Sound-Source Recognition: A Theory and Computational Model

by

Keith Dana Martin

B.S. (with distinction) Electrical Engineering (1993) Cornell University
S.M. Electrical Engineering (1995) Massachusetts Institute of Technology

Submitted to the department of
Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June, 1999

Author ....
Department of Electrical Engineering and Computer Science
May 17, 1999

Certified by ....
Barry L. Vercoe
Professor of Media Arts and Sciences
Thesis Supervisor

Accepted by ....
Professor Arthur C. Smith
Chair, Department Committee on Graduate Students

Sound-source recognition: A theory and computational model

by Keith Dana Martin

Submitted to the Department of Electrical Engineering and Computer Science on May 17, 1999, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Electrical Engineering and Computer Science.

Abstract

The ability of a normal human listener to recognize objects in the environment from only the sounds they produce is extraordinarily robust with regard to characteristics of the acoustic environment and of other competing sound sources. In contrast, computer systems designed to recognize sound sources function precariously, breaking down whenever the target sound is degraded by reverberation, noise, or competing sounds. Robust listening requires extensive contextual knowledge, but the potential contribution of sound-source recognition to the process of auditory scene analysis has largely been neglected by researchers building computational models of the scene analysis process.

This thesis proposes a theory of sound-source recognition, casting recognition as a process of gathering information to enable the listener to make inferences about objects in the environment or to predict their behavior. In order to explore the process, attention is restricted to isolated sounds produced by a small class of sound sources, the non-percussive orchestral musical instruments. Previous research on the perception and production of orchestral instrument sounds is reviewed from a vantage point based on the excitation and resonance structure of the sound-production process, revealing a set of perceptually salient acoustic features.

A computer model of the recognition process is developed that is capable of “listening” to a recording of a musical instrument and classifying the instrument as one of 25 possibilities. The model is based on current models of signal processing in the human auditory system. It explicitly extracts salient acoustic features and uses a novel improvisational taxonomic architecture (based on simple statistical pattern-recognition techniques) to classify the sound source. The performance of the model is compared directly to that of skilled human listeners, using both isolated musical tones and excerpts from compact disc recordings as test stimuli. The computer model's performance is robust with regard to the variations of reverberation and ambient noise (although not with regard to competing sound sources) in commercial compact disc recordings, and the system performs better than three out of fourteen skilled human listeners on a forced-choice classification task.

This work has implications for research in musical timbre, automatic media annotation, human talker identification, and computational auditory scene analysis.

Thesis supervisor: Barry L. Vercoe
Title: Professor of Media Arts and Sciences

Acknowledgments

I am grateful for the fantastic level of support I have enjoyed in my time at MIT. First and foremost, I thank Barry Vercoe for bringing me into his rather unique research group (known by a wide range of names over the years, including The Machine Listening Group, Synthetic Listeners and Performers, Music and Cognition, and “The Bad Hair Group”). I could not have dreamed of a place with a broader sense of intellectual freedom, or of an environment with a more brilliant group of colleagues. Credit for both of these aspects of the Machine Listening Group is due entirely to Barry. I am thankful also for his patient encouragement (and indulgence) over the past six years.

I am grateful to Marvin Minsky and Eric Grimson for agreeing to serve as members of my doctoral committee. I have drawn much inspiration from reading their work, contemplating their ideas, and adapting their innovations for my own use.

Two members of the Machine Listening Group deserve special accolades. On a level of day-to-day interaction, I thank Eric Scheirer for serving as my intellectual touchstone. Our daily conversations over morning coffee have improved my clarity of thought immensely (not to mention expanded my taste in literature and film). Eric has been a faithful proof-reader of my work, and my writing has improved significantly as a result of his feedback.

Many of my ideas are variations of things I learned from Dan Ellis, and I count him among my most influential mentors. Although my dissertation could be viewed as a criticism of his work, it is only because of the strengths of his research that mine makes sense at all.

I would also like to thank the other members of the Machine Listening Group, past and present. To my officemates, Youngmoo Kim, Bill Gardner, Adam Lindsay, and Nicolas Saint-Arnaud, thanks for putting up with my music and making the daily grind less of one. Also, thanks to Paris Smaragdis, Jonathan Feldman, Nyssim Lefford, Joe Pompei, Mike Casey, Matt Krom, and Kathryn Vaughn.

