for each response variable. The "sign" variables (SIMPSIGN, SLOWSIGN, etc) indicate only whether the responses on the corresponding sliders were below the center of the scale or above the center of the scale. The "offset" variables

(SIMPOFF, SLOWOFF, etc) indicate the magnitude of response deviation from the center of the scale on each trial, without regard to direction.

A note about loudness

One of the scales used in this experiment is labeled **soft** to **loud**. As discussed in Section 3.1.2, this exactly matches the psychophysical definition of loudness—that perceptual attribute that allows it to be positioned on the continuum from "soft" to "loud." Thus, by definition, this scale asks subjects to report the loudness of the stimuli.[20]

However, it is clear that loudness in this experiment means something different than loudness in most other psychophysical experiments. At the least, there are many more degrees of freedom in these stimuli than are typically included in psychoacoustic stimuli. It seems clear that subjects might use a variety of complex percepts, including the musical structure, genre, and style of the music, as well as physical stimulus parameters, in rating loudness. (This is in line with results presented by Fucci *et al*. (1993; 1996), who found that, other things equal, subjects' musical preferences affected their perception of the loudness of sounds).

There is nothing wrong with this. It simply highlights the difficulties that begin to confront us when we move beyond test stimuli to work with the sorts of stimuli that subjects might encounter in the real world. To the extent that loudness is a real, relevant property of sounds that subjects actually make use of in everyday judgments, we should *expect* that it will be a messy and noisy one. The only sorts of stimuli that afford us clean experimental judgments are those that eliminate the complexity of the real world.

As noted in Section 7.1.3, a simple power-based scaling was applied to normalize the signals to a roughly equal listening level. This may be taken as a simple form of loudness equalization; of course I could also use a more sophisticated model of loudness to do this scaling. But again by definition, if I could really normalize all of the stimuli according to their actual loudness, there would no longer be any sense in including loudness as a dependent variable since all stimuli would be equal. The fact that, empirically, they are not equal on this scale reveals simply that power-based equalization doesn't really capture all of the degrees of freedom used in loudness judgments.

If such factors as musical preference, structure, genre, and style really form part of the perception of loudness, as it seems they must, then to normalize for loudness must mean normalizing away the musical preference, structure, genre, and style. This is a very strange idea indeed!

For present purposes, the subject ratings presented below and the regression model presented in the next section can be seen as operating over the residual loudness left after an incomplete normalization. That is, we assume (by hypothesis) that the overall power level of the signal is going to be the most important factor in judging loudness, and so we remove this degree of freedom in order to highlight the other degrees of freedom more clearly.

7.1.6. Results

In this section, I will present the results of this experiment. Even outside of the overall context of developing psychoacoustic models, the results are of interest since, as far as I know, there have no few experiments like this reported in the literature to date.

[20] I am grateful to Malcolm Slaney for bringing this argument to my attention.

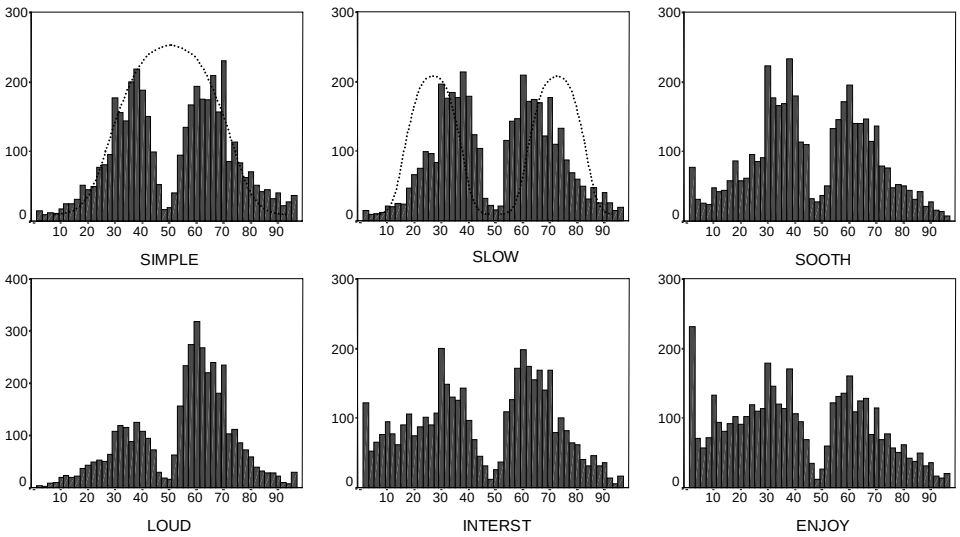


Figure 7-2: Histograms showing response distribution before rescaling, pooled for all subjects, for each of the six judgments elicited in Experiment 1. There is a characteristic shape shared by each distribution that may reflect response bias in the experimental interface. Also notice the floor effects present in ENJOY, INTERST, and SOOTH.

Response distribution

The distribution of responses in this task for each of the six features has a characteristic shape, as shown in Figure 7-2. These distributions can be modeled in two ways. First, as a normal response curve with a systematic bias away from the center (as suggested on the SIMPLE histogram). Second, as a bimodal response with a systematic bias (skew) towards the center (as suggested on the SLOW histogram). The second model was adopted for further investigation of the data. It is possible that the bimodal response pattern is an artifact of the experimental interface; this will be explored briefly in Experiment III (Section 7.5).

Since correlational study is an important part of the data-modeling approach I use, it is essential that second-order statistics be an accurate model of the data. A transformation was therefore applied to the raw results in order to remove the center-bias in the bimodal response. The center point was moved to 0 by subtracting 50 from each response, and then a nonlinear transformation (the cube-root function) was applied to each response. This maps the response scale to the range [-3.68,+3.68]. As seen in Figure 7-3, after such rescaling is applied, the data are well-modeled by a simple bimodal distribution.

All further results reported are based on the rescaled responses.

Learning and fatigue effects

Correlation (Pearson's *r*) tests were run in order to investigate relationships between the trial number (that is, the position of a particular stimulus in the random sequence of stimuli for a subject) and each dependent variable. These tests explore possible learning or fatigue effects.

There were no such effects for the primary judgments; none of the variables SIMPLE, SLOW, LOUD, or SOOTHING showed significant correlation with the trial number. On the other hand, both of the secondary judgments showed small but significant correlations with the trial number. For ENJOY, $r = -0.039$ ($p = 0.01$), and for INTEREST, $r = -0.071$ ($p < 0.001$). These results indicate that on average, as the experiment went along, stimuli were found to be less interesting and less enjoyable. This is clearly a fatigue effect. Fortunately, both effects

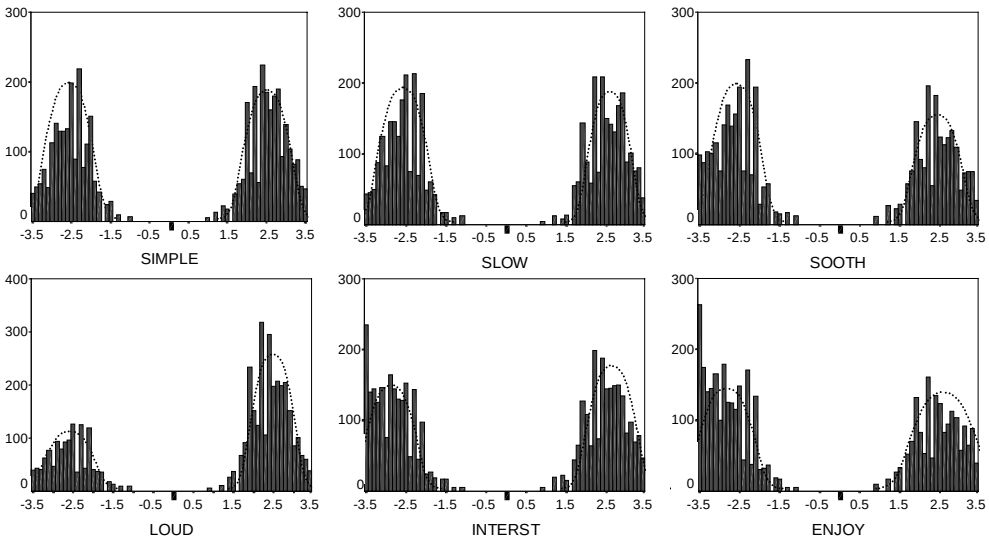


Figure 7-3: Histograms showing response data after centering and rescaling. These are the data used for further analysis. As seen by the imposed normal curves (fitted by eye, not by statistical analysis of the data), a bimodal distribution is a good model for each of the rescaled histograms.

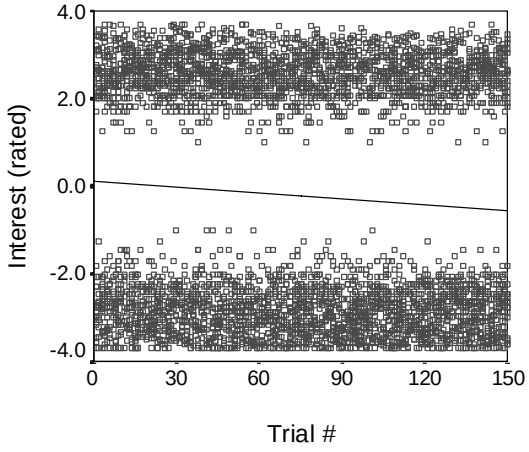


Figure 7-4: Scatterplot of trial vs. INTEREST, with imposed regression line. The correlation of these variables is significant but very small, as seen by the limited relationship of the data points to the regression line.

were very slight (the Pearson *r* is a sensitive test with so many cases), explaining only 0.2% and 0.5% of the variance in the dependent variables respectively.

To illustrate this, a scatterplot of INTEREST vs. trial is shown in Figure 7-4. As can be seen in the figure, the correlation is very slight.

Intersubject correlations

The pairwise intersubject correlations of responses to each feature judgment were calculated[21]. There are 435 intersubject pairs, so I will only present the results in summary. They are shown in Table 7-3. There are strong differences in intersubject correlation from

[21] This is not strictly proper, as the bimodal response patterns violate the assumption of normality in the correlation procedure.

Judgment	% correlations $p < 0.01$	Range of r	Mean r^2
SIMPLE	24.6 %	-0.204 – 0.525	0.035
LOUD	57.2 %	-0.203 – 0.761	0.097
SLOW	94.7 %	0.100 – 0.694	0.195
SOOTHING	95.6 %	0.089 – 0.766	0.232
INTEREST	12.9 %	-0.341 – 0.436	0.022
ENJOY	34.2 %	-0.245 – 0.566	0.049

Table 7-3: Intersubject stimulus-to-stimulus correlations on the bimodal response data. The second column shows the proportion of intersubject correlations (relative to the 435 possible pairs of subjects) that are significant at $p < 0.01$. The third column shows, for each judgment, the range of r (the correlation coefficient) over all pairs of subjects. The fourth column shows the mean of r^2 over all pairs of subjects, which indicates the average proportion of the variance in one subject's data that can be explained by the data of another subject selected at random.

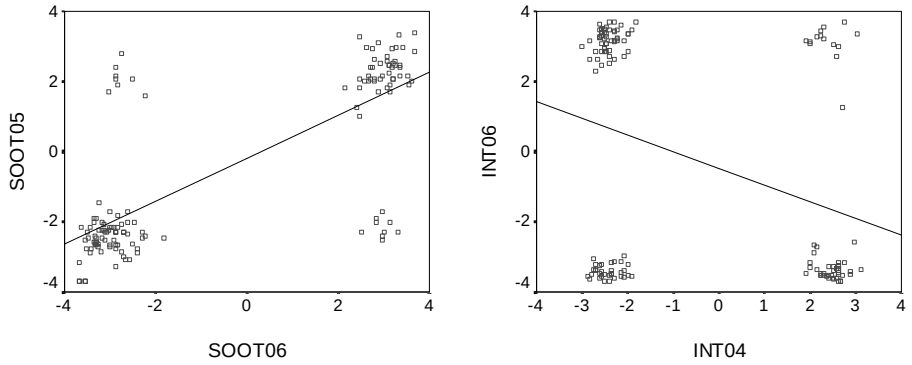


Figure 7-5: Scatterplots of intersubject correlations, with imposed regression lines, for the two most extreme cases. On the left, subjects 5 and 6 show very strong agreement (r = 0.766) on their judgment of SOOTHING. On the right, subjects 4 and 6 show significant disagreement (r = -0.341) on their judgment of INTEREST. Note the bimodal structure of these judgments, which has a strong bias on the correlation statistics.

variable to variable. For SLOW and SOOTHING, subjects generally agree on these ratings for the various musical examples. This is observed in several ways. First, for these features, nearly all of the pairs of subjects show significant correlation between their ratings. Second, even for those few pairs of subjects for which the intersubject correlation is not significant, the trend in the data is in the same direction—the r value is positive for all of the 435 pairs of subjects. For the best-matched pairs, more than 50% of the variance in one subject's data can be explained by the other subject's responses. Finally, on average, around 20% of the variance in one subject's rating of a piece (on these highly-intercorrelated judgments) can be explained by another, randomly-selected, subject's ratings of the same piece.

However, this is not the case for all features. For others, notably INTEREST, there is often disagreement among subjects. In this case, only 12.9% of the pairwise correlations are significant, the r values are sometimes negative and never get strongly positive, and on average, only a small proportion of the variance in one subject's data can be explained by randomly selecting another subject and examining his responses. Two extreme cases, one showing strong agreement and the other showing strong disagreement between subjects, are shown as scatterplots in Figure 7-5.

It is not possible to determine from these results whether the disagreement among subjects is due to different interpretations of the meaning of **interesting** and **boring**, different

psychoacoustic percepts that correspond to whatever **interesting** means, different "personal preferences" that govern what makes music interesting, or some combination of these three and other additional factors. However, it is clear that there are different types of responses among the feature scales used—some (SOOTHING, SLOW, and to some degree, LOUD) on which subjects generally agree, and some (INTEREST, SIMPLE, and ENJOY) on which they do not.

Analyses of variance

Analyses of variance were conducted to explore the relationships between the various independent and demographic variables and the subjects' responses. Due to the strongly bimodal nature of the responses, it violates the assumptions of the analysis of variance (which depends on second-order statistics only) to use the raw or rescaled responses. Instead, the two lobes of the distribution were collapsed by taking the absolute value of the rescaled response for each feature rating. I will refer to the new variables generated from this as *offset variables*. They measure the degree to which the subject moved the slider away from the center, in one direction or the other. Averaged for a subject, they indicate that subject's tendency to use the ends of the scale rather than the middle.

The existence of a significant covariate of an offset variable is a sufficient condition for there to be significant group-by-group response differences in the underlying bimodal variable. That is, we cannot improperly reject the null hypothesis of an analysis of variance (that all groups behave identically, drawn from the same probability distribution) because of this transformation. It may be the case that we fail to reject the null hypothesis in some cases in which it would be appropriate, but we will never reject it inappropriately.

The first set of analyses of variance explores the intersubject and interstimulus variances. The results are summarized in Table 7-4. As expected, there are significant effects in each of these analyses. This indicates that there are consistent differences between subjects—some subjects find all the stimuli to be simpler, and louder, and so forth than others do—and between stimuli—some stimuli are consistently rated as simpler, and louder, and so forth, than other stimuli.

The second analysis of variance explores possible dependencies of subject responses on the demographics of the subjects: musical ability, self-reported absolute pitch, native language (English or non-English), sex, and age. Age was segmented into four categories for analysis: 17-21, 22-29, 30-39, and 40+. The *p* values for all of these analyses are summarized in Table 7-5.

Dependent variable	$F_{SUBJ}(29)$	P(SUBJ)	$F_{STIM}(149)$	P(STIM)
SIMPLE	106.850	**0.000**	2.081	**0.000**
SLOW	76.684	**0.000**	2.671	**0.000**
LOUD	66.147	**0.000**	2.840	**0.000**
SOOTHING	57.125	**0.000**	5.340	**0.000**
INTEREST	64.306	**0.000**	1.525	**0.000**
ENJOY	51.990	**0.000**	3.148	**0.000**

Table 7-4: Summary of analyses of variance examining potential differences in offset response due to the stimulus and the subject. As expected, all of these analyses are strongly significant, indicating that the complexity, loudness, etc. differs from stimulus to stimulus, and that some subjects find all of the stimuli to be consistently simpler, louder, etc. than do other subjects.

Dependent variable	p(MUS)	p(AP)	P(ENG)	P(SEX)	p(AGE)
SIMPLE	**.001**	**.000**	**.002**	**.000**	**.000**
SLOW	**.001**	**.000**	**.000**	**.000**	**.000**
LOUD	**.032**	**.000**	.452	**.000**	**.000**
SOOTHING	**.000**	**.000**	**.000**	**.012**	**.000**
INTEREST	**.000**	**.000**	**.022**	**.000**	**.000**
ENJOY	**.000**	**.000**	**.**400	**.000**	**.000**

Table 7-5: *p* values from analyses of variance exploring dependencies of offset in subjective responses on the demographics of the subjects. The five demographic variables tested are: musical ability (MUS), self-reported absolute pitch (AP), native English language speaker (ENG), sex (SEX), and age, segmented into four categories (AGE). Only the *p* values are shown; most effects are significant (shown in boldface) for most demographic variables and most of the rating scales.

Again, most of the effects are significant here. This was unexpected and is somewhat difficult to interpret.

Figure 7-6 plots the overall mean rating offsets of the six judgments broken out by musical ability; Figure 7-7 shows them broken out by absolute-pitch group. As seen in the figures, the differences are consistent but very small. The differences are more evident in the ratings of INTEREST and ENJOY. For these judgments, skilled musicians and those listeners who claim to have absolute pitch[22] use the ends of the scale more, on average, than do less-skilled musicians and listeners who don't claim to have absolute pitch. The effects on the other ratings are difficult to summarize succinctly.

Since the absolute magnitude of the differences between the demographic groups was very small compared to the interstimulus differences, the subjects were pooled for further analysis, which focuses mainly on analysis of the stimulus-by-stimulus data.

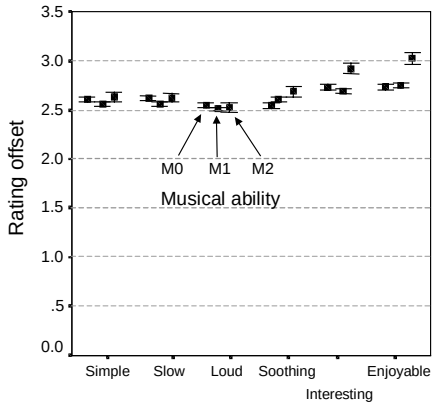


Figure 7-6: Mean and 95% confidence interval of offset ratings of the six semantic judgments, averaged across all stimuli. The y-axis shows the mean offset from the center, not the absolute mean position, for conformance with the ANOVA in Table 7-5. Each judgment is broken out by musical ability. For each judgment, there is a statistically significant (although obviously quite small) effect of the musical ability of the listener. The differences are difficult to interpret; perhaps the best that can be said is that highly-musically-trained (M2) listeners use the ends of the scale (that is, they have stronger opinions) more than do less-musically-trained (M0 and M1) listeners.

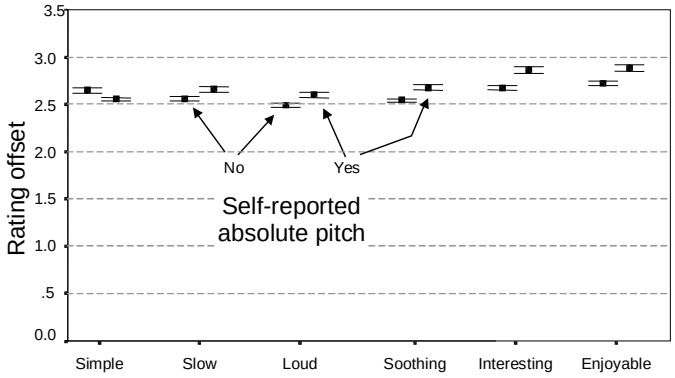


Figure 7-7: Mean and 95% confidence interval of offset ratings of the six semantic judgments, broken out by self-report of absolute-pitch ability. For each judgment, there is a statistically significant effect in whether the listener claims to have absolute pitch or not (claims were not verified). On five of the six scales, AP-claimers used the ends of the scale more than did non-AP-claimers.

Data modeling

As well as examining the relationship between the subject demographics and the observed responses, it is possible to learn about the perceptual behavior of subjects by examining the relationships *among* the different responses. I will do this in two ways in this section, first by presenting the intercorrelations among the semantic judgments, and then by performing a factor analysis of the results.

For these analyses, I will work with the mean rescaled ratings for each stimulus, averaged across all subjects. Although the rescaled subject-by-subject ratings are strongly bimodal (as shown in Figure 7-3), the mean ratings are not, as shown in Figure 7-8 (this is just a particular instance of the Law of Large Numbers). Thus, it is possible to use second-order-statistics techniques such as correlation and regression to analyze the mean ratings even though it is not proper to do so for the individual subject data.

The variable-to-variable correlations among the mean semantic judgments were analyzed, and the result is shown in Table 7-6. All except two of the pairwise correlations were significant,

	SIMPLE	SLOW	LOUD	SOOTHING	INTEREST
SLOW	**.521**				
LOUD	**-.479**	**-.741**			
SOOTHING	**.426**	**.715**	**-.882**		
INTEREST	**-.305**	**.**050	**-.168**	**.326**	
ENJOY	**.**097	**.320**	**-.506**	**.701**	**.750**

Table 7-6: Variable-to-variable correlations among the means of the perceptual judgments, calculated over the 150 musical examples (boldface indicates correlation is significant at $p < 0.05$). All but two correlations are strongly significant; variance explained ranges as high as 78% (LOUD and SOOTHING, in negative correlation).

[22] These were not the same group; of the absolute-pitch claimers, 4 were in group M0, 3 in group M1, and 2 in group M2.

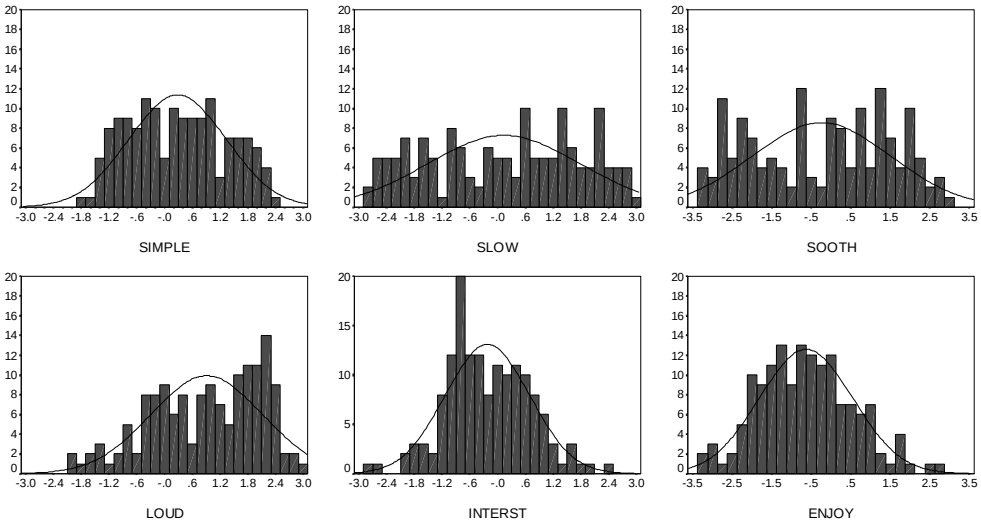


Figure 7-8: Histograms of mean across-subject rescaled ratings on the six semantic judgments for the 150 musical examples. As contrasted with the subject-by-subject ratings shown in Figure 7-3, the mean ratings are more nearly normally distributed. Imposed normal curves are fitted using the mean and variance of the data. Kolmogorov-Smirnov tests of normality reveal the following significance levels for rejecting the null hypothesis that the distributions are normal: SIMPLE, p = n.s.; SLOW, p = 0.011; SOOTHING, p = 0.001; LOUD, p = 0.001; INTEREST, p = 0.035; ENJOY, p = n.s. Thus, for two of the six judgments, the distribution can be considered normal, and for two others, the divergence from normality is mild.

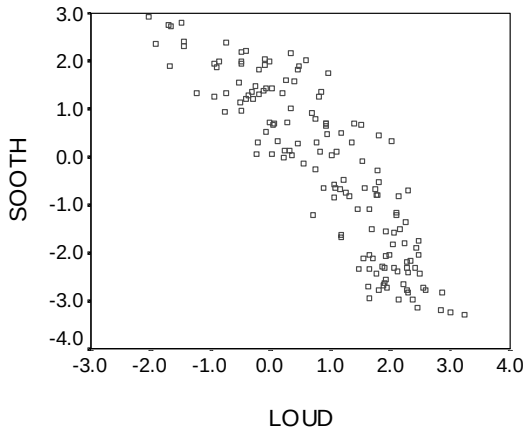


Figure 7-9: Scatterplot showing relationship between judgment of LOUD and SOOTHING. Each data point represents the mean rating taken over all subjects for one stimulus. The correlation between these two judgments is very strong: knowing the judgment on one scale explains 78% of the variance in the other.

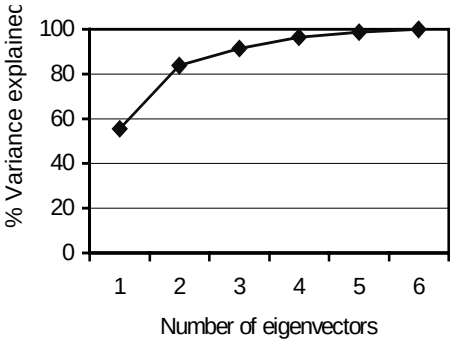
indicating that in almost all cases, the pairs of judgments are not independent. Some of these correlations are very strong—the correlation between SOOTHING and LOUD explains nearly 80% of the variance in these judgments. This indicates that the judgments of whether a piece of music is soothing or harsh, and whether it is loud or soft, are nearly the same judgement. The scatterplot of SOOTHING vs. LOUD is shown in Figure 7-9.

On one hand, it is important that not all of the results are strongly intercorrelated—this shows that there is more than one degree of freedom underlying the perceptual judgments. Previous research in the music-education literature (Madsen, 1996) had suggested that perhaps there is really only one judgment, of the nature "loud-fast-harsh-more," that underlies any semantic scale. But the fact that there are pairs of judgments that are uncorrelated is inconsistent with this hypothesis.

On the other hand, the fact that most interjudgment correlations are significant is an indication that subjects are not really expressing *six* degrees of freedom in their judgments. Thus, an interesting question is how many degrees of freedom are really present. The statistical technique called *factor analysis* can address this. The factor-analysis procedure analyzes the correlation matrix among the judgments, as shown in Table 7-6, considers it as a transformation in a high-dimensional space (6-dimensional in this case), and "rotates" the transformation so that the most important directions of variance are aligned with new *basis axes*.[23] If the judgments are really expressing only a few degrees of freedom, then only a few basis axes will suffice for explaining most of the correlation among the judgments. The results of a factor analysis of the mean ratings for the 150 musical examples is shown in Figure 7-10.

In Figure 7-10(a), a Scree plot of the factor analysis is shown. This plot shows how the proportion of variance explained increases as the number of eigenvectors (basis axes) increases. By definition, this proportion must be 100% when the number of eigenvectors is equal to the number of ratings. However, we can also observe that many fewer dimensions suffices to explain a large proportion of the variance. A two-dimensional basis space contains 84% of the covariance among the six judgments, and will be used for further analysis. Each additional dimension only explains a small amount of the residual variance.

(a)



(b)

Judgment	Corr. I	Corr. II	Var. exp.
SIMPLE	.519	-.670	71.8%
SLOW	.803	-.341	76.2%
LOUD	-.907	.172	85.3%
SOOTHING	.956	.002	91.4%
INTEREST	.384	.870	90.3%
ENJOY	.737	.582	88.2%

Figure 7-10: Factor analysis of the six perceptual judgments. (a) Scree plot, showing the proportion of covariance explained as a function of the number of eigenvectors used. Based on these data, a two-dimensional feature space that explains 84% of the overall variance is used for further analysis. (b) Table showing the partial correlation of each judgment to each of the two factors (I and II), and the proportion of variance in this feature explained by the two-dimensional reduced space. 72% of the variance in SIMPLE, 76% of the variance in SLOW, and so on, lies in the rotated two-dimensional plane spanned by the two extracted factors.

[23] For readers with a pattern-recognition background, *factor analysis* is just the term used by psychologists for principle-components analysis of the correlation matrix via the Karhunen-Loeve expansion. The resulting basis axes might be called *eigenjudgments* for this set of semantic scales.

In Figure 7-10(b), the correlation between each feature and the two basis axes is shown. This lets us determine whether the new axes are similar to any of the judgments that were the basis of the factor analysis, and how well the new feature space does at containing the variance of each of the perceptual judgments. We see that the first eigenvector is closely aligned with the judgments LOUD and SOOTHING, and that the second eigenvector is closely aligned to INTEREST, orthogonal to SOOTHING. The new feature space explains at least 70%, and in some cases more than 90%, of the variance in the original judgments.

This result indicates that, in fact, the six semantic judgments that I elicited on this task actually reflect only two underlying perceptual dimensions. The question of the physical features to which the judgments (and the underlying dimensions) correspond is the topic of the next section.

7.2. Modeling semantic features

The experimental data I presented in the previous section are useful in two ways. First, the data may be used to examine the nature of immediate judgments about music. This was the main approach pursued in Section 7.1.6. Second—more importantly from the perspective of this dissertation as a whole—we make examine the ability of psychoacoustic models to predict the data. This is the focus of the present section.

I will address three modeling questions. First, how well can the features extracted from musical signals in Chapter 6 predict the mean results of the experiment in the previous section? Second, what can be said about individual differences in model fit? And third, how do the predictions made by this feature model compare to the predictions made by other feature models?

The modeling technique that I use here is multiple linear regression, in which a set of linear equations on the predictors (feature variables) is fit to the outcomes (human judgments) in the way that minimizes mean-squared error. Naturally, there are a variety of other techniques that could be used that would enable modeling of consistent nonlinearity in the data, such as neural networks. In the future, such approaches should be considered and evaluated. However, the amount of data collected in Experiment I is probably not adequate to fit the more numerous degrees of freedom presented by these techniques.

7.2.1. Modeling mean responses

Consider the psychoacoustic features extracted in Chapter 6 as a vector for each piece of music. That is, for musical stimulus i, the vector $\mathbf{x}_i = [\mathrm{MEANIM}_i\ \mathrm{VARIM}_i\ \mathrm{MEANMOD}_i \ldots \mathrm{LOUDEST}_i]^\mathrm{T}$ contains the values of the k features for this stimulus. Then, for each perceptual feature described in the previous section, we wish to compute a coefficient vector $\beta = [\beta_1 \beta_2 \ldots \beta_k]^\mathrm{T}$ such that, if y_i is the mean human response to stimulus i on one rating scale,

$$E = \frac{\sum_i \left(\mathbf{x}_i^\mathrm{T}\beta - y_i\right)^2}{N} \tag{7-1}$$

is minimized, where N=150 is the total number of stimuli. The β in this equation give the relative weights and directions of contributions of each of the $\mathbf{x}_i$ to the overall prediction

$$\hat{y}_i = \mathbf{x}_i^\mathrm{T}\beta \tag{7-2}$$

The β are the same for each musical stimulus (thus, the matrix solution is underdetermined and only a least-error solution can be found) and differ for each perceptual feature. Post-hoc analysis of the β can be used to determine which of the predictors are making a statistically-significant contribution to the model. An additional statistic, termed R^2, can be derived from

the model. R^2 describes the proportion of variance in the independent variables that is explained (modeled) by the predictors through the regression equation (it is the square of the correlation coefficient between the predictions and the observed data). See a standard reference (for example, Amick and Walberg, 1975) for a more in-depth statistical presentation on the multiple regression procedure.

Table 7-7 shows the fit of a simple linear-regression model to each of the mean judgments. The same judgments that showed strong intersubject correlation in the previous section (SLOW, LOUD, SOOTHING) are relatively easy to predict using linear regression and the features measured in Chapter 6. The judgments with less intersubject correlation in the previous section (SIMPLE, ENJOY, INTEREST) are more difficult to predict this way. This result is consistent with the hypothesis discussed in the previous section: that SLOW, LOUD, and SOOTHING are "surface features" of music that are based on the sorts of properties extracted in Chapter 6, while SIMPLE, ENJOY, and INTEREST are based on other features, which are perhaps more variable between subjects.

For the three variables that can be modeled well this way, the model fits are quite good. Figure 7-11 shows scatterplots of predicted vs. observed data for each of the six features. The predictions of SLOW and LOUD in particular match very well with the observed data. That is, even given the likelihood that there are individual differences that have not yet been considered, and the certainty that some cognitive processing is involved in making these judgments, the surface psychoacoustic features extracted through direct signal processing predict half of the variance in the data.

To examine the role of the various predictors in the model, a slightly different regression model is useful. The *stepwise regression* procedure is a variant on the overall linear regression method that more robustly considers intercorrelation among the predictors. In this variant, the features are added one at a time. That is, first, the single feature that is best correlated with the observed data is added to the model. Based on this feature, the regression coefficient is calculated, the model predictions obtained, and the residuals (the differences between the model prediction and observed data) calculated. In each subsequent step, the single feature that best explains the *residuals* from the previous step is added to the model. The coefficients are recalculated and used to produce new residuals. The process ends when no individual feature is significantly correlated at $p < 0.05$ with the residuals. In this method, a minimal list of features that explains the data well (although perhaps not quite as well as the full regression matrix) is obtained. The results of the stepwise regression are shown in Table 7-8.

Judgment	R^2	**Best predictor**
SLOW	.531	-BESTT
LOUD	.516	-VARIM
SOOTHING	.495	MEANMOD
ENJOY	.378	MEANMOD
SIMPLE	.294	MEANIM
INTEREST	.236	PCHENT

Table 7-7: Fit of linear regression model, using the 16 psychoacoustic features presented in Chapter 6, to the mean intersubject ratings for each stimulus elicited in Experiment I. The R^2 column shows the proportion of variance in the observed data that is explained with the linear model. For the most-easily-predicted judgments, about 50% of the variance can be explained with these simple features in linear combination. The **Best Predictor** column shows the feature that is more significantly correlated with the observed data; a minus sign indicates a strong negative correlation. All regressions are significant at $p < 0.001$.

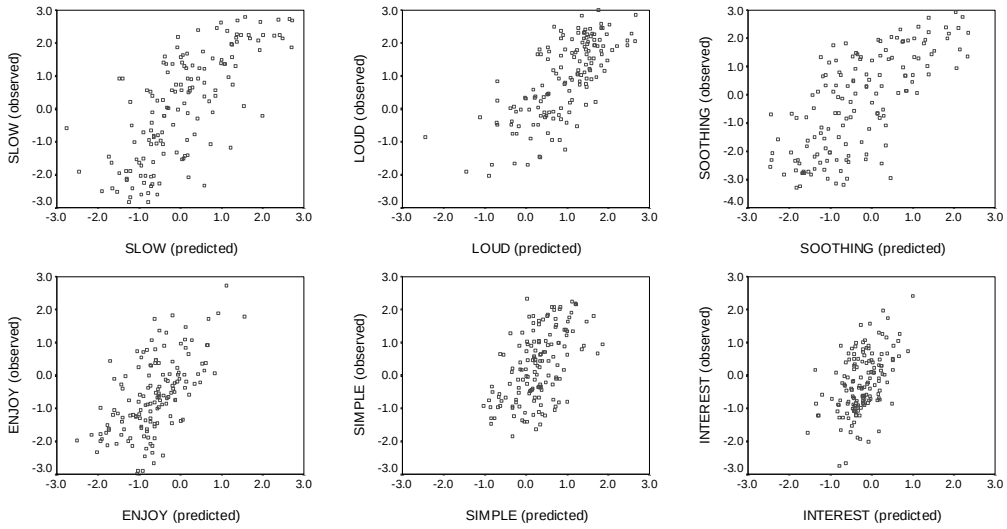


Figure 7-11: Scatterplot of predicted vs. observed values for the 150 musical trials, for each of the six feature judgments. The observed values are the mean responses of the 30 subjects to each stimulus, as described in Section 7.1. The predicted values are computed by the best regression fit from the 16 physical features described in Chapter 6 to the observed values. Notice that, for the features that are difficult to model (ENJOY, SIMPLE, and INTEREST), there is strong regression to the mean in the predictions, and to a lesser extent, in the observed data.

As seen in the table, in all cases more than half of the variance explicable with the full model can be explained with a much shorter list of features, and in one case (LOUD) the relative proportion of variance is greater than 90%. These smaller models are better models, both in the sense of being more statistically significant, and in the minimum-description-length sense of the average amount of variance explained per feature.

By examining the features entered, we can learn something about the degree to which each of them is useful. Clearly the most important feature is MEANMOD which is entered as the first predictor for four of the six regressions. For LOUD, SLOW, and SOOTHING, MEANMOD alone explains about 25% of the variance in the judgments. This is a strong indication of the importance of modulation detection (Langner, 1992) in real music, not just test signals. Most of the other psychoacoustic features are used by one model or another; at least one tempo feature is entered in each model.

Notably absent are any pitch-based features. This absence is compatible with two hypotheses. First, that pitch and pitch manipulations are not perceptually-salient aspects of the sorts of judgments elicited in this experiment. (It seems certain that pitch is an essential feature of other aspects of music perception such as melody recognition, however). Second, that the particular pitch features extracted in Chapter 6 are not robust enough, or in some other way not meaningful enough, to enable good predictions. The present data cannot distinguish these hypotheses.

As well as attempting to explain each of the individual judgments, we can continue to investigate the hypothesis, raised in Section 7.1.6, that there are only a few underlying perceptual dimensions that as serve a basis for all six of the semantic judgments. In Figure 7-10 the correlation of each of the judgments with the two-factor principle-component (factor analysis) solution was shown. We can also model these two factors directly, using linear regression on the psychoacoustic features. The result is shown in Table 7-9.

As seen in this table, it is much easier to model the first basis vector than the second. More of the variance in the first basis vector is explained using psychoacoustic features of the sounds than for any of the individual judgments. Yet the second basis vector is more difficult to explain than any of the perceptual judgments are.

It is frankly difficult, even for a human listener, to understand the projections of the musical examples onto the second basis vector. When the musical examples are sorted in order of

Judgment	Full R^2	Features	Stepwise R^2	Next feature	*p*(next)
SLOW	0.531	-MEANMOD	0.229	-PCHSTB	0.063
		-TEMPCNT	0.353		
		-BESTT	0.420		
LOUD	0.516	MEANMOD	0.261	BESTT	0.071
		TEMPCNT	0.362		
		SPECSTB	0.395		
		LOUDEST	0.427		
		-VARIM	0.470		
SOOTHING	0.495	-MEANMOD	0.238	-LOUDEST	0.157
		-TEMPCNT	0.395		
		VARIM	0.428		
		-SPECSTB	0.445		
		-BESTT	0.461		
ENJOY	0.378	-MEANMOD	0.122	BESTT	0.050
		-TEMPCNT	0.202		
		LOUDEST	0.253		
		-TEMPSTB	0.292		
		PCHENT	0.316		
SIMPLE	0.294	-SPECSTB	0.064	TEMPSTB	0.072
		VARIM	0.102		
		-LOUDEST	0.138		
		-VARIBI	0.163		
		-BESTT	0.186		
INTEREST	0.236	-SPECSTB	0.072	TEMPSTB	0.052
		VARIBI	0.106		
		LOUDEST	0.131		

Table 7-8: Results of stepwise regressions for each of the six perceptual judgments, using the physical features of chapter 6. Each regression is significant at $p < 0.001$ or better. The "full R^2" column shows (for comparison) the proportion of variance that can be explained by entering all 16 features into the regression at once. These values are the same as in Table 7-7. The "Features" column and "Stepwise R^2" column show the order in which features are entered in a stepwise regression, and the proportion of variance explained at each step. For example, for judgment SLOW, the first feature entered is MEANMOD, which is negatively correlated with the judgment and explains 22.9% of the variance. The second feature entered is TEMPCNT, which is negatively correlated with the residual after MEANMOD's effects are considered. MEANMOD and TEMPCNT together explain 35.3% of the variance in SLOW. The "Next feature" and "p(next)" columns show the feature that has the most significant correlation with the residual, after all features at the $p < 0.05$ level or better have been entered. In most cases, the *p* level of the next feature is very close to significance, indicating that there is no hard boundary between the features that are useful for explaining the perceptual data and those that are not.

their values along this vector, I cannot qualitatively tell the examples that wind up at one end from the examples that wind up at the other or to characterize their order in any useful way. Thus, the perceptual reality of an underlying dimension that corresponds to this eigenvector must be considered with some skepticism.

Rating	SOOTHING	SLOW	LOUD	SIMPLE	ENJOY	INTEREST
Mean of variance	1.898051	2.010852	2.031866	2.404179	2.488282	2.645411

Table 7-10: Means across stimulus of intersubject variances in response for each of the semantic ratings. These are calculated by computing the intersubject variance for each rating for each of the 150 stimuli, and then taking the means of the variances across stimuli. The ratings that were difficult to model in the preceding section have 20-40% more variance than the ratings that were easy to model. This indicates that subjects agree less about the meaning and psychoacoustic correlates of the difficult semantic features.

7.2.2. Intersubject differences in model prediction

A natural hypothesis stemming from the results in the preceding section is that for some judgments (LOUD, SOOTHING), subjects generally agree on the meaning and psychoacoustic correlates of the semantic scale, while for other judgments, the subjects

understand the scale differently from one another. The previous results are consistent with this hypothesis, because drawing the subjects from a homogenous group (as is the case for LOUD and SOOTHING) fits the assumptions of the regression model better than drawing the subjects from several different groups does.

Another way to observe this is simply to inspect the intersubject variances in semantic rating for each stimulus and each rating scale. These are shown in Table 7-10. As seen there, the easily-modeled ratings have much less intersubject variance than do the ratings that were more difficult to model. This is consistent with the same hypothesis: that it is individual differences make it difficult to model the overall results for some ratings.

In this section I will explore the possibility of modeling each subject's individual responses with multiple-regression models. This allows us to distinguish between two different forms of the individual-differences hypothesis. First, that the intersubject variance stems from different subjects weighting the various psychoacoustic factors differently in their judgments of what makes a stimulus complex or interesting. If this is the case, then we should be able to predict the ratings on the "difficult" scales as well subject-by-subject as for the "easy" scales. Second, that the intersubject variance stems from different methods of combining the factors, or different factors entirely (such as cognitive factors). If this is the case, then that ratings like SOOTHING and SLOW should continue to be easier to predict than SIMPLE and INTEREST, even for individual subjects.

In the previous section, I used multiple linear regression to model the mean ratings across subjects. This was an admissible procedure because (as shown in Figure 7-8) the intersubject mean ratings are roughly normally distributed. However, this is not the case for the individual

Eigenvector	Full R^2	Features	Stepwise R^2	Next feature	p(next)
First	0.587	-MEANMOD	0.277	-TEMPSTB	0.242
		–TEMPCNT	0.442		
		–BESTT	0.482		
		–SPECSTB	0.517		
		VARIM	0.541		
		–LOUDEST	0.555		
Second	0.227	LOUDEST	0.056	-TEMPSTB	0.055
		VARIBI	0.101		

Table 7-9: Results of stepwise regressions, using the psychoacoustic features to model the two basis vectors determined through principle components analysis (factor analysis) in Section 7.1.6. Specifically, for each piece of music, the six-dimensional perceptual vector is projected into the rotated two-dimensional basis space. The projected distance along each basis vector for each piece of music is used as the dependent variable in the multiple-regression analysis. All R^2 values are significant at or beyond $p < 0.01$. The first eigenvector is easier to model with psychoacoustic features than is the second eigenvector.

Judgment	Correct: Range	Correct: Mean	% significant
SOOTHING	69.8%—95.4%	77.4%	97.7%
SLOW	70.6%—96.6%	78.3%	93.3%
LOUD	65.5%—96.6%	82.0%	76.7%
ENJOY	64.4%—84.4%	73.5%	70.0%
INTEREST	61.9%—90.4%	71.4%	60.0%
SIMPLE	60.8%—93.1%	73.6%	46.7%

Table 7-11: Results from subject-by-subject logistic regressions, using the psychoacoustic features to predict a bivalent form of each of the response scales (that is, whether the rating for each stimulus is "high" or "low", without regard to the magnitude of the offset). The two results that are returned for each subject are the proportion of stimuli that are predicted correctly by the logistic model, and whether or not this level of prediction is statistically significant. The second column shows the range of proportion correct across subjects for each feature. That is, for SOOTHING, as few as 69.8% (for one subject) or as many as 95.4% (for another subject) of the bivalent ratings were correctly predicted by the model (random guessing would give 50% for each subject here). The third column shows the mean of proportion correct across subjects. That is, for SLOW, on average 78.3% of the stimuli were predicted correctly. The fourth column shows the proportion of subjects for whom the model was significant[24]. That is, for LOUD, 76.7% (23 out of 30) of the subjects could be individually modeled to a statistically significant ($p < 0.05$) extent (random guessing would give the α level, or 5%, here).

subject-by-subject data, and so multiple linear regression cannot be used. Instead, I will use a variant statistical procedure called *logistic* regression, which allows the distribution of binary variables to be modeled by continuous predictors. Since the distribution of responses collected in Experiment I was strongly bimodal, there is a natural binary variable to derive from each of the response scales. Namely, whether the response is above or below the center point. Such a model follows the hypothesis that the basic judgment made in each case is bivalent—loud vs. soft, fast vs. slow, interesting vs. boring—and that the actual magnitude offsets are less important.

The logistic regression procedure attempts to estimate the probability of response (that is, the probability of a "high" rating) as a linear regression, using an equation of the form

$$\log\frac{\pi_k}{(1-\pi_k)} = \alpha + \beta_0 x_{k0} + \beta_1 x_{k1} + \ldots + \beta_n x_{kn} \qquad (7\text{-}3)$$

where π_k is the probability of response observed in trial k, the x_{ki} are the observed independent variables (the psychoacoustic features), and the α and β values are the regression weights. Significance can be tested by comparing the estimated probability to the observed probability in light of the number of predictors.