For helping me assemble a database of recordings, I am grateful to Forrest Larson of the MIT Music Library, Juergen Herre, Janet Marques, Anand Sarwate, and to the student performers who consented to being recorded, including Petra Chong, Joe Davis, Jennifer Grucza, Emily Hui, Danny Jochelson, Joe Kanapka, Teresa Marrin, Ole Nielsen, Bernd Schoner, Stephanie Thomas, and Steve Tistaert.

I would like to thank the students in the Media Lab's pattern-recognition course who worked with some preliminary feature data and provided valuable feedback. Special thanks are due to Youngmoo Kim, who collaborated with me on my first musical instrument recognition system (Computer experiment #1 in Chapter 6).

Connie Van Rheenen, Betty Lou McClanahan, Judy Bornstein, Dennis Irving, Greg Tucker, and Linda Peterson provided essential support within the Media Lab, and I am thankful for their assistance. Other Media Labber's owed thanks include Bill Butera, Jocelyn Scheirer, and Tom Minka.

Over my many years of schooling, I have been strongly influenced or assisted by many teachers. At MIT, I have particularly benefited from interacting with Roz Picard, Whit Richards, and Bill Peake. I am grateful to two professors at Cornell University, Carol Krumhansl and Clif Pollock, both for being great teachers and for helping me get to MIT.

Of course, my most influential teachers are my parents, and I owe an immense debt of gratitude to them for their support (of all kinds) over the years. Thank you for helping me get through my many paroxysms of self doubt.[1]

Finally I wish to thank my wife and best friend, Lisa, whose emotional support during the ups and downs of graduate school made it possible for me to keep going. I promise not to drag you through this again!

This thesis is dedicated to the memory of my grandfather, Dana West Martin.

[1] A phrase gleefully stolen from David Foster Wallace, who along with Kurt Vonnegut, Andy Partridge, and the members of Two Ton Shoe, deserves a little bit of credit for helping me maintain some level of sanity during the last year or so.

Table of Contents

CHAPTER 1 Introduction

I am sitting in my office, and a Beatles compact disc is playing on my stereo. I hear many different sounds, yet I have little difficulty making sense of the mixture. I can understand the singer's words and can tell that it is Paul McCartney singing. I hear drums, electric guitars, organ, and bass guitar. In addition to the sounds reproduced by my stereo's speakers, I can hear cars driving by, the chatter of children walking home from the school bus stop, and the humidifier humming in the hallway. The telephone rings, and I answer it. I recognize my wife's voice from a single word ("Hi"), and realize that she is calling to tell me when she will be home from work. I turn down the stereo to hear her more clearly, and now I can hear that our cat is scratching the sofa in the next room.

These examples are mundane, but they illustrate how easily we gather information with our ears. The language we use to describe our perceptions is also revealing. We often describe what we hear in terms of the objects producing the sounds and the information that the sounds convey. We hear a *dog barking nervously* (or *viciously*), a *glass breaking*, an *airplane flying overhead*, a *bell ringing*, a *violinist playing a melody,* and so on. (Loudspeakers—as in the example above—are special-case sources that reproduce sounds originally produced by other sources.) We routinely understand mixtures of sounds, somehow segmenting, parsing, disentangling, or otherwise interpreting the complicated auditory scene that arrives at our ears.

Hearing is an important part of normal human interaction, yet we understand surprisingly little about how our brains make sense of sound. Our limited knowledge is partly a result of the inability to gain conscious access to our perceptual processes, but our language, far removed from sound waves, also limits us. We have difficulty explaining what something sounds like except by analogy to other sounds. Our descriptive words for sound—loud, bright, rough, cacophonous, sweet, open, dark, mellow, percussive, droning, scratchy, dull, smooth, screechy, pounding, noisy, clanging—are extremely context-dependent, and most of them have no clear relationship to properties that scientists know how to measure.

1.1 Motivation and approach

This dissertation is driven by the desire to understand how human auditory perception works. In it, I take the view that the human auditory system is a complex information-processing system. By considering the constraints under which the human system operates, the limitations of its "hardware," and the perceptual abilities and limitations of the listener, it is possible to form theories of the system's operation. The theories can subsequently be tested by constructing and evaluating computational models. In this dissertation, *theory* refers to an idea or algorithm, and *model* refers to its implementation, usually on a general-purpose computer.

Computational models are the best tools we have for understanding complex systems. By formulating a theory of a system's operation, constructing a model that embodies the theory, and then testing the performance of the model, it is possible to identify the strengths and weaknesses of the theory. Sometimes, the model will mimic some aspect of the system, and this correspondence can be taken as evidence in favor of the theory. More often, however, the model will fail to account for crucial aspects of the system's behavior. These shortcomings are valuable because they tell us about the weaknesses of the theory, often highlighting tacit assumptions made by the theorist. Models are also valuable because they can be extensively manipulated. By changing parameters or selectively enabling and disabling the model's components, it is possible to gain insight into the operation of the system as a whole.

In this dissertation, I describe a theory and computational model of auditory sound-source recognition. The theory is a response to ongoing work in the nascent field of computational auditory scene analysis (CASA), where systems are developed to model the process of understanding mixtures of sounds. By and large, current CASA models rely on handfuls of signal-processing techniques and sets of "grouping heuristics" to divide a sound signal into parts arising from independent sources. Although some authors have acknowledged the need for "top-down" processing in auditory scene analysis, current CASA models make little use of world knowledge or contextual information to aid the process of scene analysis. This contrasts starkly with the human perceptual system, for which context is indispensable. I view hearing as a complex task similar to assembling a jigsaw puzzle, where world knowledge ("ah, that's a bit of tree branch") can be used to get closer to a solution ("it must go with the other branch pieces in the corner here")[1]. In this view, recognition is intimately tied to the pro-

[1] Of course, with hearing, the puzzle is always changing, making it important to assemble the pieces quickly!

cess of understanding. Complex mixtures would be impenetrable without extensive knowledge-based inference. It remains to be seen if models without extensive world knowledge can solve any interesting—that is, *realistic*—perceptual problems. So far, there has been no existence proof.

The model I describe operates on recordings of isolated sound sources and recognizes a limited range of sound-source classes—the non-percussive orchestral instruments. It cannot be viewed as a complete model of human sound-source recognition. However, I have endeavored to construct a model that could be integrated into a suitable CASA framework. The representations described here may be extended easily to include other kinds of sound sources. Although the model is not complete without both pieces, recognizing isolated sound sources is a sufficiently complex problem to merit attention on its own. In the following discussion, I point out the additional complexities due to mixtures of sounds when it is relevant.

The ideas described in this dissertation are not the result of my efforts alone. Several of the techniques I employ are inspired by (or derived from) research in visual perception. In particular, my views on representation and modeling are strongly influenced by the work of David Marr, and several ideas have been adapted from work by Shimon Ullman and Eric Grimson. My views on auditory scene analysis are particularly influenced by the modeling work of Dan Ellis, and by the writing of Stephen Handel and Stephen McAdams. I have also drawn from the theories of mind described by Marvin Minsky and Daniel Dennett. At times, I employ language reminiscent of the writing of J. J. Gibson and the ecological psychologists; their influence is visible most clearly in my experiments, which employ real-world stimuli rather than laboratory confections.

1.2 A theory of sound-source recognition

In this section, I outline a general theory of sound-source recognition. In the rest of the dissertation, I will provide evidence that supports some of its elements, demonstrate a computational model based on its principles, and show how the model recognizes sound sources in a manner similar to humans. The general theory of sound-source recognition that I propose can be stated simply. Recognition is a *process*—not an achievement or goal. It is the process of gathering information about objects in the environment so as to more accurately predict or infer their behavior. I will use the language of classification to describe this process, but it is important to note that the theory makes no claims about the immanent status of categories. Categories, or classes, are merely groups of objects that have similar characteristics in some frame of reference. A particular *categorization*, or division of objects into classes, is useful only insofar as knowledge of an object's category label enables the perceiver to make accurate predictions about some unobserved aspect of the object.