Results from logistic regressions for each of the six semantic ratings are summarized in Table 7-11. These values were calculated by performing a full logistic regression for each rating for each subject and summarizing the results. The logistic regressions were computed using a "full-entry" method in which all 16 psychoacoustic features are entered at the same step. Stepwise regressions could also be computed, but this produces a great deal of data that is difficult to summarize and interpret, and so is left as a topic for future work. It is possible that more subjects might be predictable to a statistically significant degree in the stepwise regression models, due to the fewer degrees of freedom in the initial stages of the stepwise models.

Looking at the results shown in Table 7-11, the overall performance is quite good. Depending on the semantic judgment, anywhere from nearly half to nearly all of the subjects' ratings can be significantly predicted by the logistic model. This is despite the fact that since there is only one rating for each stimulus from each subject, it is likely that there is a fair amount of experimental error in the ratings themselves.

The results seem to be inconsistent with the first hypothesis, and consistent with the second hypothesis, regarding intersubject variance. That is, there remain differences in the quality of modeling between the "easy" scales like SOOTHING and the "difficult" scales like SIMPLE. If intersubject variability for the difficult ratings were only a matter of combining the same psychoacoustic features in different ways, then this would not be the case. From this, we can tentatively draw the conclusion that the individual differences on these scales arise from features not tested here (such as cognitive factors) or different methods of combining features, or both.

I also performed analyses of variance to explore the relationship between the subject demographics and the predictability of each scale for each subject. The null hypothesis in this case is that whether the subject is a musician or not, is male or female, is young or old, has no effect on the ability of the logistic regression model to predict his/her ratings. None of the ANOVAs were significant and so this null hypothesis cannot be rejected.

7.2.3. Comparison to other feature models

Over the long run of the experimental-modeling research program laid out here, there are a number of methods that should be used to evaluate models. One is to try to use a single model to explain different things. This is one way to think of the connections between the low-level results in Chapters 4 and 5, the feature extraction in Chapter 6, and the modeling results presented in this chapter. The psychoacoustic models of tempo and auditory-image formation are shown to be useful for explaining low-level percepts in the earlier chapters, and for higher-level percepts in this chapter.

Another important evaluation, though, is to look at the performance of models in comparison to other models that do the same thing. Considered only in its abilities to explain musical perceptions, the linear-regression model presented in Section 7.2.1 is very complex indeed. In order to extract the features needed for this model, there is a great deal of sophisticated preprocessing. The good modeling performance achieved shows only that this sort of complex model is sufficient to explain musical perceptions, not that it is necessary. It might be the case that a simpler model could do equally well. If so, Occam's razor would force a re-evaluation of the use of the complex model as the best explanation for these perceptual data.

To fully address this question is a long-term prospect, because the complex model really needs to be compared to all simpler models. There have been few models proposed that target similar questions, so it is difficult to draw from the research literature to find candidates. However, one such is found in my previous research on speech/music discrimination (Scheirer and Slaney, 1997). In that research, we found that a 13-dimensional feature space, using simple spectral-based features, was sufficient to provide approximately 4% error rate in categorizing complex ecological signals (sampled directly from a radio) as "speech" or "music."

We tested several trained classification frameworks in order to examine speech/music discrimination performance. T here were few significant differences between the frameworks in discrimination error. Here, I will apply the 13-dimensional feature space to the same linear-regression approach to modeling the perceptual data. See the original report (Scheirer and Slaney, 1997) for more extensive details on this feature set and the way we evaluated its performance on the speech/music task.

This is an unfair comparison, because of course the speech/music features were not intended to be used in this way, and there is little reason that they should work particularly well. There was no perceptual motivation in constructing the feature set for that classifier. However, at a high level of abstraction, it could be argued that the features from the speech/music discriminator capture some of the surface information in music. Thus, they serve as a point of evaluation for the more complex, perceptually-motivated, features developed in Chapter 6.

The speech/music features are calculated through continuous signal-processing of the musical signal. Each is updated 50 times per second. To calculate single values for each of the 150 test stimuli, I took the mean of each feature value over the last second of the signal. This is not necessarily the same as *processing* only the last second of the signal, since some of the features maintain a temporal history and thereby reflect the early part of the signal in part as well.

Judgment	R^2 (Psycho-acoustic)	R^2 (Speech/ music)	R^2 (Counterpart)	R^2 (Other judgments)
SLOW	.531	.414	.692	.629
LOUD	.516	.421	.578	.816
SOOTHING	.495	.408	.681	.885
ENJOY	.378	.291	.588	.824
SIMPLE	.294	.208	.291	.473
INTEREST	.236	.207	.351	.723

Table 7-12: Multiple-regression results comparing four different feature sets as predictors for the mean perceptual judgments averaged across subjects. All R^2 values are significant at $p < 0.001$. From left to right, the columns show the results of different predictor sets, to wit: (1) the 16 psychoacoustic features from Chapter 6; (2) the 13 physical features previously used to build a speech/music discriminator (Scheirer and Slaney, 1997); (3) the 6 counterpart features, that is, the mean perceptual judgments on the other excerpt from the same piece of music; and (4) the 5 other mean perceptual judgments on the stimulus.

As well as the speech/music features, I used the perceptual data itself in two other ways to model the perceptual judgments. First, since there were two segments selected from each musical selection (75 x 2 = 150 stimuli in the whole dataset), for each stimulus, the *counterpart* stimulus can be used as a reference. That is, given the perceptual ratings of selection #1, excerpt A, how well can these be used to model the perceptual ratings on selection #1, excerpt B?

Second, since there is a significant amount of intercorrelation among the judgments, for each judgment, the *other judgments* can be used as a reference. That is, given the perceptual ratings of LOUD, SLOW, SOOTHING, INTEREST, and ENJOY on a particular stimulus, how well can they be used to model the rating of SIMPLE on that stimulus? (Given the strength of the first few eigenvectors in the factor analysis in Section 7.1.6, it is to be expected that the modeling results will be very good in this method).

A comparison of model predictions using the 16 physical features from Chapter 6, the 13 speech/music features, the counterpart ratings, and the other judgments on the same stimulus, for each of the six perceptual judgments, is in Table 7-12.

The psychoacoustic features clearly fall between the speech/music features and the counterpart features in their ability to model the human perceptual judgments. They are between 14% and 41% better (mean = 26.2%) than the speech/music features, and from 1% better to 56% worse (mean = 30.5%) than the counterpart features. The other judgments on each stimulus are the best basis for modeling each perceptual judgment—this is unsurprising,

because it is likely that the different perceptual judgments elicited are actually reflecting only a few underlying percepts. The speech/music features show the poorest performance, again unsurprising, for the reasons discussed above.

7.3. Experiment II: Perceived similarity of short musical stimuli

Experiment I demonstrated that subjects can make immediate, reasonably consistent judgments about short musical stimuli in full ecological complexity. The modeling results in Section 7.2 demonstrated that, for some of these judgments, fully half the variance in the observed data can be explained through a simple linear-regression model using features extracted by a psychoacoustic model.

In this section, I will present the results from a second experiment along the same lines. The question posed in Experiment II is to what degree subjects can consistently judge the *similarity* of two pieces of music from short examples. From an applications viewpoint, the judgment of similarity holds a position of prime importance. This is because if we wish to build software agents that can search the Internet (or other musical databases) for music that is similar to music that we like, we must first understand the nature of musical similarity. This includes such crucial questions as whether there is some common judgment of similarity shared by listeners acculturated in the same musical environment, or whether the judgment of similarity differs from person to person in important ways.

To my knowledge, there has been no formal experimental study of musical similarity based on these questions. However, there is a significant and rigorous body of previous work related to this in the area of *melodic similarity* (Hewlett, 1998). Such studies have been narrower than my interest here in several important ways. First, this area focuses on melody as an aspect of music that is independent of timbre and orchestration. This may or may not prove to be empirically the case, but it seems to me a rather strange thing to assume. There are deep questions that immediately arise, for example, when we ask whether the Rolling Stones' version of "Satisfaction" is more similar to the Montovani Strings' easy-listening version of "Satisfaction," or to the Rolling Stones' song "Sympathy for the Devil." A narrow viewpoint of melodic similarity holds that the two versions of the same song must be very similar in important ways, since they have the same melody. Yet it is clear that they are also different in important ways. To understand *which ways* in which they are similar and different—for example, in terms of the behaviors that are differentially afforded by the three performances—should be a topic of more study. Similarity depends on context: it makes no sense to ask whether two pieces of music are similar without asking what they are similar *for* The experiment that I present in this section suffers from the problem of reduced context as well.

Second, all of the research on melodic similarity has taken a structuralist viewpoint, trying to explain how melodies are similar to, or different from, one another through representations of melodies as hierarchically-structured lists of notes. Again, this representation may or may not prove to be an empirically useful one, but it still begs important questions about the relationship with the hearing process and the signal processing in the auditory system. Finally, most research on melodic similarity has focused on relatively long-term structural phenomena, in which melodies go on for 10-15 seconds or more. In contrast, I am particularly interested in the judgment of immediate similarity—how a listener might decide within only a second or two of listening how a newly heard piece of music is similar or dissimilar to other music with which she is familiar.

	Female	Male
M0	9	3
M1	6	10
M2	1	3

Table 7-13: Cross-tabulation of subjects' musical ability by gender for Experiment II.

	18-25	25-30	31-40	40+
M0	8	1	2	1
M1	11	4	0	1
M2	2	0	2	0

Table 7-14: Cross-tabulation of subjects' musical ability by age for Experiment II.

7.3.1. Overview of procedure

Thirty-three musically trained and untrained subjects listened to 190 pairs of five-second musical excerpts. The pairs of stimuli represented each of the pairwise combinations of a subset of twenty stimuli drawn from the larger stimulus set used in Experiment I. The subjects used a computer interface to listen to the stimuli and judge whether the two pieces in each trial were similar or different.

7.3.2. Subjects

The subjects, 33 in all, were drawn from the MIT community, recruited with posts on electronic and physical bulletin boards. 21 of the subjects ("repeat subjects") had previously participated in Experiment I, the other 12 had not ("new subjects"). There was an interval of approximately one month between the repeat subjects' participation in Experiment I and their participation in this experiment. Most subjects (66%) were between 18 and 23 years of age, the rest ranged from 25 to 72 years. The median age was 22 years.

Of the 33 subjects, 16 were male, 16 female, and one subject declined to answer the demographic questionnaire. All but four subjects reported normal hearing. 23 reported that they were native speakers of English, and 7 reported that they were not. Seven subjects reported that they had absolute-pitch ability in response to the question "As far as you know, do you have perfect pitch?" Other than the repeat subjects' participation in Experiment I, the subjects had no consistent previous experience with musical or psychoacoustic listening tasks.

The same questionnaire regarding musical ability used in Experiment I was administered to the subjects after they completed the experiment. The subjects were categorized as M0 (N = 12), M1 (N = 16), or M2 (N = 4) subjects as described in Section 7.1.2.

Cross-tabulations of musical ability by gender and by age are shown in Table 7-13 and Table 7-14, respectively.

7.3.3. Materials

Experimental stimuli were paired 5-second segments of music. There were 190 stimuli in all, formed by taking all pairwise combinations of a 20-segment subset of the database used in Experiment I. The 20 selections used were the first excerpt from each of the songs numbered 1…20 in the original database as shown in Appendix A (after normalization as reported in

Figure 7-12: The computer interface for listening experiment II. Each subject rated each paired stimulus (marked "Play A" and "Play B" on the interface) to indicate whether he found them to be similar or different.

Section 7.1.3). There were no pairs of stimuli drawn from the same composition. Since the Experiment I stimulus database was ordered randomly (based on the sampling procedure used to obtain examples from MP3.com), this method of drawing samples from it presents no particular bias and may also be taken as a random sample of the full MP3.com database.

7.3.4. Detailed procedure

Subjects were seated in front of a computer terminal that presented the listening interface, as shown in Figure 7-12. The interface presented a slider labeled **very similar** on one extreme and **very different** on the other. The subject was instructed that his task was to listen to pairs of short musical excerpts and report his judgments about the similarity of each pair. It was emphasized to the subject that there are no correct answers on this task, and that the experiment was only designed to elicit his opinions. Three practice trials were used to familiarize the subject with the experimental procedure and to set the amplification at a comfortable listening level. The listening level was allowed to vary between subjects, but was held fixed for all experimental trials for a single subject.

The 190 trials were presented in a random order, different for each subject. Within each trial, the order of the two examples (which one was marked "A", and which one "B") was also randomized. When the subject clicked on the **Play A** or **Play B** buttons, one of the examples was presented. After the music completed, the subject moved the slider as he felt appropriate to rate the similarity of the stimulus pair. The subject was allowed to freely replay the two examples as many times as desired, and to make ratings after any number of playings. When

the subject felt that the setting of the rating slider reflected his perception of similarity, he clicked the **Next** button to go on to the next trial. The slider was recentered for each trial.

The subjects were encouraged to proceed at whatever pace was comfortable, taking breaks whenever necessary. A typical subject took about 60 minutes to complete the listening task.

7.3.5. Dependent measures

For each trial, the final setting of the slider was recorded to a computer file. The computer interface produced a value from 0 (the bottom of the slider) to 100 (the top) for each rating on each trial. A value of 0 indicates that the subject found that pair of examples to be very different, while a value near 100 indicates that the subject judged that pair to be very similar. Trials on which the subject did not move the slider (value of 50) were rejected and treated as missing data.

7.3.6. Results

This section reports the results on the similarity-matching experiment.

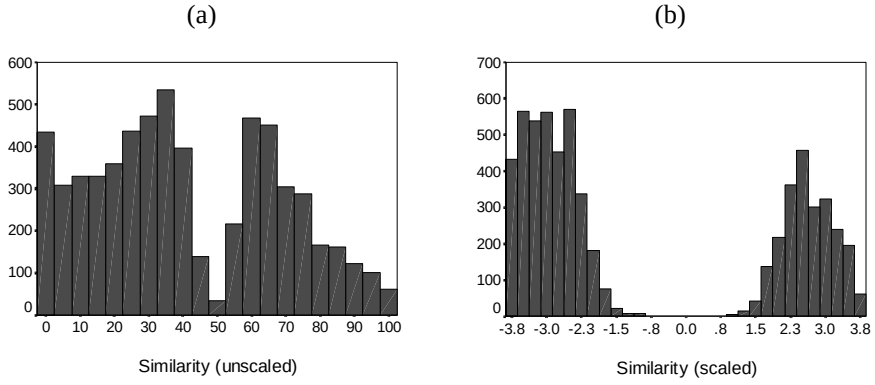


Figure 7-13: Distribution of responses in similarity-judgment task, before rescaling (a) and after rescaling (b). The same pattern of responses is seen in the unscaled data as in the responses to Experiment I, so the same rescaling function (with the cube-root nonlinearity) is used.

Response distribution

As in Experiment I, the responses in Experiment II are not distributed on a normal curve (Figure 7-13). Either of the explanations suggested for the earlier experiment would seem to apply here as well. Therefore, the same scaling function (recentering the scale to zero and applying a cube-root nonlinearity) was used for further analysis and modeling study. For these data as well, the scaling function makes the response pattern nearly a perfect bimodal distribution.

Learning and fatigue effects

As with Experiment I, the relationship between the trial number of each pair (the ordinal number of the place in the sequence of pairs to a subject that that trial occupied) and the judged similarity was investigated. Unlike in Experiment I, there was no significant correlation between trial number and similarity. An ANOVA comparing the means on each trial instance was also not significant. Both results are consistent with the null hypothesis that there is no learning or fatigue effect in this experiment.

Effects of participation in Experiment I

Since some of the subjects had participated previously in Experiment I, and some had not, it is important to know whether the earlier participation had any consistent effect on responses. A *t*-test comparing the mean offset of the judgment of similarity (that is, the rescaled similarity, collapsed around the midpoint) across all stimulus pairs between repeat subjects and new subjects was significant ($t(6106)=8.503$, $p < 0.001$). As shown in Figure 7-14, repeat subjects' responses were, on average, slightly closer to the center of the slider than were new subjects' responses. The effect is similar in absolute magnitude to the other demographic differences between subjects.

Repeat subjects were not a random sample of the Experiment I subject pool. They participated in Experiment II according to their time schedule and personal interest. Thus, there may be a consistent selection bias in the group of subjects that repeated that would explain this effect. From the available data, this hypothesis cannot be distinguished from one in which the participation in Experiment I *itself* had an effect.

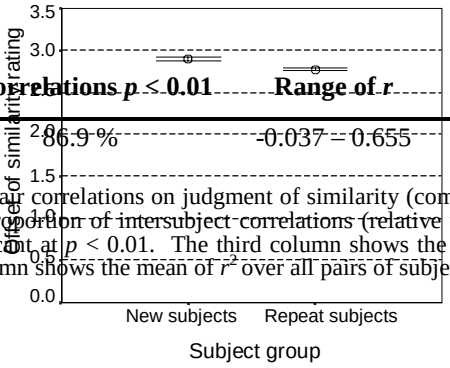


Figure 7-14: Effect of participation in experiment I on mean offset of similarity. Subjects that participated in experiment I had slightly (but significantly) less extreme opinions about the similarity of stimuli that subjects that did not.

Judgment	% correlations $p < 0.01$	Range of r	Mean r^2
Similarity	86.9 %	-0.037 – 0.655	0.163

Table 7-15: Intersubject pair-to-pair correlations on judgment of similarity (compare to Table 7-3). The second column shows the proportion of intersubject correlations (relative to the 496 possible pairs of subjects) that are significant at $p < 0.01$. The third column shows the range of r over all pairs of subjects. The fourth column shows the mean of r^2 over all pairs of subjects.

Intersubject correlations

As with the semantic-judgment data, the intersubject correlations regarding the judgment of similarity can be analyzed. The results are in Table 7-15.

The intersubject correlations on the similarity judgments resemble the results for the "easy" semantic judgments (LOUD, SOOTHING) more than they do the "difficult" ones (INTEREST, ENJOY). There is a high degree of correspondence between subjects, only a few of the intersubject pairs are negatively correlated (and none significantly), and one subject's data explains a good proportion (16%) of the variance in another randomly-selected subject's data. This makes it seem likely that there is general agreement among subjects regarding the meaning of similarity and the features of sounds that correlate with it.

Analyses of variance

I conducted analyses of variance to explore the relationship between the demographics of the subjects and their judgments of similarity. As with the analyses of variance for the semantic-feature judgments, I collapsed the similarity ratings across the center of the scale, to create a normally-distributed variable indicating the offset from the center.

Table 7-16 shows the results of analyses of variance examining the relationship between the stimulus pair, the subject, and the similarity rating. As seen there, there are strongly significant effects of both stimulus pair (indicating that some pairs are consistently judged to

be more similar that others) and subject (indicating that some subjects consistently find all of the pairs to be more similar than other subjects do).

Dependent variable	$F_{SUBJ}(32)$	*p*	$F_{STIM}(189)$	*p*
Similarity	20.453	**0.000**	13.975	**0.000**

Table 7-16: Summary of analyses of variance examining potential response differences based on the stimulus and the subject. As expected, both of these analyses are strongly significant, indicating that the similarity differs from one stimulus pair to another, and that some subjects find all of the stimulus pairs to be consistently more similar than do other subjects.

Independent variable	df	F	p
MUS	2	104.536	**0.000**
AP	1	71.631	**0.000**
ENG	1	3.730	0.053
SEX	1	5.579	**0.018**
AGE	3	1.562	0.197

Table 7-17: Summaries of analyses of variance exploring dependencies of similarity response on the demographics of the subjects. The five demographic variables tested are: musical ability (MUS), self-reported absolute pitch (AP), native English language speaker (ENG), sex (SEX), and age, segmented into four categories (AGE). Effects are strongly significant for MUS and AP, mildly significant for SEX, trending towards significant for ENG, and not significant for AGE.

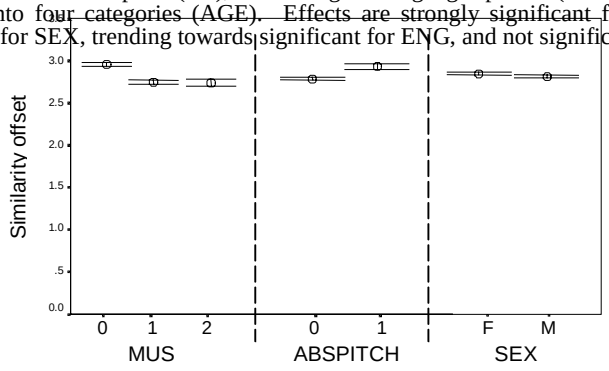


Figure 7-15: Mean and 95% confidence interval of similarity offset, averaged over all stimulus pairs, broken down three ways. (Left) Similarity offset broken down by musical ability. Non-musicians use the ends of the scale more than do musicians. (Center) Similarity offset broken down by self-reported absolute-pitch (AP) ability. Subjects who claim AP use the ends of the scale more than subjects who do not. (Right) Similarity offset broken down by sex. Female subjects use the ends of the scale slightly more than male subjects do.

The second analysis of variance explores dependencies of adjudged similarity on the demographics of the subjects: musical ability, self-reported absolute pitch, native language (English or non-English), sex, and age. Age was segmented into four categories for analysis: 17-21, 22-29, 30-39, and 40+. These analyses are summarized in Table 7-17.

As with the Experiment I demographics, it is difficult to give a coherent interpretation of these effects. From their strengths, and their consistent appearance in two different experiments, it seems likely that they are real, but more research is needed to explain their origins. Figure 7-15 shows the differences in similarity offset ratings as they covary with musical ability, absolute-pitch ability, and sex. Different than Experiment I (see Figure 7-6), in Experiment II it was the *non-musicians* that used the ends of the scale more.

7.4. Modeling perceived similarity

In this section, I will develop models of the data collected in Experiment II; that is, models that predict the perceived similarity of short musical experiments. First, I will use the psychoacoustic features of Chapter 6 to develop a multiple-regression model, as I did above for the Experiment I data. Then, I will use the Experiment I results to model the Experiment II results, and discuss the implications of such a model. Finally, I will explore the structure of the similarity responses through multidimensional-scaling analysis.

7.4.1. Predicting similarity from psychoacoustic features

To continue the approach presented in Section 7.2, I will attempt to model the results of the similarity-judgment experiment with the psychoacoustic features developed in Chapter 6. The hypothesis embodied in this model is that the perceived similarity of a pair of musical examples is determined by comparing the psychoacoustic properties of the two excerpts.

I used the psychoacoustic features from the two examples in each pair to derive *difference* features, which are simply the absolute difference between the values of the features for the two excerpts. That is,

$$\begin{aligned} \Delta\text{MEANIM} &= \left|\text{MEANIM}_1 - \text{MEANIM}_2\right| \\ \Delta\text{VARIM} &= \left|\text{VARIM}_1 - \text{VARIM}_2\right| \\ &\vdots \\ \Delta\text{LOUDEST} &= \left|\text{LOUDEST}_1 - \text{LOUDEST}_2\right| \end{aligned} \tag{6-4}$$

where the index indicates the two members of each stimulus pair.

As with the derivations of the psychoacoustic features themselves (Chapter 6), it is possible to devise other methods of computing distance features. For example, the squared differences could be used, or the features grouped into subspaces and various norms computed within those subspaces. I intend no claim that the method presented here is the best.

The difference features can now be used as predictors in a linear multiple-regression model of the similarity judgments. The multiple-regression procedure determines a linear equation in the difference features that predicts the similarity judgment for each stimulus pair with minimal mean-squared error. The results of the full multiple-regression and the stepwise regression are shown in Table 7-18.