I adopt the viewpoint that a sound-producing object belongs to various categories at different levels of abstraction. An illustration of this idea, synthesized from drawings and discussion by Bobick (1987) and Minsky (1986), is shown in Figure 1. Some general properties of this organization are worth observing. The particular categories shown are not the only possible choices—others might include "things Bill likes to listen to," "brown wooden things," or "things with ornate shapes," but these latter sorts are not as useful to the recognition process because they do not permit as many inferences. At the top of the figure is a single category, labeled "Sound Source," that contains all sound-producing objects. It does not, however, allow the perceiver to make inferences much stronger than "vibrates somewhere in the frequency range that can cause a human eardrum to move." At the very bottom are categories containing only a single object making

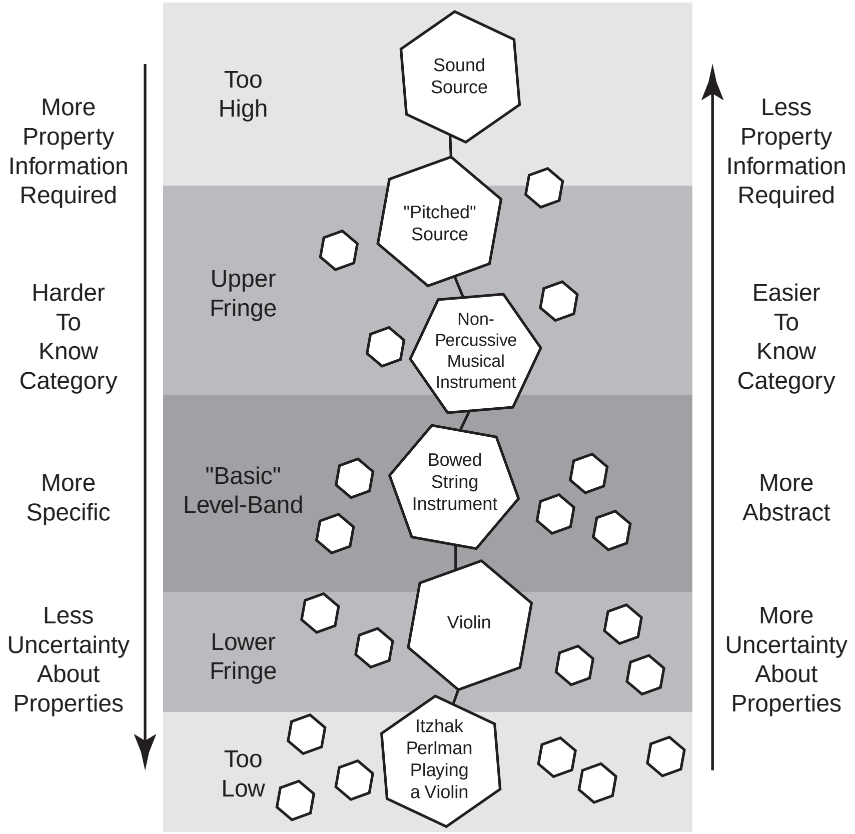
More Property Information Required
Harder To Know Category
More Specific
Less Uncertainty About Properties
Too High
Upper Fringe
"Basic" Level-Band
Lower Fringe
Too Low
Sound Source
"Pitched" Source
Non-Percussive Musical Instrument
Bowed String Instrument
Violin
Itzhak Perlman Playing a Violin
Less Property Information Required
Easier To Know Category
More Abstract
More Uncertainty About Properties

FIGURE 1. Recognition as classification in a category-abstraction space. The illustration is a synthesis of drawings from Bobick (1987) and Minsky (1986). A particular sound source—Itzhak Perlman playing a violin—is a member of different categories at different levels of abstraction. The arrows indicate that a change in the level of abstraction affects both the difficulty of determining the category of an object and the amount of information represented by knowledge of an object's category. The shaded regions and their labels correspond to Minsky's "level-bands." Minsky (1986) argues that there is a privileged level for reasoning and recognition that occurs at an intermediate level of abstraction.

a specific sound. In between, as the level of specificity increases (and the level of abstraction correspondingly decreases), more specific details are known but more information is required to choose among the categories. As the level of abstraction increases, less information is required to classify an object, but the classification does not provide the same predictive strength.

The process of recognition begins at an intermediate level of abstraction, where classification is relatively easy but still yields useful information about unobserved properties. It then proceeds to more specific categories as warranted by the needs of the listener. Sensory data accumulate, and increasingly specific classifications are made when they are useful. This approach has the benefits that it requires less effort when less-specific information is needed, and that the perceiver need never examine every possible categorization directly.

In this outline, I have purposely provided little detail about the various parts of the process. In the rest of the dissertation, I fill in the details by proposing a computational model of the recognition process. A small set of sound sources, the non-percussive orchestral instruments, are considered in depth, and the model is tested with natural recordings of sounds produced by these instruments. Its performance on a battery of classification tasks is compared to the performance of human listeners on similar tasks, highlighting the strengths and weaknesses of the model.

This dissertation will not address the acquisition of the category-abstraction structure or the development of new feature detectors. These difficult problems are left for future research.

1.3 Applications

The primary goal of this research is scientific: to present a theory of sound-source recognition and test it with a computational model. There are also several practical areas in which such a model might be applied, including:

- **Media annotation:** Over the last two decades, digital media have proliferated. For example, my personal digital-audio library includes well over 500 compact discs, and my laptop computer stores a wide variety of digital image, video, and audio files. To the computer or compact-disc player, however, these are merely streams of bits in some coding scheme. They are converted into images or sounds when I decide to play them. Today, we have internet search engines that can identify text documents matching a user's query, but multimedia documents are opaque to search engines. Today's systems have no way of discovering if a spoken phrase in a recording or an object in an image matches a query and retrieving the relevant document.

  Recently, efforts have begun that will result in standardized "descriptors," or *meta-data* formats, for multimedia data (MPEG Requirements Group, 1999). However, for most of the descriptors we would like to use—in queries such as "find the cadenzas of all the Mozart concertos in the database, and sort them by instrument" or "find all the photographs of Abraham Lincoln"—we have no tools that can extract the relevant information automatically. The producer of the data must add the meta-data by hand. Sound-source recognition—at the level achieved by the model described in Chapters 4 and 5—could be used at the point of production, where sounds are often isolated on separate channels of a multi-track recording system. Meta-data could be added before the sounds are mixed together and preserved throughout the production process. Better yet, recordings could be distributed in *structured* formats (Vercoe et al., 1998) that preserve the isolation of individual sounds until the time of playback, and then techniques like those described here could be applied by the end-user.

- **Talker identification:** Identifying a particular human voice is the one example of sound-source recognition that has received considerable attention in the scientific literature (e.g., Reynolds, 1995). The theory of sound-source recognition described in this dissertation is a general one, and as such can be viewed as a generalization of theories of talker identification. However, the techniques used here are very different from those typically used to build talker recognition systems. Some of the acoustic properties determined to be important for recognizing musical instruments may also be important for recognizing human talkers, and the hierarchical classification framework described here might be put to good use in speech systems as well.
- **Music transcription:** The process of listening to a piece of music and reconstructing the notated score is known as *transcription*. More generally, transcription is the process of determining *which* musical notes were played *when* (and by *what* instrument) in a musical recording or performance. In the general case of music played by multiple instruments (or a single polyphonic instrument such as a guitar or piano), the task is one of polyphonic pitch tracking. This is extraordinarily difficult—humans require extensive training in order to transcribe music reliably. However, because transcription is an important tool for music theorists, music psychologists, and musicologists—not to mention music lovers who want to figure out what their favorite artists are playing in rapid passages—it would be wonderful to have tools that could aid the transcription process, or automate it entirely. State-of-the-art polyphonic pitch tracking research demonstrates that the task is made simpler if good—and explicit—models of the sound sources (the musical instruments) are available (Kashino & Murase, 1998). By integrating sound-source recognition with a transcription engine, the end result can be improved dramatically.
- **Structured-audio encoding:** As noted above, structured-media formats make automatic multimedia annotation easier. In addition, they give the end user more control over the media playback. For example, an audio enthusiast could take better advantage of a seven-speaker playback setup if the audio material was not pre-mixed for stereo playback. Movie soundtracks could include speech tracks in multiple languages, enabling distributors to provide only one version of a movie for international presentation. Amateur musicians could “mute” a particular part of a recording and play along.

  Although structured formats provide immense advantages over their non-structured counterparts (such as the current generation of compact discs and videotapes), we currently have no way of automatically adding structure to an unstructured recording. In the future, by combining robust tools from sound-source recognition, CASA, music transcription, and speech recognition, it may be possible to build fully or partly automated tools for unstructured-to-structured encoding.