As seen in this table, the physical features can be used to predict the judgment of similarity, about as well as for the most difficult (INTEREST—see Table 7-7) of the semantic judgments. Thus, the hypothesis that perceived similarity between short musical excerpts can

Judgment	Full R^2	Features	Stepwise R^2	Next feature	p(next)
Similarity	**.246**	-ΔMEANMOD	**.104**	-ΔTEMPSTB	0.099
		ΔLOUDENT	**.138**		
		-ΔMEANIBI	**.169**		
		ΔVARIBI	**.186**		
		ΔNUMB	**.204**		

Table 7-18: Multiple-regression results predicting mean similarity from the difference features (compare Table 7-8). All R^2 values are significant at the p = 0.001 level or better. Two of the factors entered in the stepwise regression have negative β values; this is the logical direction, since the greater the difference of features, the less similar the stimuli should be. The three features entered in a positive direction must have a complicated partial-correlation relationship with the residual.

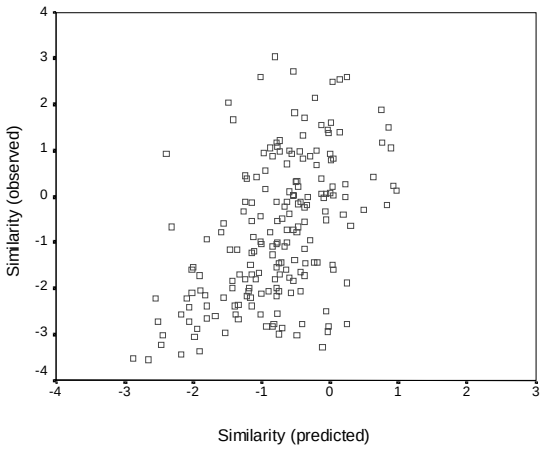


Figure 7-16: Predicted vs. observed similarity for the mean judgments in Experiment II. Each data point here corresponds to one pair of stimulus examples. The x-axis gives the similarity predicted for that pair via linear multiple regression on the differences between the signal processing features for the two excerpts in the pair. The y-axis is the mean rating of similarity across listeners. Note the strong regression to the mean—the regression model never predicts that stimuli will be very similar or very different. This happens because the regression model is unable to make consistently correct predictions for extreme cases, and so the safest thing for it to do (to minimize mean-squared error) is simply to use values near the mean as the predictions.

be explained on the basis of the similarity between their surface features is confirmed. A scatterplot of the predicted vs. observed similarity judgments for the 190 paired stimuli is shown in Figure 7-16.

As with the models of the musical judgments, MEANMOD is the most important predictor of predicting perceived similarity. That is, it is more likely for a pair of stimuli whose MEANMOD values are close to be judged similar, than for any other predictor. The other features entered are mostly tempo features. Once again, no pitch features were used.

Note also that three of the five entered features have positive β values. This means that the more *dissimilar* the stimuli are according to these features (because the difference feature is large), the more similar the stimuli are (because the similarity rating is also large). This outcome is not as unlikely as it seems, because in stepwise regression, the predictors after the first are not actually predicting the main effect directly. Rather, they are predicting whatever residual remains at that step. So, for example, the positive β value on LOUDENT means that once the effects of MEANMOD are accounted for, then stimuli that appear different on LOUDENT are actually quite similar.

7.4.2. Predicting similarity from semantic judgments

A second method of modeling the similarity ratings is to use the *semantic feature judgments* from Experiment I as predictors. In contrast to the one-stage model suggested in the previous section (the psychoacoustic features are compared to derive similarity), this model is a two-stage model. In the first stage, the psychoacoustic features are perceptually analyzed and combined to arrive at the semantic features studied in Experiment I. In the second stage, the semantic features of the two musical examples are compared to derive the similarity of the pair. The data here are not rich enough to allow me to evaluate these hypotheses against each other, but I will explore the two-stage model to make it concrete.

Table 7-19 shows two different analysis using the perceptual ratings. First, on the left, the results of using the semantic features elicited from subjects in Experiment I are shown. In this analysis, I computed the differences between the mean semantic ratings on each stimulus in a

pair, and entered them as predictors of the mean similarity. This works quite well; nearly half the variance in the similarity ratings can be predicted using the differences between the semantic features.

	Semantic judgments			**Modeled judgments**		
	Full R^2	Features	Stepwise R^2	Full R^2	Features	Stepwise R^2
Similarity	**.467**	-ΔSOOTH -ΔLOUD -ΔINTEREST	**.393** **.444** **.456**	**.091**	-ΔLOUD	**.072**

Table 7-19: Multiple-regression results predicting mean similarity from the semantic features of Experiment I (left) and from the psychoacoustic-model-based predictions of the semantic features derived in Section 7.2.1 (right). All R^2 values are significant at $p = 0.002$ or better. Both models are able to predict the similarity judgments; the semantic features are much better predictors.

Thus, if we were able to perfectly predict the semantic feature judgments, we could use these predicted judgments to do better than we presently can by directly modeling the perception of similarity directly from the psychoacoustic data. However, of course the present feature set cannot perfectly predict the semantic judgments (as summarized in Table 7-7), so it is interesting to explore how well we can do this two-step process.

The result is shown in Table 7-19(right). In this model, the differences between the predicted results obtained for the semantic features through multiple regression are themselves used as predictors for the similarity judgments. As seen in the table, while the results are not nearly as good as the direct psychoacoustic model, they are still significant. Notably, only one predictor is entered in the stepwise regression.

A similar analysis as shown on the left of Table 7-19 can be conducted subject-by-subject for those subjects that participated in both experiments. That is, we can use the individual subject's responses on Experiment I (rather than the means) to try to predict his own responses on Experiment II (rather than the means). This works (R^2 is statistically significant) for 18 of the 23 repeat subjects, but again not as well as using the means to predict the means. The R^2 values in this case range over the interval [0.042,0.310] with a mean of 0.122.

7.4.3. Individual differences

Again following the modeling paradigm used for the semantic judgments, I will explore the use of multiple logistic regression to model individual subjects' ratings of similarity. In this case, we use the difference features to construct a model that predicts whether a subject will judge the similarity for each pair to be higher than average, or lower than average.

Two sets of difference features were used: the set of 16 psychoacoustic features, and the set of 6 semantic judgments, for those subjects that participated in both Experiment I and Experiment II. In the latter case, we are attempting to use the subjects' *own* semantic judgments to model their perception of similarity. The results are summarized in Table 7-20.

Even though there are only 40% as many degrees of freedom in the feature space created by the perceptual ratings, they produce nearly as many correct answers as do the psychoacoustic features. It seems easier (in terms of the number of subjects that can be successfully predicted) to model the individual ratings of similarity than the individual ratings of the difficult semantic judgments ENJOY and INTEREST.

7.4.4. Multidimensional scaling

A natural way to explore the results from a similarity-rating experiment is the statistical procedure known as *multidimensional scaling* (MDS). MDS assumes that judgments of

Predictors	df	N	Correct: Range	Correct: Mean	% significant
Psychoacoustic features	16	33	60.1%—93.9%	72.5%	63.6%
Semantic judgments	6	21	59.3%—87.8%	70.58%	77.8%

Table 7-20: Results from subject-by-subject logistic regressions using the psychoacoustic features and semantic judgments (compare to Table 7-11). The first row shows regression results where the predictors are the 16 psychoacoustic features. The second row shows regression results where the predictors are the 6 observed semantic judgments from Experiment I. Only the 21 repeat subjects could be used for the second regression.

similarity reflect the *perceptual distance* between the stimuli. Based on these distances, the stimuli can be placed in a multidimensional space such that the interstimulus distances approximate the similarity judgments as nearly as possible. If it is really the case that the stimuli perceptually occupy a low-dimensional Euclidean space, then the locations derived through MDS will accurately characterize the similarity judgments. In this case, the *stress* of the MDS solution will be low. (Stress is a sort of goodness-of-fit measure). Naturally, the stress of a solution is guaranteed to decrease with increasing dimensionality—that is, the more dimensions are used for analysis, the better the fit will be.

It can be very difficult to interpret the solution returned by the MDS procedure for a given set of similarity judgments. The MDS procedure can be applied to any set of data and is guaranteed to return a result, but of course there is no particular reason that the result must be meaningful in any way. Typically, when MDS is used in the context of analyzing psychological data, we are not interested in the location of the stimuli in the new space as much as we are the nature and scaling of the axes themselves. What we hope is that these axes can be interpreted in terms of the results of independent experiments and that the results scale to include new stimuli that were not included in the original experiment, to predict the similarity of other stimuli.[25]

The mean data collected in the similarity experiment were converted to dissimilarity (distance) judgments by subtracting the scaled values from 3.685 (the scaled endpoint). For example, a raw similarity judgment of 75 is scaled to 2.92 as described in Section 7.1.5, and then converted to a distance of 0.76. On the other side of the scale, a raw similarity judgment of 20 is converted to a distance of 6.79. Naturally, there are many other possible ways to convert the similarity judgments to distances. For this analysis, the data were pooled and the means across subjects analyzed.

Figure 7-17 shows the proportion of the overall variance in the dissimilarity matrix that can be explained with an MDS solution with various numbers of dimensions. As seen in this figure, there is no clear knee in the curve that indicates an optimal number of dimensions to use, trading off the generality of description against the complexity. Rather, each added dimension explains a small proportion (about half) of the remaining variance in the data. This is one indication that the similarity data are not modeled well using the multidimensional-scaling approach. For the analyses presented below, the two-dimensional MDS solution is used as a starting point for modeling.

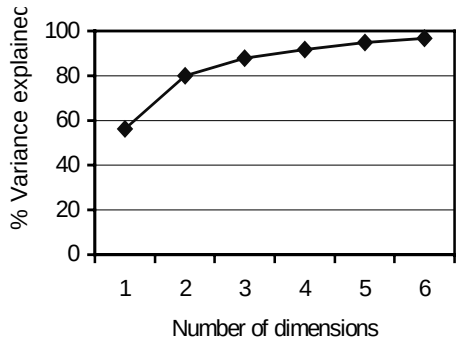


Figure 7-17: Scree plot showing the proportion of the overall variance in the dissimilarity matrix that can be explained with an MDS solution with various numbers of dimensions. There is no obviously-best fit to these dissimilarity; after the first dimension, each additional dimension explains about half of the remaining variance. The two-dimensional solution is selected for further analysis, based mostly on convenience.

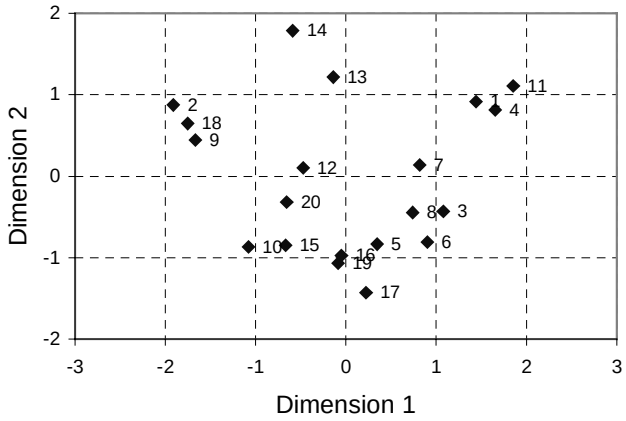


Figure 7-18: 2-dimensional MDS solution for the dissimilarity ratings computed from the human similarity judgments of the 20 similarity-test stimuli. The dimensions are in arbitrary units that could be scaled to recover the dissimilarity ratings in the original data. This solution explains 80.05% of the variance in the full 20x20 similarity matrix.

The two-dimensional MDS solution is shown in Figure 7-18. In this figure, the points (corresponding to the 20 musical stimuli used in the experiment) have been placed in the two-dimensional configuration that minimizes the average error in pairwise distances. The axes are in arbitrary units that could be rescaled to recover the approximate pairwise similarity ratings.

It is possible to interpret this solution by eye and ear to a certain degree. Most notably, the clusters that can be seen in the solution space seem to correspond either to musical genres, or to the instruments present in the samples (or both, as these two descriptors are certainly conflated). There is a group containing stimuli #2, #18, and #9, which are the three hard-core rock-and-roll examples, with noisy distorted guitars and screaming vocals (#9 to a lesser degree than #2 or #18). There is a second group containing #1, #4, and #11, which are the three "classical" sounding stimuli, with long harmonic notes and violins. The group at the bottom containing #17, #19, #16 and branching out into #15, #5, #10, and #6 all contain acoustic guitar prominently.

These sorts of groupings seem to correspond to something that other writers (Hajda *et al.*, 1997) have argued about MDS solutions for timbre-dissimilarity data. Namely, that the dissimilarity judgments really reflect source-model-based factors at least as much as acoustic factors. Perhaps the hypothesis reflected in my approach (that "surface features" of the music are the primary contributor to the immediate perception of short musical stimuli) is completely wrong, and instead listeners are really evaluating similarity based upon their perceptions of the *instruments* being used in the music. This will have to remain a possibility for future work.

Another fruitful approach is to try to model the dimensions returned by the MDS procedure. If successful, this result indicates a possible basis for the judgment of similarity in music. The models of each dimension can serve as hypotheses for future experiments. In the approach I have presented here, there are two possible bases for explanation. The first is the set of semantic judgment ratings obtained in Experiment I. The second is the set of psychoacoustic features that have already been used to model the semantic-judgment and similarity ratings.

Table 7-21 shows the results of using linear multiple regression to model the axes derived in the two-dimensional MDS solution. That is, I take the horizontal position in the MDS solution space of each the 20 stimuli placed in the space as a new feature of each stimulus, and similarly for the vertical position. I then attempt to model these features through linear regression on the semantic judgments, and on the psychoacoustic features.

As seen in this table, the first dimension corresponds very well to the semantic feature LOUD, and therefore also to the signal-processing feature MEANMOD, which was (in Table 7-7) the best predictor of LOUD among the signal-processing features. It is likely that if there were more musical stimuli in Experiment II (and thus more statistical power), the other signal-processing features that predict LOUD and SOOTHING well would become significant predictors of Dimension I. In fact, SPECSTB and BESTT, both of which were predictors of LOUD and SOOTHING in Table 7-8, are the next two predictors approaching the significance level required for entry in this model as well (p = .084 and p = .149 as the second factor for potential entry, respectively).

Neither the semantic-judgment model nor the signal-processing model is able to predict the second axis in the two-dimensional MDS solution. Thus, the existence of a two-dimensional space underlying the judgment of musical similarity must be considered questionable until more data are collected or more features examined.

	Semantic judgments			**Signal-processing features**		
	Full R^2	Features	Stepwise R^2	Full R^2	Features	Stepwise R^2
Dimension I	**.875**	-LOUD	**.804**	.860	-MEANMOD	**.422**
Dimension II	.495	(none)	N/A	.889	(none)	N/A

Table 7-21: Feature models for each of the axes in the two-dimensional MDS solution. Statistically-significant R^2 values are in bold. The "Full R^2" column shows the proportion of variance along each axis that can be explained with the full models. The "Features" and "Stepwise R^2" columns show the order of entry of features entered in a stepwise regression and the proportion of variance explained by each subset of features. Left, model based on the 6 semantic judgments. The first dimension anticorrelates strongly with LOUD, which alone explains 80% of the variance along the first axis. The second dimension cannot be explained with the semantic-judgment ratings. Right, model based on the 16 signal-processing features. Note that the full model does not reach statistical significance in this case (p = 0.479)—this is because explaining 20 data points with 16 features is a very easy task. The first dimension anticorrelates significantly with MEANMOD, which is the only feature entered. The second dimension cannot be explained with the signal-processing features.

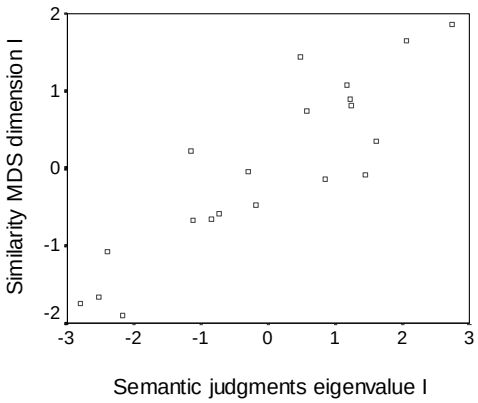


Figure 7-19: Scatterplot of the relationship between the principal eigenvector, determined by factor analysis, of the six-dimensional semantic judgment (Experiment I) and the first axis in the two-dimensional MDS solution for the dissimilarity ratings derived from similarity judgments (Experiment II). There is a very strong ($r(19) = .885$, $p < 0.001$) relationship between these two models of the data, indicating that these data are likely reflecting something about the underlying musical percept.

Convergence between MDS and semantic-judgment factor analysis

Notably, there is a very significant correspondence ($r = .885$, $p < .001$) between the first eigenvector determined by factor analysis of the semantic judgments in Section 7.1.6, and the first axis of the two-dimensional MDS solution. This is a strong result, because it represents a convergence in which two different experimental paradigms give the same answer. A scatterplot of the relationship between these two models is shown in Figure 7-19.

The primary axis (shown as the *x*-axis in Figure 7-19) is readily interpreted as the "quiet-soothing vs. loud-harsh" axis, with all of the noisy and disturbing examples to the left and the soothing and quiet examples to the right. This can be seen in the weights of the factor analysis in Figure 7-10.

There is no such relationship between the second eigenvector and the second MDS axis ($r(19) = -.128$, p = n.s.). This continues the trend that the second axis cannot be not well-connected to the actual behaviors of subjects observed in this experiment, and is consistent with two competing hypotheses. First, that similarity cannot really be approximated by distance in a two-dimensional feature space. Or second, that the psychoacoustic and semantic features spaces used here are not rich enough to model similarity. (The second hypothesis is not inconsistent with the modeling results using the semantic judgments, reported in Table 7-18. The amount of variance explained there is not more than the amount of variance explained by a one-dimensional MDS model, as shown in Figure 7-17).

7.5. Experiment III: Effect of interface

In both Experiment I and Experiment II, there was a consistent bimodal response pattern. I conducted a short post-pilot to investigate whether this response pattern was an artifact of the experimental method, or whether it was actually part of the human perception. In particular, since each experimental trial in Experiments I and II began with the response sliders reset to the center (see Figure 7-1), it is possible that the use of sliders biased the results around the center of the scale.

This experiment tests the hypothesis that the response pattern will be the same even if there are no sliders used. If this hypothesis cannot be rejected, then we would conclude that there was no slider-based bias around the middle of the scale.

Overview of procedure

Five subjects listened to the same 150 musical excerpts used in Experiment I. Experimental procedure was identical to Experiment I, except that the response interface did not include visible sliders on the rating scales.

Subjects

The five subjects were drawn from the MIT community. None had participated in Experiment I or Experiment II. Three were male and two female; two were M0, two M1, and one M2.

Materials

The same experimental stimuli from Experiment I were used.

Detailed procedure

Experimental procedure was the same in all regards as Experiment I, except that the listening interface was as shown in Figure 7-20.

Dependent measures

For each trial, the final ratings on each scale were recorded to a computer file. The computer interface produced a value from 0 (the bottom of the scale) to 100 (the top) for each scale on each trial.

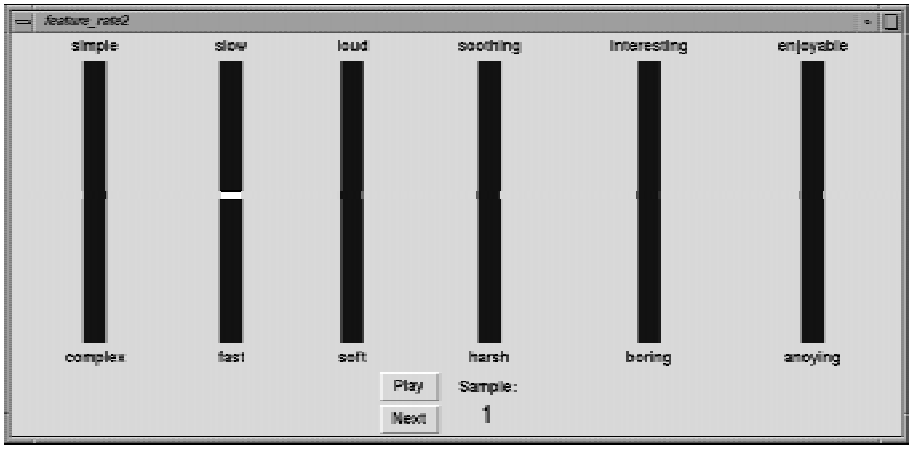


Figure 7-20: The experimental interface for Experiment III (compare Figure 7-1). The response scales did not use sliders; rather, the subject was free to click anywhere within the scale. Upon doing so, an indicator bar (as shown for the **slow-fast** scale) appeared as visual feedback. The subject could click again in a different place to change his/her judgment; final ratings were recorded when the subject clicked **Next**. Clicking **Next** was prohibited until the subject made ratings on all six scales.

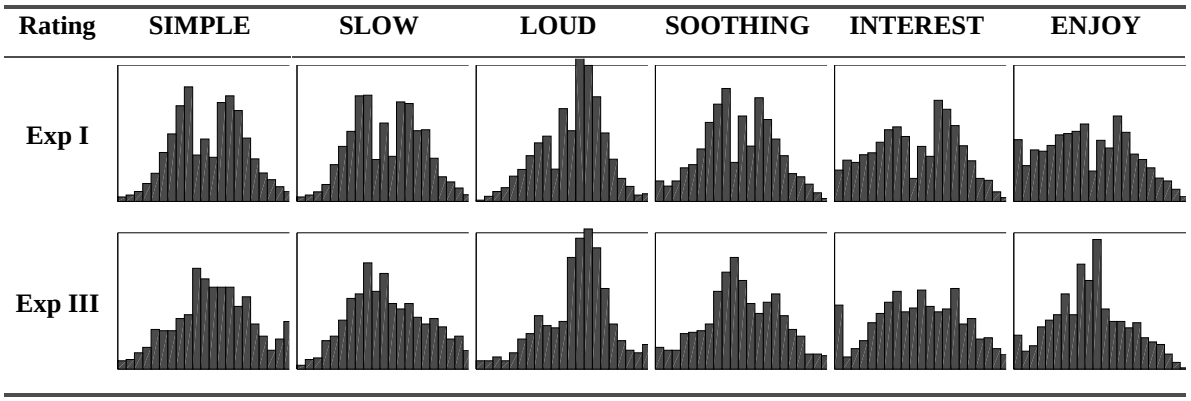


Figure 7-21: Comparison of response distributions between Experiment I and Experiment III. Each panel is a histogram of the unscaled responses pooled across all subjects and stimuli on each of the six rating scales. Responses are taken for 31 subjects in Experiment I, and 5 subjects in Experiment III. At top, in Experiment I, there is a clear response bias against the center of the scale. At bottom, in Experiment III, there is no such bias. (Experiment I results are slightly different than shown in Figure 7-2 because centered responses are not removed here.)

	SIMPLE	SLOW	LOUD	SOOTHING	INTEREST	ENJOY
Unscaled	6.67%	4%	4.67%	4%	6.67%	1.3%
Scaled	3.33%	2%	1.33%	4%	3.33%	2%

Table 7-22: Proportion of significant differences between response means for Experiment I and Experiment III. Each cell shows the proportion of *t*-tests (out of 150 tests, one for each stimulus) reaching significance at $p < 0.05$ for the given rating scale and dependent variable. "Unscaled" data are the values recorded directly by the subjects, as shown in Figure 7-21. "Scaled" data are the transformed values by removing centered responses, recentering the scale, and scaling by the cube root, as described in Section 7.3.6. Note that with $\alpha = 0.05$, by chance 5% of the *t*-tests will reach significance.

Results

Results are shown in Figure 7-21. As seen there, the hypothesis governing this experiment may be rejected immediately—the response patterns are quite different for Experiment I and Experiment III. For each scale, there is a clear bias away from the center of the scale in Experiment I that is not present in Experiment III. This difference must be due to the different response methods used. Other than the anti-center bias, the shapes of the response distributions are very similar.

A more pressing question regards the means of the rescaled distributions, since most of the important results in Section 7.2 (and in Section 7.4) were obtained with these. If the means were significantly different, then this would call the results in these sections into question. To test this, I performed independent-sample *t*-tests comparing the means by subject for each stimulus on Experiment I to the means on Experiment III (note that the number of subjects differs in the two conditions). The results are summarized in Table 7-22 and show that only the proportion of tests expected due to chance variation reach significance. Thus, a weaker version of the Experiment III hypothesis cannot be rejected—namely, that the different forms of rating interface have no effect on the mean responses.

This result still leaves open the possibility that regression modeling of subject-by-subject ratings (which, as discussed in Section 7.2.2, couldn't be conducted on the unscaled data) could be better modeled if data were obtained with the revised experimental interface. Future research examining such questions should not use a slider-based interface.

7.6. General discussion

The experiments and models reported in this section allow us to develop a clearer picture of the early sensory stages of music-listening. The most fundamental result is that computational models of psychoacoustics can be used to predict the behavior of listeners engaged with real musical sounds. It is crucial to understand that I do not claim that no higher-level processing or cognitive structure is involved in music-listening. To do so would be absurd. But the statistical results shown here demonstrate that significant proportions—in the strongest cases, about half—of the variance in human judgments can be explained *without* recourse to cognitive models.

In other words, I have demonstrated that the models presented here suffice to explain significant proportions of the variance in these judgments. The only explanatory space left to cognitive models remains in the residual. Thus, if we accept the traditional precept in comparative psychology holding that all things equal, a sensory explanation is a simpler one than a cognitive explanation, the models I have presented must be considered very attractive.

The modeling and interpretation of the semantic-judgment data is more convincing than for the similarity-matching data. For example, it is notable that while the similarity ratings are as consistent (subject-to-subject) as the "easy" semantic ratings (compare Table 7-3 and Table 7-15), they are as difficult to model as the "difficult" semantic ratings. This is a strong indication that it is not because of intersubject variability that the similarity judgments, at least, are difficult to model. Rather, it seems likely that there are consistent but unmodeled factors as work, such as cognitive structures or instrument identities.

Part of the difficulty is clearly the relatively small number of sounds used in Experiment II. As seen in Figure 7-16, there are very few pairs of sounds that are judged to be very far on the "similar" end of the scale. A comparison with timbre-similarity experiments—for example, (Grey, 1977)—is illustrative here. When studying the similarity of the sounds of musical instruments, we can afford to use perhaps 30 sounds as a test set, since each is very short. This gives 435 pairwise combinations; if each takes 5 sec to rate, then the whole experiment takes about an hour. Further, it is a reasonable working hypothesis that 30 sounds can cover most of the space of pitched orchestral musical instruments, which are the class of sounds that have been studied most in similarity experiments.

However, this is not the case for comparison of full musical excerpts. In retrospect, it is entirely unsurprising that the average similarity is low. I have collected 20 pieces out of the space of all music, and asked whether any two of them are similar. The space of music is far too broad to be effectively sampled in 20 points, and this is naturally reflected in the subjective data. But it seems unlikely that doing the experiment again with 30 samples, or even with the full set of 75 from Experiment I, would provide very much better coverage.

Trying to use more than a hundred samples in a full-pairwise-combination paradigm would be completely impractical. A hundred excerpts yield 4950 pairs, each taking about 20 sec to rate, or more than 20 hours per subject. The amount of time required increases as the square of the size of the sample set. Thus, it seems that this laboratory methodology is unlikely to give very good results in the future. On the other hand, many people listen to at least this much music for recreation in a week—they just don't provide experimental feedback. If it were possible to develop experimental measures that somehow leveraged off of natural listening behavior (which also connects back to the idea of emphasizing ecologically significant behaviors and judgments), then subjects could provide experimental data in their day-to-day listening activities, even outside the laboratory. Perhaps this idea could be made fruitful in collaboration with a corporate partner that wanted to develop musically intelligent agents for practical search-and-retrieval applications.

7.7. Applications

In this section, I will briefly discuss the connections between the modeling results that I have reported in this chapter and three interesting applications for music-listening systems. For the first, I have implemented a simple system as a demonstration and evaluation; for the second and third, I only sketch how such applications could be built.

7.7.1. Music retrieval by example

I will consider a different approach to the problem of similarity-matching one stimulus to another. Rather than using the human-perception data as ground truth, we can make use of the paired excerpts from the stimulus database collected for Experiment I. That is, we can assume a ground-truth in which the two excerpts from the first musical selection belong together, the two from the second belong together, and so forth. Various methods of calculating the distance between two excerpts can be tested to see how often they find that these pairs are close together.

The simplest thing to do is simply to treat feature vectors as points in N-dimensional pattern space, and to see how well various distance metrics work according to this non-perceptual evaluation metric. This is a non-perceptually-based approach because it is not necessarily the case that two segments from the same piece of music will actually sound similar, or have anything in particular to do with each other. However, this is a readily-quantifiable task that may correspond to the way music-listening systems will be deployed in applications in the future.

I used four feature spaces:

(1) a 6-dimensional space using the semantic judgments elicited in Experiment I

(2) a 16-dimensional space using the psychoacoustic features from Chapter 6

(3) a 6-dimensional space using the linear-regression predictions of the semantic judgments. These are the predictions evaluated in Table 7-7.

(4) a 13-dimensional space, for comparison, based on the speech vs. music features reported in (Scheirer and Slaney, 1997).

For each, I tested three distance metrics: Euclidean distance (2-norm), vector correlation, and Mahalanobis distance. For the vector correlation, each component (feature) was normalized by the mean and variance of that component, estimated over the 150 test stimuli. For the Mahalanobis distance, the covariance matrix was estimated as the covariance of the 150 data points in the test set.

	Euclidean distance		Vector correlation		Mahalanobis distance	
	% hits	Avg. rank	% hits	Avg. rank	% hits	Avg. rank
Perceptual judgments	31.3%	20.4	23.3%	26.9	22.0%	28.6
Psychoacoustic features	16.0%	51.7	24.0%	33.5	22.7%	38.2
Predicted judgments	16.0%	40.3	12.7%	44.4	17.3%	40.2
Speech/music features	13.3%	52.4	27.3%	26.5	22.7%	23.6

Table 7-23: Evaluation of non-perceptual music-matching task. Four feature spaces are compared for each of three distance metrics. The "% hits" column indicates the proportion of the time that the counterpart to the target stimulus is one of the five closest stimuli according to the given metric. The "Avg. rank" column indicates the mean position of the counterpart in the rank-ordering of all stimuli to the target, across the 150 test stimuli.

I evaluated these twelve metrics in the following way. For each of the 150 musical stimuli, I calculated the distance from that stimulus (the "target") to each of the other stimuli. Then I ranked the other stimuli in increasing order of distance. I counted a hit for each time that the other excerpt from the same selection (the "counterpart") was one of the five-closest stimuli to the target. I also computed the average rank in the distance table for the counterpart across all 150 examples.

The results are summarized in Table 7-23. Note that random guessing would give 3.3% hits with average rank of 75.

Another way evaluate at performance on this task is to examine the growth in the hit rate as a function of the "hit window" length. That is, if we require that the counterpart be returned in the first 2, or first 20, closest stimuli to the target, how does the result compare? This is plotted for four of the feature space/distance metric combinations, plus the baseline from random guessing, in Figure 7-22.

Not surprisingly, the perceptual judgments from Experiment I are best able to perform this simple task, as shown in Table 7-23 and Figure 7-22. I hypothesize that if Experiment II were conducted on the entire set of 150 stimuli (at great experimental cost), the direct similarity results would give nearly perfect performance on this task. This because listeners would readily identify the paired segments from the same song and give them very high similarity ratings. However, given this hypothesis, it is striking that the perceptual judgments do not perform better than they do. Apparently, either the set of judgments elicited in Experiment I is not rich enough to cover those perceptual features of music that are used to judge similarity, or the two excerpts from each piece of music are not all that similar, on average. (The results shown in Table 7-12 for the "counterpart" features are compatible with the latter hypothesis, as they show that the perceptual features on one excerpt are significantly, but not completely, the same as those on the counterpart.)

The psychoacoustic features perform very well at this task, giving the same hit rate on the normalized metrics as the perceptual judgments do. (Their performance with a simple Euclidean distance is artificially depressed because the features have widely different scaling.) It is not presently known what sort of performance on a task like this would be required to give satisfying interaction with a musical search engine or other application. The predicted judgments perform about half as well as the observed judgments, which makes sense since that the model predicts about half of the variance in these judgments.

The speech/music features also perform well at this task, particularly when vector correlation is used as the distance metric. Likely, this is because these non-perceptual features can pick up information in the musical examples that have to do with the signals themselves. This information is missed by more perceptual approaches. To take one example, some of the

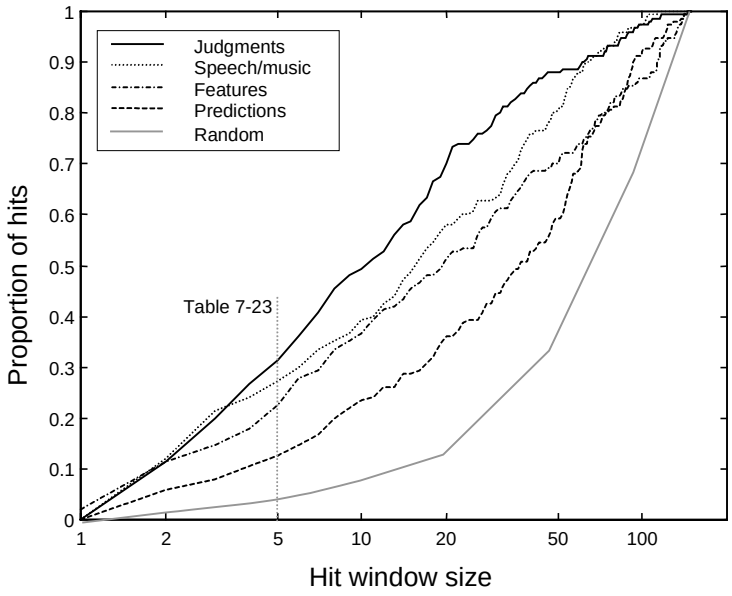


Figure 7-22: Growth in counterpart retrieval rate as a function of *N*, the length of the hit window, averaged over the 150 musical examples. The y-axis shows the proportion of the time that the counterpart was returned as one of the first *N* closest matches. The five curves are, from top to bottom: (1) perceptual judgments with Euclidean distance metric; (2) speech/music features with correlation distance metric; (3) psychoacoustic features with Mahalanobis distance metric; (4) predicted perceptual judgments with correlation distance metric; (5) random guessing. Note that other distance metrics for each feature set give different results (as shown in Table 7-23), and so this figure should not be taken as showing the best possible performance for each feature set.

stimuli were digitized (in the source recordings) at a lower sampling rate that the others. The perceptual features (and, presumably, the human listeners, since results were not noticeably different for these stimuli on average) do not pay attention to this. But a simple measure of signal bandwidth then has an advantage in trying to match up the "two halves" of a single musical example, since both excerpts will have this signal modification.

These results highlight the difference between evaluating performance on perceptually-motivated and non-perceptually tasks. In non-perceptually-motivated tasks, such as this one, there may be simple tricks that give good results due to artifacts in the stimuli. And if the applications we want to build suggest that non-perceptually-motivated evaluation is appropriate, then we should make use of these tricks whenever we can. But when we evaluate a perceptually-motivated task such as modeling human judgments, there are fewer such tricks to use. Thus, in such a case, perceptually-motivated features perform better.

7.7.2. Parsing music into sections

Various music-psychology studies (Clarke and Krumhansl, 1990b; Deliège *et al.*, 1996) have examined the ability of listeners to divide music into sections. This is a different problem than the one that I have been considering, but stands as the most robust study of ecological listening behavior presently in the literature. It seems likely that when music-listeners hear a long piece of music, they do not perceive the music as a single, undifferentiated stream. Rather, they perceive that it is divided into segments of various lengths, with short segments subsumed into longer ones.

Most of these studies have tested hypotheses regarding the use of musical structure to create perceived segmentation. For example, Clarke and Krumhansl used their results to evaluate

the structural model put forward by Lerdahl and Jackendoff (1983). A competing hypothesis, though, is that it is not musical structure at all, but rather changes in the musical surface, such as tempo, orchestration, dynamic, activity, and so forth, that are the primary basis for the perception of musical segmentation. (Naturally, there is a wide middle ground between these endpoints, in which surface perception and structural cognition play shared roles, differing by listener and circumstance, in determining segmentation).

It would be straightforward to apply the psychoacoustic features, which I claim are part of the music surface, to music segmentation. The most straightforward way to do this would be the following two-step process. First, modify the feature extraction algorithms in Chapter 6 so that rather than producing a single result for each feature for a whole piece of music, they produce a time-series of changing feature estimates. This is done simply by calculating the feature values within a short time window. The result is a sequence of feature vectors. Second, use dynamic programming to divide the sequence of vectors into homogenous groups. There are several ways to accomplish this; the basic idea is to find the optimal segment boundaries such that the set of vectors within each segment is as homogenous (in terms of spread, or covariance) as possible.

Once a piece of music is automatically segmented in this way, the result may be compared to human results on the same task, and to the segmentations produced by other automatic methods.

7.7.3. Classifying music by genre

A final application of interest is one in which music is automatically categorized by the musical "genre." The remarkable ability of humans on this task has recently been formally investigated.(Perrott and Gjerdigen, 1999). The division of music into genre categories can be somewhat arbitrary. For example, in the list of musical stimuli in Appendix A, we see that MP3.com includes as part of their ontology the categories "Heavy Metal," "Power Metal," and "Metalcore." It is likely that such fine distinctions are lost except to all but the most discriminating aficionado.

Regardless, the prevalence of the music-genre-categorization in music stores, radio stations, and now on the Internet is reasonably good evidence that categories of musical style are a perceptually useful starting point. The great literature on classification and categorization systems (Therrien, 1989) could be readily applied with the psychoacoustic features in order to develop automated systems. They would have application in music-database systems, and in systems to automatically create playlists for radio or other listening environments.

The approach I used for speech/music discrimination (Scheirer and Slaney, 1997) would be the right starting point. A large database of music with known genre labels collected and used to partition the psychoacoustic-feature space into regions associated with each genre. There are a variety of ways to do this. Then, based on their locations in the feature space according to sound analysis, new songs can be classified according to the similarity to the original prototypes. Such a system could be evaluated according to perceptual criteria (for example, by similarity to the results obtained by Perrott and Gjerdigen) or engineering ones (for example, ability to match ground-truth labels given by the owner of the music in the database.

7.8. Chapter summary

In this chapter, I have presented the results from two pilot experiments studying the human perception of complex musical stimuli. In the first, listeners rated each of 150 short (5-sec) stimuli on 6 semantic scales. In the second, listeners judged the similarity of each of 190 pairs of musical examples. The data collected in these experiments are the first available from a study that compares such a wide range of musical styles directly. Further, the

modeling results presented here are the first time that interesting musical percepts have been connected to psychoacoustic models for real musical stimuli in their full ecological complexity.

I will briefly summarize the major results from this chapter.

(1) Human listeners were able to make consistent judgments regarding the loudness, speed, complexity, soothingness, interestingness, and enjoyability of short musical examples selected at random from a large Internet database. The judgments were consistent across listeners and across multiple excerpts from the same piece of music.

(2) The six semantic judgments elicited were not independent from one another. A two-dimensional basis space derived by factor analysis explains 84% of the covariance among them.

(3) The intersubject means of the semantic judgments can be predicted by a signal-processing-based psychoacoustic model that processes the sound signal and produces 16 features. In linear multiple-regression, such a model explains from 24% to 53% of the variance in the data, depending on the particular feature. This set of perceptually-motivated features explains more (15-40% more) variance than does a simpler set of features whose construction was not perceptually motivated.

(4) The six semantic judgments fall into two categories: those that are easy to model and extremely consistent across subjects—loudness, soothingness, and speed—and those that are difficult to model and only somewhat consistent across subjects—complexity, interestingness, and enjoyability. A natural hypothesis stemming from this is that the first sort of judgment is close to the musical surface, while the second sort involves more cognitive processing.

(5) The directions (up or down) of many of the individual subjects' responses on the semantic scales can be predicted with a logistic-regression model based on the 16 psychoacoustic features. Such a model predicts from 61% to 97% of the bivalent responses correctly, depending on the subject and semantic scale, and can predict the responses of 47% to 98% of the subjects statistically significantly well, depending on the semantic scale. The fact the same ratings are as difficult to predict subject-by-subject as for the pooled means is not consistent with the hypothesis that intersubject differences on such ratings take the form of different weights applied to a single set of features.

(6) Human listeners were able to consistently judge the similarity of pairs of short musical examples. The judgments were consistent across listeners.

(7) The perceptual similarity of pairs of musical examples can be predicted by the differences between the psychoacoustic features of the two examples, and by the differences between the semantic judgments reported for those examples. In linear multiple-regression, such models explain about 25% of the variance, and about 50% of the variance respectively.

(8) The primary axis produced in a two-dimensional multidimensional-scaling (MDS) analysis of the similarity data corresponds closely to the differences in ratings on perceived loudness and soothingness. Positions of stimuli along this axis can be predicted well by linear regression from the perceptual judgments (88% variance explained) or from one of the psychoacoustic features (44% variance explained). Further, the primary MDS axis corresponds strongly ($r = 0.895$) to the principal eigenvalue derived from factor analysis of the semantic judgments. This is strong converging evidence that a surface feature that is something like soothingness plays an important role in the immediate perception of music.

However, beyond the direct results, the methodological implications are important as well. These experiments, preliminary in nature as they are, tend to bring about more questions than

answers. Certainly, I have shown that such an experimental paradigm works to give data that are useful for further analysis. We are sorely in need of more data to which music signal-processing and pattern-recognition systems can be compared. It is essential that more experiments of this sort be run, more rigorously and wider in scope where possible.

The most obvious methodological lack here is a sophisticated treatment of individual differences. Where possible (in Sections 7.2.2 and 7.4.4 in particular), I have tried to do simple statistical analysis to discover trends in individual data, but sorely missing are any good independent variables that might relate to individual differences. Particularly when thinking from an applications standpoint (the notion of "musically intelligent agents"), it is essential that our systems be able to induce and make use of the vast differences between listeners.

The prospect of dividing all listeners into three musical "classes" (my M0, M1, M2 groups) and thereby learning something about their preferences is absurdly limiting; however, I note that this is one *more* class than most other studies have considered when they consider individual differences at all. A truly ecological psychology of music-listening would have to be able to account for the evolving musical preferences and behaviors of a given human over time. We are still quite some distance from such a goal!

CHAPTER 8 CONCLUSION

My dissertation has presented a variety of signal-processing techniques, methodological stances, and psychoacoustic and music-perception theories that I loosely connect together under the single term *music-listening systems*. In this chapter, I will summarize the material, first as a high-level story that shows where all the pieces fit, and then in a more detailed itemization of the specific contributions I believe that I have made. In many ways, I view the questions that I am asking and other theoretical contributions to be more important than the practical results themselves. A lengthy discussion of future directions that warrant more study therefore concludes the chapter and the dissertation.

8.1. Summary of results

As outlined back in Chapter 1, this is a dissertation that rests on three main disclipinary pillars: the studies of music psychology, psychoacoustics, and music signal processing. The fundamental result that I have presented is a demonstration that it is possible to connect these approaches to the study of sound. I have constructed signal-processing algorithms that analyze complex musical sounds, and argued that these algorithms are reasonably taken as models of the psychoacoustic behaviors they target. Further, I have demonstrated through human listening experiments that interesting high-level musical behaviors can be partly explained with computer programs based on these algorithms.

I believe that this is the first research to show how a theory of music perception can be grounded in a rigorous way on a compatible theory of psychoacoustics. It is also the first research to show computer modeling of such interesting human musical percepts directly from audio signals. Finally, it is the first research that explores psychoacoustic modeling of such complex sound signals. Thus, the connections between these fields have been advanced, in some cases a small amount and in some cases a larger amount, by my research.

The reason that I have successfully made these connections is in large part due to the novel methodological approaches I have presented. By considering a broad range of musical stimuli and a broad subject population, I believe that my work has broader relevance to real human music-listening behavior than do many other theories of music perception. Also, the approach that I call *understanding without separation* is an important viewpoint, both for theories of psychoacoustics and auditory scene analysis, and for the practical construction of computer systems for working with musical sounds.

These ideas are naturally connected: it is only with the separationless approach that it is possible to consider analysis of such a broad range of ecological stimuli as I do here. There are few, if any, other results reported that attempt to admit the whole of music for computational or psychological study, and this is because previous approaches have depended too much on assumptions that do not scale well.

8.2. Contributions

In this section, I briefly summarize the contributions to the research literature that I have made in my dissertation.

Music signal-processing and pattern recognition

I have developed two new signal-processing approaches to the computational analysis of complex musical signals. The first, reported in Chapter 4, extracts the tempo and beat of musical signals. I have demonstrated that the results of this extraction are similar to the perceptions of human listeners on a variety of ecological sounds. This model performs more robustly on a wider variety of signals than other systems reported in the literature. The second, reported in Chapter 5, uses the principle of dynamic detection of comodulation among subbands in the autocorrelogram domain to allocate energy from auditory scenes to auditory images for analysis. I have demonstrated that this model can be used to explain the percepts of a number of important psychoacoustic stimuli.

As well as the processing principle in the auditory-image-formation model, I have developed a new pattern-recognition framework. This framework performs unsupervised clustering of nonstationary data given certain constraints that apply from one point in time to the next. It accepts as input an unstructured sequence of input data in feature space and time, which corresponds to modulation features extracted from each cochlear channel. From this, it dynamically estimates the number of clusters, their positions in feature space, and the assignment of cochlear channels to images over time. Although the development of this method must be considered preliminary (I have not done any principled testing of this framework in its own right outside of its application to the particular musical systems considered here), there are relatively few approaches to such non-stationary data in the pattern-recognition literature.

I have also developed simple feature extractors, described in Chapter 6, that apply to the output of the tempo and image-formation models. I have shown in Chapter 7 that these features can be used to explain the immediate perception of musical sounds. They can be used as the basis for modeling high-level semantic judgments on several perceptual scales, and for modeling (less well) the perception of musical similarity among short stimuli.

Psychoacoustics

The models reported in Chapters 4 and 5 can be taken not only as pure signal-processing approaches to the study of sound, but as psychoacoustic theories of sound perception. In both cases, I have related the construction of these models to the existing scientific discourse on the perception of complex sounds. In particular, I have shown that the model I described for tempo and beat analysis of acoustic signals bears much similarity to existing subband-periodicity models of pitch perception.