- **A composer's workbench:** The research described in this dissertation embodies a viewpoint on musical-instrument sound that is informed by knowledge of human perception. The techniques used to recognize sounds could be inverted and used to create new sounds based on natural, verbal descriptions. With a single workstation including analysis and synthesis tools, a composer could more easily create a wide variety of new sounds. Virtual instruments could be created—for example, "like a very large brass instrument, but with a percussive attack and pronounced vibrato"—without extensive physical modeling. Automatic indexing would be a valuable tool, enabling automatic responses to naturally posed requests such as "play back the part where the clarinet comes in."
- **Environment monitoring:** One of the most obvious applications of sound-source recognition is environmental monitoring. A home-monitoring system could alert the homeowner if there is someone knocking at the door, if the baby is crying, or if water is boiling over on the stove. Such systems could be used as the basis of prostheses for listeners with severe hearing loss, converting auditory information into another medium, such as a visual display.
- **Synthetic listeners and performers:** Endowing computer systems with the ability to recognize sounds and understand the information they convey would enable a host of exciting applications. We could build virtual music instructors (with unending patience!), virtual orchestras to conduct, and virtual performers to jam with. Although these applications may sound far-fetched, each has already been demonstrated in some form (Vercoe, 1984; Vercoe & Puckette, 1985).

1.4 Overview and scope

This dissertation makes contributions to modern hearing science at several levels, ranging from practical signal-processing techniques to a new philosophical viewpoint. Among the contributions are:

- A review of musical instrument sound production and perception from a unified viewpoint, based on the excitation and resonance structures of the sound sources and on modern hearing models.
- A psychophysical experiment testing human abilities on instrument-classification tasks using realistic—that is, musical—recordings of orchestral instruments as stimuli.
- A demonstration of the extraction of perceptual features from realistic recordings of orchestral instruments made in realistic—that is, noisy and reverberant—environments.

- A broad theory of sound-source recognition with applications to human talker identification, multimedia annotation, and other areas.
- A computational framework based on the theory, with behavior similar to that of humans in several important ways.

This dissertation is conceptually divided into three parts. The first part, consisting of Chapters 2 and 3, reviews human and machine sound-source recognition abilities, highlighting many of the constraints under which sound-source recognition systems operate. The second part, consisting of Chapters 4 and 5, describes a computational architecture for a novel model of sound-source recognition. The third part, consisting of Chapter 6, compares the abilities of the artificial system to those of humans on a variety of classification tasks.

In Chapter 2, **Recognizing sound sources**, I review the psychophysical evidence that shows that a sense of hearing is used to make inferences about objects in the world, and that these inferences are based on categorization at various levels of abstraction. I claim that knowledge of class membership can be used to help sort out the contributions of various sound sources in a complex auditory scene, and that previous research in computational auditory scene analysis has suffered by ignoring or postponing the potential contributions of sound-source recognition. I describe recognition as a process of refinement that begins at an appropriate level of abstraction and gradually becomes more concrete until sufficiently powerful inferences can be made for achieving the listener's goals. I present a set of criteria for evaluating sound-source recognition systems, and, in light of these criteria, compare the state-of-the-art in artificial systems to human abilities. I conclude with the observation that current artificial systems can recognize either a small number of sound-source classes with reasonable generality or a larger number of classes with very limited generality. One of the challenges for the rest of the dissertation—and for the next generation of sound-source recognition systems—is to increase the number of classes of sound while maintaining the ability to generalize.

In Chapter 3, **Recognizing musical instruments**, I restrict attention to a limited set of sound sources consisting of the common non-percussive musical instruments. I review the extensive literature on the production and perception of orchestral-instrument sound, highlighting the constraints of the sound production process and the perceptual limitations of human listeners. These are summarized from a viewpoint centered on the excitation and resonance structure of the instruments, which strongly supports the traditional division into instrument families. One of the core theses of this dissertation is that many sound sources—including the non-percussive orchestral instruments—are recognized primarily by perception of their excitatory and resonant structures.

In Chapter 4, **Representation**, I describe a series of representational transformations, beginning with an acoustic waveform generated by an isolated sound source and resulting in an abstract model of the source's excitation and resonance structure based on perceptually salient acoustic features. The representations are functionally matched to current models of the human auditory system, becoming increasingly speculative with each level of abstraction away from the sound wave. The description of a particular sound source is refined over time as sounds produced by the source are heard. The chapter concludes with a description of a taxonomic inheritance hierarchy that contains abstract models for a variety of sound sources. This hierarchy comprises the knowledge base used during the recognition process.