This observation brings about the question of whether pitch and tempo could be related perceptual phenomena, which in turn relates to the general discourse on the perception of modulation in the auditory system. I believe that it is likely that both pitch and tempo perception will be understood in the future as simply two particular special cases of a general modulation-detection ability. However, this is pure speculation at the present time, and while

the results in Chapters 4 and 5 are consistent with this hypothesis, taken alone they are rather weak evidence.

The auditory-image-formation model presented in Chapter 5 is also based on the subband-periodicity model of pitch. It can be seen as a bridge between, on one hand, static models of the pitch-perception process, and on the other, dynamic models suitable for explaining auditory scene analysis. It is the first model capable of explaining, based on acoustic input, the wide range of psychoacoustic and auditory-scene-analysis phenomena that it does. To be sure, there is a great deal more work required to evaluate and study the theoretical implications of this model. However, at the least it shows potential to become part of a theory of the perception of complex sound scenes. It is one step (how large a step remains to be seen) along the difficult path of developing a complete psychoacoustics of real-world ecological sounds.

Finally, through the modeling results reported in Chapter 7, I have made connections between the physical, acoustic nature of musical sounds, and the high-level percepts that they elicit. This can be seen as the study of the physical correlates of new kinds of sensory phenomena (such as musical similarity and semantic feature judgment) that have not been previously studied.

For the time being, the sorts of features extracted in Chapter 6 are too far removed from theories of hearing to be considered as contributions to psychoacoustics per se.

Music psychology

In chapter 7, I have reported experimental results for two new music-listening experiments. The results are consistent with two new hypotheses about music perception. First, that human listeners are able to make rapid, immediate judgments about musical sounds from relatively short stimuli. Such judgments include both what I term *semantic features* of music and also the perception of the similarity of musical segments, and can be consistently modeled across listeners and across stimuli. Second, that there are important aspects of music-listening that are sensory in nature. That is, a large proportion of the variance in the results elicited in these experiments can be explained from a model based only upon sensory features, that does not include any modeling of musical structure.

The perceptual models that I have developed bear a closer connection to the musical signal that most other models of music perception that have been previously reported. I regard the development of connections between the psychophysics of sound perception and the formation of immediate musical judgments as the strongest theoretical contribution of this dissertation.

Philosophically, I have contrasted two sorts of models of the music-listening process. The first, most common in the literature, is the structuralist approach to music perception, in which perceptual judgments about music are explained as stemming from the structural relationships among mental representations of musical entities. The second, developed for the first time here, is the sensory approach to music perception, in which perceptual judgments about music are explained as the direct and immediate results of low-level processing in the auditory system.

Of course it may be the case that human listeners use both sorts of models to some degree. However, I claim that the second model is now the one that has been more robustly connected to the hearing process and the acoustic signal. At least for the time being, as there are no structuralist models of the formation of perceptual judgments about music from the acoustic signal, and as the acoustic signal cannot be ignored in a rigorous theory of music perception, I submit that the burden of proof must now shift to those who claim that the basis of music perception is fundamentally based upon mental structures built out of symbolic elements.

Music-listening systems

The modeling and perceptual results that I have reported and discussed can be used as the basis for constructing machine models of the perception of music. As I outlined briefly at the end of Chapter 7, it would be quite straightforward to use these sorts of perceptual models to build computer systems that can perform useful musical tasks such as enabling content-based retrieval of music on the Internet, and the segmentation of music into sections. These are the first music-listening systems in the literature that can perform these tasks across such a broad range of musical signals, and with such demonstrated connection to human listeners on these tasks.

Methodological issues

Particularly in Chapter 3, I articulated several ideas about approaching the study of music that are relatively new. Theories and methodologies can never be completely novel in the way that working computer programs can be novel, but I have tried to argue more explicitly on behalf of three approaches than have previous researchers.

I use the phrase *understanding without separation* to refer to the idea that it is possible (and usually desirable) to construct theories and computer models that analyze sounds holistically and directly, without first separating them into notes, voices, tracks, or other entities. This is scientifically appropriate, because the human listener does not separate the sound in this sense. It is also practically useful, because the effort required to build sound-separation systems sits as an unnecessary barrier to the construction of practical tools that might otherwise be straightforward to build. This approach stands in contrast to the majority of music-perception theories that assume that music is first parsed into elements by some unspecific auditory agency, and only later perceived.

I have also argued for a broader and more generally inclusive approach in the music sciences. This goes both for the construction of musical-signal-processing algorithms and for the development of theories of music perception. The former typically suffer from either or both of two problems. First, they are only constructed in reference to restricted classes of signals (based on timbre, degree of polyphony, and so on) and their scalability to other sorts of signals is limited. Second, as the algorithms are built, they are highly tuned to a few select "test" signals and poorly evaluated for a broader range of cases, even within the restricted classes that they target.

To be sure, there is value in working very hard to see just how much information can be extracted, or how sophisticated a system build, in restricted cases. But this should not be the main thrust of the field, only one approach to be contrasted with a more inclusive approach to the world of audio signals and musical styles. For the models that I have built, particularly in Chapters 4, 6, and 7, the domain of music signals that is supposed to be admitted is, simply, all of them. In the formal model evaluations in Chapters 4 and 7, I have created databases that I claim are truly representative of the musical world and used them for evaluation.

The lack of attention to music other than "the classics" in the formal study of music psychology is a more serious problem. It is indefensible that almost no research has focused on the actual listening behaviors of actual music listeners, using the kinds of musical materials that are most relevant to them as individuals. Music psychology, as a scientific inquiry, has spent far too long as a sort of offshoot of aesthetic approaches to music, with its primary goal the justification of one interpretative stance or another.

The argument that is most easily brought against this viewpoint is that the scant time and energy of researchers should not be wasted on the study of inferior music, because music can be such a rich and subtle window on the highest human emotions. While the latter point is beyond question, the fact is that music is also a rich and subtle window into the quotidian emotions. If we wish to include musical study as part of a serious science of human behavior, we should begin with the sorts of musical behaviors that occur most often and in the greatest

number of people and proceed from there. To do otherwise is to inappropriately privilege the judgments and emotions of a specially-selected non-representative group that comprises a vanishingly small proportion of the human population.

8.3. Future work

It goes without saying that this dissertation presents only a very preliminary step towards the development of the kinds of systems and theories it treats. All of the results must be evaluated much more extensively, applied to build systems that could be tested with real users, and especially, used to form hypotheses that can be scientifically tested. Each of Chapters 4-7 probably could serve in its own right as a the basis for a dissertation-length investigation of music perception and music signal-processing.

This section explores several of the directions that I think present interesting opportunities for future research. I will proceed generally in order from the most specific and direct to the most large-scale and abstract.

8.3.1. Applications of tempo-tracking

Many of the individual signal-processing tools that I have developed could be incorporated into interesting musical applications in their own right, independent of their utility in the overall context of music-listening systems. For example, it would be possible to use the tempo-tracker presented in Chapter 4 as the basis for a number of interesting multimedia systems. Since the model works predictively in real-time, it would be an appropriate basis for scheduling dynamic computer-graphics displays to be synchronized with unknown music (as Goto (1999) reports for his beat-tracking work).

Also, any method of finding structure in musical signals can then be used to allow musicians to *manipulate* structure in a composition tool. For example, the automatically-determined beat locations can be used for automatic synchronization and mixing of two separate musical tracks, or for otherwise manipulating the rhythmic properties of existing musical signals. Unpublished tempo-tracking work by Jean Laroche (personal communication) has been used in this manner in a commercially-available digital sampling tool.

Finally, the tempo itself is a useful piece of musical *metadata*. Many researchers feel that there will be a growing importance on the extraction, representation, and application of metadata for all kinds of audiovisual data in coming years. As standards efforts such as MPEG-7 (Pereira and Koenen, 2000) evolve, it is natural to imagine including the isolated feature extractors developed here into an automated tool for multimedia annotation in one or more standardized formats.

8.3.2. Applications of music-listening systems

Section 7.7 briefly presented the simplest form of three applications for music-listening systems: music retrieval, music segmentation, and music classification. Naturally, there are many possible ways to extend these systems and develop others. The prospect of creating *musically intelligent agents* that can search through databases of music on the Internet or in a music library, by listening to the sound and making human-like decisions under the guidance of a human supervisor, is an exciting one both for researchers and for the development of practical applications.

To move in this direction will require both continuing study of human music perceptions and how they are applied in music-search tasks, and also a more sophisticated focus on issues of usability, customizability, user preference, and so forth. To take one simple example, it seems unlikely that a single similarity/dissimilarity metric will suffice for all users. Different music-

listeners use different criteria to decide what makes two pieces of music similar or not. (The results from Experiment II, Section 7.3, are compatible with this hypothesis, but of course do not prove it in any rigorous sense). Thus, to build systems that are useful in practice will require the creation of user models that guide the different sorts of similarity measurements that will be needed.

This point suggests that a useful direction of progress would be to try to merge the concept of *collaborative filtering* (Maes *et al.*, 1997), which has been previously used for music recommendation, with music-listening systems of the sort I have demonstrated. The collaborative-filtering component would learn about user preferences and figure out how to best adapt the similarity model to a particular user's needs, and then the music-listening component would implement that similarity model in relationship to a large database of music.

8.3.3. Continued evaluation of image-formation model

The model that I presented in Chapter 5 is appealing because of its apparent ability to connect several different psychoacoustic results. However, it needs a great deal more evaluation in order to stand as a robust contribution. The set of stimuli that I have presented here was, naturally, chosen because the results are positive. For a more thorough evaluation, more types of stimuli, and more stimuli of each type, should be tested. Many of the psychoacoustic effects described in Chapter 5 are known to be robust within a large set of slightly-different stimuli; it is crucial to test whether the model is as well.

In the long run, it is important to evaluate the model competitively against other models that attempt to cover a similar area. This is a difficult process, simple for the practical reason that as models become large it is difficult to make them work well. It is enough of a challenge to get one model working properly, let along to try to test several models on the same set of stimuli.

The fundamental scientific purpose of model is to make us think of new hypotheses to test with experiments. As the image-segmentation model I have presented matures, and our experimental techniques for dealing with complex stimuli do likewise, it seems likely that numerous experiments could be devised to test various new predictions that the model can make. In particular, I have worked from a number of unexamined assumptions regarding modulation processing that could probably be evaluated experimentally immediately.

From an engineering point of view, the model contains several subparts. The two main ones are the periodicity analysis technique and the dynamic-programming technique for estimating the number of images in the scene. It would certainly be possible to remove these pieces from the large model and use them, together or separately, in other psychoacoustic and pattern-recognition systems.

8.3.4. Experimental methodology

The experimental results presented in Chapter 7 are only preliminary. It is unfortunately the case that the methodological approach that I have taken here is somewhat naïve and probably needs to be reconsidered entirely. I can only plead that this is the first time the particular questions addressed here have been considered experimentally, and so a study that is somewhat exploratory is warranted, even if it is not completely desirable. Future research by more skilled experimentalists will help to refine the simple understanding of the sorts of percepts I have considered.

It is impossible to determine from only the results in Sections 7.1.6 and 7.3.6 what aspects of the subjects' ratings are actually response biases stemming from the experimental interface used. For example, the bimodal response pattern consistently observed seen raises the question whether the underlying percepts are actually bimodal (or even categorical) in nature,

or whether the slider-rating method lends a bias in this direction. Similarly, it is also possible that the judgment-to-judgment correlations in the data are artifacts stemming from the multiple-slider interface. Perhaps if subjects made each judgment separately, the correlations would vanish or be reduced.

The consistent response variances due to demographic variables (age, sex, musical ability, native language, self-reported absolute pitch) reported in Section 7.1.6 are difficult to interpret and a bit disturbing. Personally, I don't believe that these are real effects of demographics, but are actually reflecting some unaddressed covariate of individual differences (about which see below)—perhaps musical subculture or daily listening behavior. Nonetheless, the present data do have strongly significant variances of this sort, and so it would be interesting to see if a more well-principled experimental study could replicate them with a larger and more diverse subject population and, ideally, interpret them in terms of music-listening behavior.

8.3.5. Data modeling and individual differences

A second advantage of using larger sample sizes, both in terms of the number of subjects and the number of test stimuli, would be the possibility of exploring more-sophisticated data-modeling techniques. The multiple-regression framework that I have used here has the positive attribute of being relatively simple to implement and evaluate, but of course can only explain linear relationships among the data. A trained neural-network classifier or other non-parametric scaling technique could be used to search for more complex relationships among hypothesized sensory features and perceptual judgments.

But to do this, more training data are required. There is not enough information in the present data set to fully fill out all of the degrees of freedom in a neural network with hidden layers. Collection of human-response data for training and modeling was the most time-consuming part of the modeling research reported here. One possibility for collecting training data would be to include the collection of responses in a deployed music-retrieval system on the Internet. That is, as well as providing content-based music indexing services to users, the system would collect some judgments from the listeners in order to build a database of training data and user information.

It is likely that in this scenario, listeners would not be willing to sit for an hour or two hours apiece (as did the subjects in this research), so a more complex model to induce underlying perceptions from limited data would be required.

8.3.6. Integrating sensory and symbolic models

The long-term goal of which this dissertation forms a part is to try to understand what aspects of music perception can be best understood through sensory processing, and what aspects can be best understood through symbolic processing. To the extent that the research I have presented can be viewed as grounding the sensory study of real music in a coherent starting point, it is now possible to imagine trying to interrelate these topics. Although it is only possible to speculate on the form of such theories and models, I believe that this is the correct direction for contemporary research on music perception—to focus on both the signal-based sensory aspects and high-level cognitive aspects, while constantly maintaining connections between them.

APPENDIX A: MUSICAL STIMULI

The table contained in this Appendix shows the set of 75 musical examples from which the 150 stimuli were drawn. Two segments, each 5 sec long, were selected from each example: the first starting at 60 sec into the song, and the second starting at 120 sec into the example. All of the stimuli can be heard in MP3 and WAV format from my website at http://sound.media.mit.edu/~eds/thesis/ .

For each example, the title, performing artist or group, and genre is listed, as well as the URL at which the entire performance can be downloaded from MP3.com, and whether the excerpt has vocals or not. The musical genres are those under which the songs are listed on MP3.com—they were originally provided by the artists and so may or may not correspond precisely to the genres that listeners would associate with the songs. I provide them so that readers may observe the diversity of genres represented.

Song	Title	Artist	URL	Genre	Vocals?
1	Lascia chĭo Pianga	Luv Connection	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQmxrBQDABG5vcm3CBGV4dHJD0vyUOCqzl.ei8rdfz_IMTY3pQhc-/lascia_chio_pianga.mp3	Europop	Yes
2	Competition Orange	aka Pawn	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQoxPBADABG5vcm3CBGV4dHJDuwaOONJSH10_ZepGJO2Wv3VPLr0-/competition_orange.mp3	Metalcore	Yes
3	Love that you need	Phantoms in Orbit	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQusWBADABG5vcm3CBGV4dHJDeuiNOEpTIJlVB1muJGJtH_I9_1c-/love_that_you_need.mp3	Alternative	Yes
4	A secred faoth in d Mol	Slow Motion	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQhZ7BQDABG5vcm3CBGV4dHJDTgyOOAM48AheHyMV.lOZiNzlU24-/a_secred_faoth_in_d_mol.mp3	Goth	No
5	Slip and Slide	Hurler	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQtG1BQDABG5vcm3CBGV4dHJDyWKOOAq_hbT_9UEfRgCMD0107NI-/slip_and_slide.mp3	Acoustic rock	Yes
6	"Brothers, Blood, and Bone"	Dan Treanor	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQm.dAQDABG5vcm3CBGV4dHJD9..NOODArymmGk23xVrzCGE1ihI-/brothers_blood_and_bone.mp3	Electric blues	Yes
7	I worship the ground that you	Lee Harris	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQvXUAADABG5vcm3CBGV4dHJDfviTOHRRd.Tb_qWtf7re70PPGsg-/i_worship_the_ground_that_.mp3	Power pop	No
8	Thanks	Danel	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQivlBQDABG5vcm3CBGV4dHJDjcePOO8_pqj0Qw92Alj1	Spiritual pop	Yes

			N8yZCSw-/thanks.mp3		
9	Hot Tommy	Super Love Master Force	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQhLtAQDABG5vcm3CBGV4dHJDo_KNOKDqJ2jvcX7DwaaBkYFnK0Q-/hot_tommy.mp3	Power metal	Yes
10	The Bear	The Swinging Hemphills	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQj3KAgDABG5vcm3CBGV4dHJDT_uNOAnmWLvrFvAcSQeGAeTVEPg-/the_bear.mp3	Political humor	Yes
11	Rossini Intro/Theme	Cantus	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQh8HBADABG5vcm3CBGV4dHJDvQOOOEO0mul54ra6dbxR015X7TY-/rossinis_introduction_t.mp3	Classical	No
12	Marvin	John Martinez	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQrN5AADABG5vcm3CBGV4dHJDt.GNOAV5eG2tfbfirxZ9odHJuNM-/marvin1.mp3	Smooth jazz	No
13	Inside your Heart	Hypnofunk	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQhV.AwDABG5vcm3CBGV4dHJDyemNOD1h5PUzJB3aBSsbKuNKYMg-/inside_your_heart1.mp3	Jazz	Yes
14	Robot Love	The Fuzzy Bunnies	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQhLxAwDABG5vcm3CBGV4dHJDLV6YOHitHHbXk_QyWdb5d9IO6es-/robot_love.mp3	Alternative	Yes
15	Better off Dead	The Snapdragons	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQkArAwDABG5vcm3CBGV4dHJD2oeQOI8qvAGBPi6GwSW0RD2wLzc-/better_off_dead.mp3	Rock	No
16	Still there's something	Kerry Lee	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQpPBAADABG5vcm3CBGV4dHJD8.WNOHOCB31UygVGA3JjJnMBN00-/still_theres_something1.mp3	Folk Rock	Yes
17	Imaginary Conversations	Lou Hevly	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQtSTAwDABG5vcm3CBGV4dHJDk_.NOLpOXlJOQkKGp2fzmkTJEy0-/imaginary_conversations.mp3	Americana	Yes
18	All About Nothing	Satyriasis	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQsQ4BgDABG5vcm3CBGV4dHJDr32UOPKB.T3cGuqw3RMSnnwb4JY-/all_about_nothing.mp3	Psychedelic Rock	Yes
19	Silver Wings	M-word	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQkaLBgDABG5vcm3CBGV4dHJDMDGROIt99fN8d4bGdocJUCZsIq0-/silver_wings.mp3	Electric Blues	No
20	The Devil is Sitting in my Fav	Nick Becker	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQpUzAgDABG5vcm3CBGV4dHJDIfWNOIkHoZeXCqo6GfCdiYodwzU-/the_devil_is_sitting_in_m1.mp3	Indie	No
21	Tuanis	Letho	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQvcsAwDABG5vcm3CBGV4dHJDLPKNOEHfIzGJPpUrbifjQ_Dt4jk-/tuanis.mp3	Reggae	Yes
22	Mean Mean Woman	Harpin Tracy Herron	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQp8zBADABG5vcm3CBGV4dHJDlgWOOLEN_MIbMQ1zl2nzVgKzW4w-/mean_woman.mp3	Blues	Yes
23	I Need Love	DFC (Da Funky Clowns)	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQtZ7AwDABG5vcm3CBGV4dHJD3gCOOMZATdLq92IXzV3yaDdoCSw-/i_need_love.mp3	R & B	Yes
24	I Won't Tell You	Swine Cadillac	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQhDLBQDABG5vcm3CBGV4dHJDhQ6OOCkuzDU3UG5HvsnmdW1fl4I-/i_wont_tell_you.mp3	Electric Blues	No
25	Baby Be Mine Forever	Gene Dawson	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQiaqAwDABG5vcm3CBGV4dHJDfD2ROMReKeeo.sgXJJTZNBCuJW0-/baby_be_mine_forever1.mp3	Blues	No
26	I've Changed	Energy	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQlYYBgDABG5vcm3CBGV4dHJDbwKOOGc2oaoMkqJrSs37XrH78CA-/ive_changed.mp3	R & B	Yes
27	Pasaje de Ida	Havana Clowns	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQrR7BgDABG5vcm3CBGV4dHJDkfSNOAS2aLH9TTIcc5.J09ipZww-/pasaje_de_ida.mp3	Cuban	Yes
28	Leary	Born Naked	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQjmcAADABG5vcm3CBGV4dHJDy.ONON6dVQQFN7wQWI6lLuhnvaw-/Leary.mp3	Heavy Metal	Yes
29	For you	Peter Seltzer	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQjB5BgDABG5vcm3CBGV4dHJDBhGOOBZaRAHUlOwahA5CP7FogHs-/for_you.mp3	Mood Music	Yes
30	Where do the Cowboys Ride	Max's Supper	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQjEpBADABG5vcm3CBGV4dHJDUfeNOPtKA.wLEirqjH0gSLZcezY-/where_do_the_cowboys_ride.mp3	Rock	Yes

31	Miss match	Kirawarebito	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQiHNAQDABG5vcm3CBGV4dHJDJvGNOCi7JDBZHXxvCq1S1KpblRg-/kirawarebito.mp3	Symphonic	Yes
32	Lost at Sea	RusTnale	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQkYnBQDABG5vcm3CBGV4dHJDCmuYOELDWaccNZAZn9UPKdahfyA-/lost_at_sea.mp3	Rock	No
33	I am a DJ (You are here to please me)	Vampirus Sceleratum	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQie1BADABG5vcm3CBGV4dHJDqwiOOOaXu6LRVf5glw1XsjmulsI-/i_am_a_dj_you_are_here_to_.mp3	Electronica	No
34	Mission to Earth	Futurtek industrie	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQoB4AgDABG5vcm3CBGV4dHJDB_iNOFQjAouIRz6lCWDVeN7PeV0-/mission_to_earth.mp3	Electronica	No
35	Payday	Bernie Stocks	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQo1.AADABG5vcm3CBGV4dHJDO.KNOC9WMeZYUzWR_wPl0WPd0vI-/payday.mp3	Folk	No
36	Wallace Lake	Freakishly Big	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQhHrAgDABG5vcm3CBGV4dHJDIvuNOGFbkQ_hupTzgyDoWRtkaJw-/wallace_lake.mp3	Mood Music	No
37	In Orbit	electronatomic	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQqHhBQDABG5vcm3CBGV4dHJDng2OODFoXMa.6U1MxzlzHH_.R64-/in_orbit.mp3	Industrial Electronic	No
38	No. Five	Spiral Motion	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQkeYAwDABG5vcm3CBGV4dHJDXgGOOLEUwqBi9aHO5eCv9BU2Nxs-/no_five.mp3	Pop	Yes
39	Dinero	Bomba de Tiempo	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQqxWAADABG5vcm3CBGV4dHJDRN.NOJBmqpT6BsUe5T8Rr5dnmnU-/dinero.mp3	Rock en Español	Yes
40	Corruptor	Mainstay	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQivbBADABG5vcm3CBGV4dHJDMAqOOABjXuVqPkDWFJH6boAIU.I-/corruptor.mp3	Rapcore	Yes
41	Chicken Coop	W.A.V.E. Compilation	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQrjkBQDABG5vcm3CBGV4dHJDlw2OOE2hscjvtg1wc6VZ7v9VhWs-/chicken_coop_by_biffy_perd.mp3	Folk Punk	Yes
42	I Love You Googleplex	Googol Press	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQrZNAQDABG5vcm3CBGV4dHJD0OyNOLonxY4CcgEvmL.s7iKfWak-/i_love_you_googolplex.mp3	Pop Vocal	Yes
43	South Presa Man	Los #3 Dinners	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQuesBQDABG5vcm3CBGV4dHJDlvSNOPZe9TYJa9tQRR8K8nmfQZ0-/south_pressa_man.mp3	Reggae	Yes
44	Stereo Crush	The Bogs	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQlV6AQDABG5vcm3CBGV4dHJDfe6NOIYzhdeApk435qWiR.QdFus-/the_bogs.mp3	Alternative	Yes
45	Winters Morning	Worlds Apart	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQj.8AgDABG5vcm3CBGV4dHJD_qWSOB9TTzwC0Nb7SG_VlQaQOaA-/winters_morning.mp3	Progressive Rock	Yes
46	Sunday Morning	imagineering	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQsQmBQDABG5vcm3CBGV4dHJDgAaOOEeLnE3DNsxidLyMAdefS6I-/sunday_morning.mp3	Classical Guitar	No
47	Golden Bird	Dreamweaver	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQjxUBQDABG5vcm3CBGV4dHJDCYePOLgKw5l8f0DNNAR85YVQCXc-/golden_bird.mp3	New age	No
48	I'll be there	Simply Blue	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQrBaAgDABG5vcm3CBGV4dHJDfPWNOFzXEcl5FhLrXeFYN5IEyqY-/ill_be_there.mp3	Gospel	Yes
49	Sospeso	Mauvaise	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQud4BQDABG5vcm3CBGV4dHJDGw2OOERRKsNNb3U4mNhAFPwcBk8-/sospeso3.mp3	Grunge	Yes
50	I Can't Take it Anymore	Nitrous	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQv.8AgDABG5vcm3CBGV4dHJDrfqNOAd4Uq4gWukp0.5ZbGs68do-/i_cant_take_it_anymore.mp3	Folk Punk	Yes
51	The Minimalist Experimentalist	Forsaken Lemonade	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQolrAQDABG5vcm3CBGV4dHJDCOqNOB0Y15xrghWaGqRWp2fZxPo-/the_minimalist_experimenta.mp3	Noise	No
52	Why	Dahminionz	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQttKAgDABG5vcm3CBGV4dHJDWvCNOKsOIwP1HrRJUzxml18hgGg-/why.mp3	Hip-hop	Yes
53	Illegal Milkshake	Liquid Evergreen	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQvdcAwDABG5vcm3CBGV4dHJDD.2VOP4KF2mQgn_Itn	Tropical	No

			nzbhdqGIU-/illegal_milkshake_live.mp3		
54	Save	Jonnie Axtell	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQuR6BQDABG5vcm1QBAAAAFLWPgAAUQIAAABDvxPOOCPjWmwWqrWX.UxLjRhuHfg-/save.mp3	Rock	No
55	Bug	Ashtray Babyhead	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQuuEBgDABG5vcm3CBGV4dHJDLhGOOIuHZ.ikamoQEbE0m6Avhmo-/bug.mp3	Power Pop	Yes
56	Diva	Miles Blacklove	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQkQ2AQDABG5vcm3CBGV4dHJDs.uNOBe_xXCmprye6jUnw_AzxGk-/diva1.mp3	West Coast Hip-hop	No
57	FunkRide	Foul Playaz	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQptyBADABG5vcm3CBGV4dHJDsAeOONGqyIWWAHJxS_GmvUqSC6c-/funkride.mp3	R & B	Yes
58	Don't Know Why	Uncle Salty's Cabin	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQmlaAADABG5vcm3CBGV4dHJDGuCNOFDyG0V95WAfca20SrEqaKc-/dont_know_why.mp3	Blues Rock	Yes
59	Acid Blue	Random Axe of Noise	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQhOFAgDABG5vcm3CBGV4dHJDjfiNODU3zW0FpxJpMkzY.PyFa8E-/acid_blue.mp3	Indie	No
60	I Smoke Dope	Antilife	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQkWVBQDABG5vcm3CBGV4dHJDn22QOJMSXaPLYjdrEZnsbD6hN4M-/i_smoke_dope.mp3	Rock	No
61	Laten	Peter Karlsson	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQgf1AwDABG5vcm3CBGV4dHJDBu6WONLwli3xyc1lYODfgYUmhUo-/lten.mp3	Smooth Jazz	No
62	Theif	John Riley	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQmGNAQDABG5vcm3CBGV4dHJDWq.XOOEowwgvexsnmVGw9p9tNVQ-/theif.mp3	Spiritual Rock	Yes
63	Beating	Byt	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQv.sAADABG5vcm3CBGV4dHJDf2OVOG6o_8cySfj9m09EJTr8IWQ-/beating.mp3	Soft rock	No
64	My Girlfriend Died	4GND	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQv_rAwDABG5vcm3CBGV4dHJDfQKOOARim.OOIJvkYMKBhOXVCkg-/my_girlfriend_died.mp3	Hip-hop	Yes
65	Jamm	Akapella	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQl5QAgDABG5vcm3CBGV4dHJDX_aNOD729vH.6ZESB7o4OE0ehBE-/jamm.mp3	East Coast Hip-hop	Yes
66	Asi Estas Tu	Los Pecadores	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQvJbBADABG5vcm3CBGV4dHJDBAaOOJzHdr9bdeWuhO6TN49.0lw-/asi_estas_tu2.mp3	Rock en Español	Yes
67	Blues Town	Carlos T.	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQrXhAADABG5vcm3CBGV4dHJDr.eNOC4MO90Pic2W69kMIwNFzB4-/blues_town.mp3	Blues	Yes
68	Gangsta	Travesty	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQhDQBADABG5vcm3CBGV4dHJDKf2NONjOKuG_owsQ5AgSIAmrukQ-/gangsta.mp3	Hardcore Rap	Yes
69	Quit Stressin'	Oramismo	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQsXuBADABG5vcm3CBGV4dHJDrgqOOA0owhrZFuYbnvxzKbSU1X8-/quit_stressin_rock_steelo_.mp3	R & B	Yes
70	Bush Pianino	Smile Street Noboru	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQvBLAgDABG5vcm3CBGV4dHJDcvWNOEvqtcZgKEQQLZSF7ZoTW8M-/bush_pianino.mp3	Trip-hop	No
71	No Depression in Heaven	Limestone Cowboy	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQugsBQDABG5vcm3CBGV4dHJD6QuOOF9aqhF8dlLQeaYoMWDoF7Q-/no_depression_in_heaven.mp3	Bluegrass	Yes
72	All through loving you	Guy Schwartz & New Jack Hi	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQr0kBgDABG5vcm3CBGV4dHJDAhCOOJqino5qo_4iHpBuLIehLPo-/all_through_loving_you.mp3	Jazz Fusion	No
73	Can't get enough of your love	De'Lane	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQm0sAwDABG5vcm3CBGV4dHJDQu.NONtvnR8phRWURAFs4.Xjv3Y-/cant_get_enough_of_your_.mp3	R & B	Yes
74	Another Lost Cause	Detour	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQskUAgDABG5vcm3CBGV4dHJD7fKNOOu9vHOk3IQ1G5_fKhQysPA-/another_lost_cause1.mp3	Country	Yes
75	Endlessly	Macca Cartny	http://chooser.mp3.com/cgi-bin/play/play.cgi/AAIAQkWyBADABG5vcm3CBGV4dHJDOwmOOIyXH2hXUGB2QUYrVSYA8aw-/endlessly.mp3	Soft rock	Yes

APPENDIX B: SYNTHESIS CODE

This Appendix contains the SAOL code that was used to synthesize the test sounds in Chapter 5. Using this code, the exact sounds can be easily re-created and modified by other researchers. SAOL is the MPEG-4 Structured Audio Orchestra Language (Scheirer and Vercoe, 1999); several software tools are available for creating sound from SAOL descriptions. See the MPEG-4 Structured Audio homepage at http://sound.media.mit.edu/mpeg4 for more information.

For each example, an *orchestra* and one or more *scores* is shown. The orchestra describes the particular synthesis techniques that are used to create the sound. The score describes the way in which those synthesis techniques are used to generate a particular sound.

B.1. McAdams oboe

The "McAdams oboe" sound is used in Section 2.1.1 as a test sound to illustrate the description of the subband-periodicity model of pitch, and then again in Section 5.4.1 for testing the auditory-image model in Chapter 5. In the latter case, two variants are used, one with wide vibrato and one with narrow vibrato.

Orchestra code

```
global {
  srate 24000;
  krate 1000;
}

instr oboe(f0,t1,td,t2,mod,modf) {
  // play 10 harmonics :
  // play them flat for t1 sec
  // then glide a modulator from 0 to mod in td sec
  // and apply it to the even harmonics
  // hold the maximum modulation for td sec

  table modsine(harm,128,1);
  table oddharm(harm,2048,1,0,1,0,1,0,1,0,1,0,1,0,1,0);
  table evenharm(harm,2048,0,1,0,1,0,1,0,1,0,1,0,1,0,1);
```

```
  ksig m;
  asig odd,even;

  m = koscil(modsine,modf) * kline(0,t1,0,td,mod,t2,mod);

  odd = oscil(oddharm,f0) / 50;
  even = oscil(evenharm,f0 * (1 + m)) / 50;

  output(odd+even);
}
```

Score code (normal)

```
0 oboe 10 220 2 4 4 0.02 5
```

Score code (narrow vibrato)

```
0 oboe 10 220 2 4 4 0.002 5
```

B.2. Temporal coherence threshold

Several alternating-tone-sequence stimuli were used in Section 5.4.2 to examine the auditory-image model's threshold of temporal coherence. These stimuli are similar to those used by Van Noorden (1977) and others in experimental work on this topic. To generate the five stimuli, the same orchestra is used with five different scores.

Orchestra code

```
global {
  srate 24000;
  krate 200;

  table sint(harm,2048,1);
}

instr tone(f0,amp) {
  asig a,env;
  imports table sint;

  env = end_env(0.005);
  a = oscil(sint,f0) * amp * env;
  output(a);
}

aopcode end_env(ivar end) {
  asig ltime;
  asig env;

  env = aline(0,end,1,dur-end*2,1,end,0);

  return(env);
}
```

Score code

Note that the scores for stimuli S1 and S2 seem longer because the tempo is twice as fast.

Stimulus S1	Stimulus S2	Stimulus S3	Stimulus S4	Stimulus S5
0.0 tempo 120	0.0 tempo 120	0.0 tempo 60	0.0 tempo 60	0.0 tempo 60
0.1 tone 0.08 950 0.5	0.1 tone 0.08 500 0.5	0.1 tone 0.04 950 0.5	0.1 tone 0.04 500 0.5	0.1 tone 0.04 200 0.5
0.2 tone 0.08 1000 0.5	0.2 tone 0.08 1000 0.5	0.2 tone 0.04 1000 0.5	0.2 tone 0.04 1000 0.5	0.2 tone 0.04 1000 0.5
0.3 tone 0.08 950 0.5	0.3 tone 0.08 500 0.5	0.3 tone 0.04 950 0.5	0.3 tone 0.04 500 0.5	0.3 tone 0.04 200 0.5
0.5 tone 0.08 950 0.5	0.5 tone 0.08 500 0.5	0.5 tone 0.04 950 0.5	0.5 tone 0.04 500 0.5	0.5 tone 0.04 200 0.5
0.6 tone 0.08 1000 0.5	0.6 tone 0.08 1000 0.5	0.6 tone 0.04 1000 0.5	0.6 tone 0.04 1000 0.5	0.6 tone 0.04 1000 0.5
0.7 tone 0.08 950 0.5	0.7 tone 0.08 500 0.5	0.7 tone 0.04 950 0.5	0.7 tone 0.04 500 0.5	0.7 tone 0.04 200 0.5
0.9 tone 0.08 950 0.5	0.9 tone 0.08 500 0.5	0.9 tone 0.04 950 0.5	0.9 tone 0.04 500 0.5	0.9 tone 0.04 200 0.5
1.0 tone 0.08 1000 0.5	1.0 tone 0.08 1000 0.5	1.0 tone 0.04 1000 0.5	1.0 tone 0.04 1000 0.5	1.0 tone 0.04 1000 0.5
1.1 tone 0.08 950 0.5	1.1 tone 0.08 500 0.5	1.1 tone 0.04 950 0.5	1.1 tone 0.04 500 0.5	1.1 tone 0.04 200 0.5
1.3 tone 0.08 950 0.5	1.3 tone 0.08 500 0.5	1.3 tone 0.04 950 0.5	1.3 tone 0.04 500 0.5	1.3 tone 0.04 200 0.5
1.4 tone 0.08 1000 0.5	1.4 tone 0.08 1000 0.5	1.4 tone 0.04 1000 0.5	1.4 tone 0.04 1000 0.5	1.4 tone 0.04 1000 0.5
1.5 tone 0.08 950 0.5	1.5 tone 0.08 500 0.5	1.5 tone 0.04 950 0.5	1.5 tone 0.04 500 0.5	1.5 tone 0.04 200 0.5
1.7 tone 0.08 950 0.5	1.7 tone 0.08 500 0.5	1.7 tone 0.04 950 0.5	1.7 tone 0.04 500 0.5	1.7 tone 0.04 200 0.5
1.8 tone 0.08 1000 0.5	1.8 tone 0.08 1000 0.5	1.8 tone 0.04 1000 0.5	1.8 tone 0.04 1000 0.5	1.8 tone 0.04 1000 0.5
1.9 tone 0.08 950 0.5	1.9 tone 0.08 500 0.5	1.9 tone 0.04 950 0.5	1.9 tone 0.04 500 0.5	1.9 tone 0.04 200 0.5
2.1 tone 0.08 950 0.5	2.1 tone 0.08 500 0.5			
2.2 tone 0.08 1000 0.5	2.2 tone 0.08 1000 0.5			
2.3 tone 0.08 950 0.5	2.3 tone 0.08 500 0.5			
2.5 tone 0.08 950 0.5	2.5 tone 0.08 500 0.5			
2.6 tone 0.08 1000 0.5	2.6 tone 0.08 1000 0.5			
2.7 tone 0.08 950 0.5	2.7 tone 0.08 500 0.5			
2.9 tone 0.08 950 0.5	2.9 tone 0.08 500 0.5			
3.0 tone 0.08 1000 0.5	3.0 tone 0.08 1000 0.5			
3.1 tone 0.08 950 0.5	3.1 tone 0.08 500 0.5			
3.3 tone 0.08 950 0.5	3.3 tone 0.08 500 0.5			
3.4 tone 0.08 1000 0.5	3.4 tone 0.08 1000 0.5			
3.5 tone 0.08 950 0.5	3.5 tone 0.08 500 0.5			
3.7 tone 0.08 950 0.5	3.7 tone 0.08 500 0.5			
3.8 tone 0.08 1000 0.5	3.8 tone 0.08 1000 0.5			
3.9 tone 0.08 950 0.5	3.9 tone 0.08 500 0.5			

B.3. Alternating wideband and narrowband noise

This sound was used in Section 5.4.3 as a test case for the auditory-image-formation model.

Orchestra code

```
global {
  srate 24000;
  krate 200;
}

instr noise(amp) {
  output(arand(1) * amp);
  }

instr cutnoise(amp) {
  // cut at 2000 Hz -- filter sampled at 24000 Hz
  table cutb(data,-1,0.0114775652507270, -0.0226355072073073, 0.0139639593437336,
0.0139639593437336, -0.0226355072073073, 0.0114775652507270);
  table cuta(data,-1,1.0000000000000000, -4.2574131484303770, 7.6101067304623626, -
7.0921304456448810, 3.4391630697019426, -0.6941141713147412);

  asig a;
  a = arand(1) * amp;
  output(iirt(a,cuta,cutb));
}
```

Score code

```
0.0 tempo 24
0.0 cutnoise 0.1 0.2 2000
0.1 noise 0.1 0.2
0.2 cutnoise 0.1 0.2 2000
0.3 noise 0.1 0.2
0.4 cutnoise 0.1 0.2 2000
0.5 noise 0.1 0.2
0.6 cutnoise 0.1 0.2 2000
0.7 noise 0.1 0.2
0.8 cutnoise 0.1 0.2 2000
0.9 noise 0.1 0.2
1.0 cutnoise 0.1 0.2 2000
1.1 noise 0.1 0.2
1.2 cutnoise 0.1 0.2 2000
1.3 noise 0.1 0.2
1.4 cutnoise 0.1 0.2 2000
1.5 noise 0.1 0.2
1.6 cutnoise 0.1 0.2 2000
1.7 noise 0.1 0.2
1.8 cutnoise 0.1 0.2 2000
1.9 noise 0.1 0.2
```

B.4. Comodulation release from masking

Comodulation release from masking (CMR) is an important phenomenon that is believed to relate to more general mechanisms for auditory processing of spectrotemporally complex signals. The phenomenon was discussed in general in Section 2.1.3, and one particular form was used for evaluating the auditory-image model in Section 5.4.4. There are many other sounds known to evoke similar behavior, but I only use one here for simplicity.

Orchestra code

There are two main instruments in the orchestra: trans_coh_noise(), which generates transposed coherent (TC) noise, and trans_rand_noise(), which generates transposed random (TR) noise.

```
global {
  srate 8000;
  krate 200;

  table sint(harm,2048,1);

// filters are sampled at 8 kHz

// 'low' is a 400 Hz lopass filter
  table lowa(data,-1, 1.00000000000000000, -5.60291532564091100, 13.25857989351412600,
   -16.94974710199009800, 12.34135219199334600, -4.85159245990200900,
   0.80447671437945212);
  table lowb(data,-1, 0.01091590233425196, -0.05556437820252266, 0.12607543471834312,
   -0.16273166077176202, 0.12607543471834309, -0.05556437820252266, 0.01091590233425195);

// 'high' is a 300 Hz hipass filter
  table higha(data,-1, 1.00000000000000000, -5.25090555275225320, 11.67458024025534000, -
   14.06969729423221000, 9.70514541733755460, -3.64078644674440130, 0.58237332335537029);
```

```
  table highb(data,-1, 0.58296186069447542, -3.44548924601284500, 8.53662075613657830, -
    11.34817964791776200, 8.53662075613658010, -3.44548924601284590,
    0.58296186069447553);

  // 'cut' is a 1350 Hz lowpass filter ellip(6,1,60,1350/4000)
  // 'cut2' is a 650 Hz hipass filter ellip(6,1,60,650/4000,'high')
table cuta(data,-1,1.000000000000000000, -3.403652194615361700, 6.040583475484091900,
   -6.536759755492672900, 4.525210297126158700, -1.886690344384180000,
   0.377554980242819030 );
table cutb(data,-1,0.009334414205569915, 0.010027794571888913, 0.022510171651681572,
   0.019860004210555920, 0.022510171651681597, 0.010027794571888893,
   0.009334414205569937 );
table cut2a(data,-1,1.000000000000000000, -3.764010552695741500, 6.413383082502888600,
   -6.167356008575000100, 3.585605607894495800, -1.223663081518635300,
   0.214584714639968120 );
table cut2b(data,-1,0.331211869820866030, -1.907140130890776900, 4.652812380824443900,
   -6.153709688044611600, 4.652812380824414600, -1.907140130890753000,
   0.331211869820859710 );

}

instr trans_coh_noise(first,step,n,amp) {
  asig out, ln, i, sum, q;
  oparray oscil_randph[10];
  imports table sint, lowa, lowb, cuta, cutb, cut2a, cut2b;

  ln = iirt(iirt(arand(1)*amp,higha,highb,9),lowa,lowb,9);

  sum = 0; i = 0; while (i < n) {
    sum = sum + oscil_randph[i](sint,first+step*i);
    i = i + 1;
  }

  out = iirt(iirt(sum * ln,cuta,cutb),cut2a,cut2b);
  output(out);
}

instr trans_rand_noise(first,step,n,amp) {
  asig out, ln, i, sum, q;
  oparray oscil_randph[10],iirt[20];
  imports table sint, lowa, lowb, cuta, cutb, cut2a, cut2b;

  sum = 0; i = 0; while (i < n) {
    ln = iirt[i](iirt[n+i](arand(1)*amp,higha,highb,9),lowa,lowb,9);
    sum = sum + ln * oscil_randph[i](sint,first+step*i);
    i = i + 1;
  }

  out = iirt(sum,cuta,cutb);
  output(out);
}

aopcode oscil_randph(table t, asig r) {
  asig init, ph, out;

  if (!init) {
    init = 1;
    ph = arand(0.5)+0.5;
  }

  out = tableread(t,ph*ftlen(t));
  ph = ph + r / s_rate;
  if (ph > 1) {
    ph = ph - 1;
  }
```

```
  return(out);
}
```

Score code (Stimulus S1)

As with the alternating-tone-sequence examples above, the same orchestra is used with six different scores to produce the six stimuli. The only difference between the scores here is the amplitude of the tone (the third parameter to tone()) and the type of noise used.

```
0.0 trans_rand_noise 2.5 1050 100 7 0.5
1.0 tone 0.3 1000 0.1
```

Score code (Stimulus S2)

```
0.0 trans_coh_noise 2.5 1050 100 7 0.5
1.0 tone 0.3 1000 0.1
```

Score code (Stimulus S3)

```
0.0 trans_rand_noise 2.5 1050 100 7 0.5
1.0 tone 0.3 1000 0.2
```

Score code (Stimulus S4)

```
0.0 trans_coh_noise 2.5 1050 100 7 0.5
1.0 tone 0.3 1000 0.2
```

Score code (Stimulus S5)

```
0.0 trans_rand_noise 2.5 1050 100 7 0.5
1.0 tone 0.3 1000 0.35
```

Score code (Stimulus S6)

```
0.0 trans_coh_noise 2.5 1050 100 7 0.5
1.0 tone 0.3 1000 0.35
```

REFERENCES

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.

Allen, J. B. (1999). Psychoacoustics. In J. G. Webster (ed.), *Wiley Encyclopedia of Electrical and Electronics Engineering* (pp. 422-437). New York: John Wiley & Sons.

Amick, D. J. & Walberg, H. J. (1975). *Introductory Multivariate Analysis*. Berkeley, CA: McCutchan Publishing

Balkwill, L.-L. & Thompson, W. F. (in press). A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. *Music Perception*.

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

Beauvois, M. W. & Meddis, R. (1996). Computer simulation of auditory stream segregation in alternating-tone sequences. *Journal of the Acoustical Society of America* **99**(4), 2270-2280.

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

Berg, B. G. (1996). On the relation between comodulation masking release and temporal modulation transfer functions. *Journal of the Acoustical Society of America* **100**(2), 1013-1023.

Bigand, E., Parncutt, R. & Lerdahl, F. (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.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. New York: Oxford University Press

Brandenberg, K. (1998). Perceptual coding of high quality digital audio. In M. Kahrs & K. Brandenburg (eds.), *Applications of Digital Signal Processing to Audio and Acoustics* (pp. 39-83). New York: Kluwer Academic.

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

Brooks, R. A. (1999). *Cambrian Intelligence*. Cambridge MA: MIT Press

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

Brown, G. J. & Cooke, M. (1994b). Perceptual grouping of musical sounds: A computational model. *Journal of New Music Research* **23**, 107-132.

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

Brown, J. & Puckette, M. S. (1989). Calculation of a 'narrowed' autocorrelation function. *Journal of the Acoustical Society of America* **85**(5), 1595-1601.

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. (1993). Determination of the meter of musical scores by autocorrelation. *Journal of the Acoustical Society of America* **94**(4), 1953-1957.

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

Buus, S. (1985). Release from masking caused by envelope fluctuations. *Journal of the Acoustical Society of America* **78**(6), 1958-1965.

Cabe, P. A. & Pittenger, J. B. (2000). Human sensitivity to acoustic information from vessel filling. *Journal of Experimental Psychology: Human Perception and Performance* **26**(1), 313-324.

Cariani, P. A. (1996). Temporal coding of musical form. In *Proceedings of the 1996 International Conference on Music Perception and Cognition* (pp. 425-430). Montreal.

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

Carlyon, R. P. (1991). Discriminating between coherent and incoherent frequency modulation of complex tones. *Journal of the Acoustical Society of America* **89**(1), 329-340.

Carlyon, R. P. (1994). Further evidence against an across-frequency mechanism specific to the detection of frequency modulation (FM) incoherence between resolved frequency components. *Journal of the Acoustical Society of America* **95**(2), 949-952.

Carter, N. P., Bacon, R. A. & Messenger, T. (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. & Jaffe, D. (1986). Source separation and note identification in polyphonic music. Stanford University CCRMA Technical Report #STAN-M-34, Palo Alto, CA.

Chafe, C., Jaffe, D., Kashima, K., Mont-Reynaud, B. & Smith, J. (1985). Techniques for note identification in polyphonic music. In *Proceedings of the 1985 ICMC* (pp. 399-405). Tokyo.

Chafe, C., Mont-Reynaud, B. & Rush, L. (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 CCRMA Technical Report #STAN-M-64, Palo Alto, CA.

Churchland, P. S., Ramachandran, V. S. & Sejnowski, T. J. (1994). A critique of pure vision. In C. Koch & J. L. Davis (eds.), *Large-Scale Neuronal Theories of the Brain* (pp. 23-60). Cambridge, MA: MIT Press.

Clarke, E. F. (1986). Theory, analysis and the psychology of Music: A critical evaluation of Lerdahl, F. and Jackendoff, R., *A Generative Theory of Tonal Music*. *Psychology of Music and Music Education* **14**, 3-16.

Clarke, E. F. & Krumhansl, C. L. (1990b). Perceiving musical time. *Music Perception* **7**(3), 213-252.

Clarke, E. F. & Krumhansl, C. L. (1990a). 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

Cook, N. (1994). Perception: A perspective from music theory. In R. Aiello & J. A. Sloboda (eds.), *Musical Perceptions* (pp. 64-95). New York: Oxford University Press.

Cook, N. (1998). *Music: A Very Short Introduction*

Cooke, M. (1993). *Modelling Auditory Processing and Organisation*. Cambridge, UK: University of Cambridge 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., Thom, B. & Watson, D. (1997). A machine learning approach to musical style recognition. In *Proceedings of the 1997 ICMC* (pp. 344-347). 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. (1997). Concurrent vowel identification. III. A neural model of harmonic interference cancellation. *Journal of the Acoustical Society of America* **101**(5), 2857-2865.

de Cheveigné, A. (1998a). The auditory system as a separation machine. In *Proceedings of the 1998a ??? .*

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

de Cheveigné, A. & Kawahara, H. (1999). Multiple period estimation and pitch perception model. *Speech Communication* **27**(3-4), 175-185.

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

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

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

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society (series B)* **39**(1), 1-38.

Dennett, D. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press

Desain, P. (1995). A (de)composable theory of rhythm perception. *Music Perception* **9**(3), 439-454.

Desain, P. & Honing, H. (1992a). *Music, Mind, and Machine: Studies in Computer Music, Music Cognition, and Artificial Intelligence*. Amsterdam: Thesis Publishers

Desain, P. & Honing, H. (1992b). Tempo curves considered harmful. In P. D. a. H. Honing (ed.), *Music, Mind and Machine: Studies in Computer Music, Music Cognition, and Artificial Intelligence* (pp. 25-44). Amsterdam: Thesis Publishers.

Desain, P. & Honing, H. (1994). Can music cognition benefit from computer music research? From foot-tapper systems to beat induction models. In *Proceedings of the 1994 ICMPC* (pp. 397-398). Liege BE.

Desain, P. & Honing, H. (1999). Computational models of beat induction: The rule-based approach. *Journal of New Music Research* **28**(1), 29-42.

Desain, P., Honing, H., van Thienen, H. & Windsor, L. (1998). Computational modeling of music cognition: Problem or solution? *Music Perception* **16**(1), 151-166.

Divenyi, P. L., Carre, R. & Algazi, A. P. (1997). Auditory segregation of vowel-like sounds with static and dynamic spectral properties. In *Proceedings of the 1997 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Mohonk, NY.

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

Drake, C. (1998). Psychological processes involved in the temporal organization of complex auditory sequences: Universal and acquired processes. *Music Perception* **16**(1), 11-26.

Duda, R. O. & Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. New York: John Wiley and Sons

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

Ellis, D. P. W. (1994). A computer implementation of psychoacoustic grouping rules. MIT Media Laboratory Perceptual Computing Technical Report #224, Cambridge, MA. Available from http://vismod.www.media.mit.edu/vismod/publications.

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 *Proceedings of the 1996b ESCA workshop on the Auditory Basis of Speech Perception.* Keele UK.

Ellis, D. P. W. (1997). The weft: A representation for periodic sounds. In *Proceedings of the 1997 Int. Conf. on Acoust. Speech and Sig. Proc.* (pp. 1307-1310). Munich.

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

Engelmore, R. & Morgan, T. (eds.) (1988). *Blackboard Systems*. Wokingham, UK: Addison-Wesley.

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

Essens, D.-J. P. a. P. (1985). Perception of temporal patterns. *Music Perception* **2**(3), 411-440.

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

Foote, J. (1999). An overview of audio information retrieval. *Multimedia Systems* **7**(1), 2-10.

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

Freed, D. J. (1990). Auditory correlates of perceived mallet hardness for a set of recorded percussive sound events. *Journal of the Acoustical Society of America* **87**(1), 311-322.

Fucci, D., Harris, D., Petrosino, L. & Banks, M. (1993). The effect of preference for rock music on magnitude-estimation scaling behavior in young adults. *Perceptual and Motor Skills* **76**(3), 1171-1176.

Fucci, D., Petrosino, L., Banks, M., Zaums, K. & Wilcox, C. (1996). The effect of preference for three different types of music on magnitude estimation-scaling behavior in young adults. *Perceptual and Motor Skills* **83**(1), 339-347.

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

Gardner, H. (1985). *The Mind's New Science*. New York: Basic Books

Garner, W. R. (1978). Aspects of a stimulus: Features, dimensions, and configurations. In E. Rosch & B. B. Lloyd (eds.), *Cognition and Categorization* (pp. 99-133). Hillsdale NJ: Lawrence Erlbaum.

Gibson, J. J. (1986). *The Ecological Approach to Visual Perception*. Hillsdale, NJ: Lawrence Erlbaum Publishers

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 CCRMA Technical Report #STAN-M-38, Palo Alto CA.

Goto, M. (1999). Real-time beat tracking for drumless audio signals: Chord change detection for musical decisions. *Speech Communication* **27**(3-4), 311-335.

Goto, M. & Hayamizu, S. (1999). A real-time music scene description system: Detecting melody and bass lines in audio signals. In *Proceedings of the 1999 International Joint Conference on Articifial Intelligence Workshop on Computational Auditory Scene Analysis* (pp. 31-40). Stockholm.

Goto, M. & Muraoka, Y. (1995). A real-time beat tracking system for audio signals. In *Proceedings of the 1995 ICMC* (pp. 171-174). Banff.

Goto, M. & Muraoka, Y. (1998). Music understanding at the beat level: Real-time beat tracking for audio signals. In D. F. Rosenthal & H. Okuno (eds.), *Readings in Computational Auditory Scene Analysis* (pp. 157-176). Mahweh, NJ: Lawrence Erlbaum.

Green, D. M. (1996). Discrimination changes in spectral shape: Profile analysis. *Acustica* **82**, S31-S36.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America* **61**(5), 1270-1277.

Gries, D. (1981). *The Science of Programming*. New York: Springer-Verlag

Hajda, J. M., Kendall, R. A., Carterette, E. C. & Harshberger, M. L. (1997). Methodological issues in timbre research. In I. Deliège (ed.), *Perception and Cognition of Music* (pp. 253-306). Hove, UK: Psychology Press.

Hall, J. W., Haggard, M. P. & Fernandes, M. A. (1984). Detection in noise by spectro-temporal pattern analysis. *Journal of the Acoustical Society of America* **76**(1), 50-56.

Handel, S. (1989). *Listening: An Introduction to the Perception of Auditory Events*. Cambridge, MA: MIT Press

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

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.

Helmholtz, H. L. F. (1885 / 1954). *On the Sensations of Tone*. New York: Dover Publications (Trans. A. J. Ellis).

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

Hermes, D. J. (1993). Pitch analysis. In M. Cooke, S. Beet & M. Crawford (eds.), *Visual representations of speech signals* (pp. 3-25). London: John Wiley & Sons Ltd.

Hewlett, W. B. (ed.) (1998). *Melodic Similarity: Concepts, Procedures, and Applications*. Computing in Musicology. Cambridge, MA: MIT Press.

Hirsch, I. J. & Watson, C. S. (1996). Auditory psychophysics and perception. *Annual Review of Psychology* **47**, 461-484.

Hofstadter, D. R. (1979). *Godel, Escher, Bach: An Eternal Golden Braid*. New YOrk: Basic Books

Holleran, S., Jones, M. R. & Butler, D. (1995). Perceiving musical harmony: The influence of melodic and harmonic context. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **21**(3), 737-753.

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

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

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

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

Johnson-Laird, P. N. (1991a). Jazz improvisation: A theory at the computational level. In P. Howell, R. West & I. Cross (eds.), *Representing musical structure* (pp. 291-326). London: Academic Press.

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

Jones, M. R. & Boltz, M. (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. & Murase, H. (1997). Sound source identification for ensemble music based on the music stream extraction. In *Proceedings of the 1997 Int. Joint Conf. on AI Workshop on Computational Auditory Scene Analysis* (pp. 127-134). Tokyo.

Kashino, K., Nakadai, K., Kinoshita, T. & Tanaka, H. (1995). Application of Bayesian probability network to music scene analysis. In *Proceedings of the 1995 Int. Joint Conf. on AI Workshop on Computational Auditory Scene Analysis* (pp. 52-59). Montreal.

Kashino, K. & Tanaka, H. (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. & Tanaka, H. (1993). A source source separation system with the ability of automatic tone modeling. In *Proceedings of the 1993 ICMC* (pp. 248-255). Tokyo.

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

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

Kohenen, T. (1995). *Self-Organizing Maps*. Berlin: Springer-Verlag

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

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 M. Leman (ed.), *Music, Gestalt, and Computing: Studies in Systematic and Cognitive Musicology* (pp. 294-320). Berlin: Springer.

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

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

Ladefoged, P. (1982). *A Course in Phonetics*. San Diego: Harcourt Brace Jovanovich

Langner, G. (1992). Periodicity coding in the auditory system. *Hearing Research* **60**(1), 115-142.

Large, E. W. & Kolen, J. F. (1994). Resonance and the perception of musical meter. *Connection Science* **6**(2), 177-208.

Lee, X. F., Logan, R. J. & Pastore, R. E. (1991). Perception of acoustic source characteristics: Walking sounds. *Journal of the Acoustical Society of America* **90**(6), 3036-3049.

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: Cognitive Foundations of Systematic Musicology*. Berlin: Springer-Verlag

Lerdahl, F. & Jackendoff, R. (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. & Cook, P. R. (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 S. S. Stevens (ed.), *Handbook of Experimental Psychology* (pp. 985-1035). New York: Wiley.

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

Liu, Z., Wang, Y. & Chen, T. (1998). Audio feature extraction and analysis for scene segmentation and classification. *Journal of VLSI Signal Processing* **20**(1-2), 61-79.

Longuet-Higgens, H. C. & Lee, C. S. (1984). The rhythmic interpretation of monophonic music. *Music Perception* **1**(2), 424-441.

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

Madsen, C. K. (1996). Empirical investigation of the "aesthetic response" to music: Musicians and nonmusicians. In *Proceedings of the 1996 Int. Conf. Music Perception and Cognition* (pp. 103-108). Montreal.

Maes, P., Lashkari, Y. & Metral, M. (1997). Collaborative interface agents. In M. H. Huhns & M. P. Singh (eds.), *Readings in Agents* . New York: Morgan Kaufmann Publishers.

Maher, R. C. (1990). Evaluation of a method for separating digitized duet signals. *Journal of the Audio Engineering Society* **38**(12), 956-979.

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

Mani, R. (1999). Knowledge-based processing of multicomponent signals in a musical application. *Signal Processing* **74**(1), 47-69.

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. Available from http://vismod.www.media.mit.edu/vismod/publications.

Martin, K. D. (1996b). A blackboard system for automatic transcription of simple polyphonic music. MIT Media Laboratory Perceptual Computing Technical Report #385, Cambridge MA. Available from http://vismod.www.media.mit.edu/vismod/publications.

Martin, K. D. (1999). *Sound-Source Recognition: A Theory and Computational Model.* Ph.D. thesis, MIT Department of Electrical Engineering and Computer Science, Cambridge, MA.

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

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

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

McAdams, S. & Saariaho, K. (1985). Qualities and functions of musical timbre. In *Proceedings of the 1985 ICMC* (pp. 367-374). Burnaby BC, CA.

McAulay, R. J. & Quatieri, T. F. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **34**(4), 744-754.

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

Meddis, R. & Hewitt, M. J. (1991). 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. & Hewitt, M. J. (1992). Modeling the identification of concurrent 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.

Meyer, L. B. (1956). *Emotion and Meaning in Music*. Chicago: University of Chicago Press

Minami, K., Akutsu, A., Hamada, H. & Tonomura, Y. (1998). Video handling with music and speech detection. *IEEE Multimedia* **5**(3), 17-25.

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

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

Minsky, M. & Laske, O. (1992). Foreword: A conversation with Marvin Minsky. In M. Balaban, K. Ebcioglu & O. Laske (eds.), *Understanding Music with AI: Perspectives on Music Cognition* (pp. ix-xxxviii). Cambridge MA: MIT Press.

Moore, B. C. J. (1997). *An Introduction to the Psychology of Hearing*. San Diego: Academic 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., Espy-Wilson, C. Y., Mani, R. & Bitar, N. N. (1998). Knowledge-based analysis of speech mixed with sporadic environmental sounds. In D. F. Rosenthal & H. Okuno (eds.), *Readings in Computational Auditory Scene Analysis* (pp. 177-193). Mahweh, NJ: Erlbaum.

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

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

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

Parncutt, R. (1994a). A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception* **11**(2), 409-464.

Parncutt, R. (1994b). Template-matching models of musical pitch and rhythm perception. *Journal of New Music Research* **23**, 145-167.

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

Patterson, R. D., Allerhand, M. H. & Giguere, C. (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.

Patterson, R. D., Robinson, K., Holdsworth, J. *et al* (1992). Complex sounds and auditory images. In Y. Cazals, K. Horner & L. Demany (eds.), *Auditory Physiology and Perception* (pp. 429-446). Oxford: Pergamon Press.

Pereira, F. & Koenen, R. H. (2000). MPEG-7: Status and directions. In A. Puri & T. Chen (eds.), *Advances in Multimedia: Signals, Standards, and Networks* (pp. 611-630). New York: Marcel Dekker.

Perrott, D. & Gjerdigen, R. O. (1999). Scanning the dial: An exploration of factors in the identification of musical style. In *Proceedings of the 1999 Society for Music Perception & Cognition* (pp. 88 (abstract)). Evanston, IL.

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

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

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

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

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

Povel, D. & Okkerman, H. (1981). Accents in equitone sequences. *Perception & Psychophysics* **30**(3), 565-572.

Povel, D.-J. & Essens, P. (1985). Perception of temporal patterns. *Music Perception* **2**(2), 411-480.

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

Quatieri, T. F. & McAulay, R. J. (1998). Audio signal processing based on sinusoidal analysis/synthesis. In M. Kahrs & K. Brandenburg (eds.), *Applications of Digital Signal Processing to Audio and Acoustics* (pp. 343-411). New York: Kluwer Academic.

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

Reed, E. S. (1996). *Encountering the World*. New York: Oxford University Press

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

Roads, C., Pope, S. T., Piccialli, A. & de Poli, G. (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.

Rose, M. M. & Moore, B. C. J. (1997). Perceptual grouping of tone sequences by normally hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America* **102**(3), 1768=1778.

Rosenthal, D. F. (1992). *Machine Rhythm: Computer Emulation of Human Rhythm Perception.* Ph.D. thesis, MIT Media Laboratory, Cambridge, MA.

Rosner, B. S. (1988). Music perception, music theory, and psychology. In E. Narmour & R. A. Sobie (eds.), *Explorations in Music, the Arts, and Ideas: Essays in Honor of Leonard B. Meyer* (pp. 141-175). Stuyvesant: Pendragon Press.

Rossi, L., Girolami, G. & Leca, M. (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 *Proceedings of the 1996 International Conference on Music Perception and Cognition* (pp. 317-322). Montreal: Society for Music Perception and Cognition.

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

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

Scheirer, E. D. & Slaney, M. (1997). Construction and evaluation of a robust multifeature speech/music discriminator. In *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 1331-1334). Munich: IEEE.

Scheirer, E. D. & Vercoe, B. L. (1999). SAOL: The MPEG-4 Structured Audio Orchestra Language. *Computer Music Journal* **23**(2), 31-51.

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

Schmuckler, M. A. & Boltz, M. G. (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 M. Leman (ed.), *Music, Gestalt, and Computing* (pp. 117-143). Berlin: Springer-Verlag.

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

Shamma, S. A. (1996). Auditory cortical representation of complex acoustic spectra as inferred from the ripple analysis method. *Network - Computation in Neural Systems* **7**(3), 439-476.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J. & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science* **270**, 303-304.

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. Available from http://www.interval.com/~malcolm.

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

Slaney, M. (1998). A critique of pure audition. In D. F. Rosenthal & H. G. Okuno (eds.), *Computational Auditory Scene Analysis* (pp. 27-42). Mahweh, NJ: Lawrence Erlbaum.

Slaney, M. & Lyon, R. F. (1990). A perceptual pitch detector. In *Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing* (pp. 357-360). Albuquerque: IEEE.

Slaney, M. & Lyon, R. F. (1991). Apple Hearing Demo Reel. Apple Computer, Inc. Technical Report #25, Cupertino CA. Available from malcolm@interval.com.

Slaney, M., Naar, D. & Lyon, R. F. (1994). Auditory model inversion for sound separation. In *Proceedings of the 1994 ICASSP*. Adelaide AU.

Sloboda, J. (1985). *The Musical Mind: The Cognitive Psychology of Music*. Oxford, UK: Clarendon Press

Smith, G., Murase, H. & Kashino, K. (1998). Quick audio retrieval using active search. In *Proceedings of the 1998 ICASSP*. Seattle.

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

Smoliar, S. W. (1991). The Analysis and Cognition of Basic Melodic Structures: The Implication-Realization Model. E. NarmourIn Theory Only. **12:** 43-56.

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

Snyder, J. S. & Krumhansl, C. L. (1999). Cues to pulse-finding in piano ragtime music. In *Proceedings of the 1999 Society for Music Perception and Cognition* (pp. 1). Evanston, IL.

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

Stevens, S. S. (1975). *Psychophysics*. New York: John Wiley & Sons

Summerfield, Q., Lea, A. & Marshall, D. (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 M. Clynes (ed.), *Music, Mind, and Brain: The Neuropsychology of Music* (pp. 353-369). New York: Plenum Press.

Terhardt, E. (1991). Music perception and sensory information acquisition: Relationships and low-level analogies. *Music Perception* **8**(3), 217-240.

Terhardt, E., Stoll, G. & Seewann, M. (1982). Algorithm for extraction of pitch and pitch salience from complex tonal signals. *Journal of the Acoustical Society of America* **71**(3), 679-688.

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. & Parncutt, R. (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.

Todd, N. P. M. (1994). The auditory "primal sketch": A multiscale model of rhythmic grouping. *Journal of New Music Research* **23**, 25-70.

Van Immerseel, L. M. & Martens, J.-P. (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 M. Clynes (ed.), *Music, Mind, and Brain: The Neuropsychology of Music* (pp. 251-269). New York: Plenum Press.

van Noorden, L. P. A. S. (1977). Minimum differences of level and frequency for pereeptual fission of tone sequences ABAB. *Journal of the Acoustical Society of America* **61**(4), 1041-1045.

Vercoe, B. L. (1984). The synthetic performer in the context of live performance. In *Proceedings of the 1984 International Computer Music Conference* (pp. 199-200). Paris: International Computer Music Association.

Vercoe, B. L. (1985). Csound: A Manual for the Audio-Processing System (rev. 1996). Program reference manual, Cambridge MA, MIT Media Lab.

Vercoe, B. L. (1988). Hearing polyphonic music on the connection machine. In *Proceedings of the 1988 First AAAI Workshop on Artificial Intelligence and Music* (pp. 183-194). Minneapolis.

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

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

Vercoe, B. L. & Puckette, M. S. (1985). Synthetic rehearsal: Training the synthetic performer. In *Proceedings of the 1985 ICMC* (pp. 275-278). Burnaby BC, Canada.

Verhey, J. L., Dau, T. & Kollmeier, B. (1999). Within-channel cues in comodulation masking release (CMR): Experiments and model predictions using a modulation-filterbank model. *Journal of the Acoustical Society of America* **106**(5), 2733-2745.

Versnel, H. & Shamma, S. A. (1998). Spectral-ripple representation of steady-state vowels in primary auditory cortex. *Journal of the Acoustical Society of America* **103**(5), 2502-2514.

Vliegen, J. & Moore, B. C. J. (1999). The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task. *Journal of the Acoustical Society of America* **106**(2), 938-945.

Vliegen, J. & Oxenham, A. J. (1999). Sequential stream segregegation in the absence of spectral cues. *Journal of the Acoustical Society of America* **105**(1), 339-346.

Vos, P. G. & Van Geenen, E. W. (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.

Wang, K. & Shamma, S. A. (1995). Spectral shape-analysis in the central auditory system. *IEEE TSAP* **3**(5), 382-395.

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

Warren, R. M. (1999). *Auditory Perception: A New Analysis and Synthesis*. Cambridge UK: Cambridge University Press

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

Warren, W. H. & Verbrugge, R. R. (1984). Auditory perception of breaking and bouncing events: A case study in ecological acoustics. *Journal of Experimental Psychology: Human Perception and Performance* **10**(4), 704-712.

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.

Windsor, W. L. (1995). *A Perceptual Approach to the Description and Analysis of Acousmatic Music*. Ph.D. thesis, City University Deparment of Music, London.

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

Yost, W. A. (1991). Auditory image perception and analysis: The basis for hearing. *Hearing Research* **56**, 8-18.

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