In Chapter 5, **Recognition**, I present a computational framework for sound-source recognition, based on the theory outlined in Section 1.2 and using the representation scheme described in Chapter 4. The framework has conceptual ties to the theories of decision trees, spreading activation, and taxonomic Bayesian belief networks. It employs *maximum a posteriori* classification within a taxonomy of sound-source classes. The basic algorithm is extended with context-dependent feature selection and beam search. This *improvisational* algorithm is robust, scalable, and flexible. It is sufficiently general to be expanded to a wide range of sound-source categories, and it does not depend on a fixed set of features.

In Chapter 6, **Evaluation**, the recognition framework is tested on a battery of classification tasks, and its performance is compared to that of human listeners on similar tasks. A listening experiment is performed to evaluate human abilities on musical instrument recognition tasks using both isolated tones and real music as stimuli. The model described in Chapters 4 and 5 is tested on a forced-choice classification task using the same stimuli and is shown to exhibit performance competitive with experienced musical listeners with both types of stimuli. Further, the model performs as well or better—and satisfies the evaluation criteria outlined in Chapter 2 more thoroughly—than previous sound-source recognition systems.

Finally, in Chapter 7, **Conclusions and extensions**, I evaluate the potential of the theory of recognition and identify several directions for extending the research presented here. Among the conclusions prompted by this work are that “timbre” is useless as a scientific concept, and that an ability to resynthesize acoustic wave-forms is not a necessary component of machine-listening systems.

CHAPTER 2 Recognizing sound sources

For hearing to serve as a useful sensory modality, the listener must be able to make inferences about sound-producing objects. By recognizing the kind of object that is producing a sound, a skilled listener can predict properties of other sounds the object might produce. Of course, these inferential capabilities are not limited to sonic properties. Knowledge of sound-source identity can be used to infer other characteristics of the sounding object, or can invoke behaviors in the listener himself. For example, an animal in the wild might recognize that a particular sound, a “growl,” has been produced by a large nearby predator, and this recognition might trigger a “fleeing” behavior. The inferential abilities enabled by sound-source recognition confer an immense selective advantage to animals that possess them.

This chapter has four main components. First, the complexity of the sounding world is considered, and some strategies for coping with mixtures of sounds, as revealed by research in *auditory scene analysis*, are presented. Attempts at constructing artificial listeners based on the principles of auditory scene analysis are considered, and sound-source recognition is acknowledged as an essential missing component of existing systems. Second, the constraints of the sounding world are considered, and a set of criteria for evaluating listening systems, both biological and machine, is presented. Third, the abilities of human listeners are reviewed in light of these criteria. Artificial recognition systems constructed in several domains are similarly reviewed. Finally, the common weaknesses of the artificial systems are highlighed.

2.1 Understanding auditory scenes

The sounding world is complex. In a typical environment, many objects produce sound simultaneously, and the listener must somehow organize the complicated *auditory scene* in such a way that the contributions of each sound source are comprehended. *Auditory scene analysis*, an area of psychophysical research, attempts to explain how a listener understands a continuous sound mixture as arising from a set of independent sources.

The task of auditory scene analysis is made difficult in part by sound's transparent nature. Each sound source creates small variations in the ambient air pressure—sound waves—which travel away from the source. The difficulty arises because the sound waves from independent sources arrive at the ear as a sum of the individual sound waves, and the listener has access only to the mixture. As Helmholtz observed more than a century ago:

> "The ear is therefore in nearly the same condition as the eye would be if it looked at one point on the surface of the water through a long narrow tube, which would permit of seeing its rising and falling, and were then required to take an analysis of the compound waves." (Helmholtz, 1954, p. 29)

Even without this additional complexity, auditory scene analysis has much in common with visual scene analysis, which is by no means an easy problem to solve.

2.1.1 Exploiting environmental constraints

The structure of the world places constraints on sound production. As a field of study, auditory scene analysis is concerned with identifying these constraints, their effect on sound mixtures, and possible strategies for exploiting them to aid understanding. In his book that named the field, Bregman (1990) presents a set of such constraints and strategies, along with evidence of their use by human listeners.

For example, only rarely will independent events appear to be synchronized, so sound components that start, end, or change together are likely to have arisen from the same source. The human auditory system is exquisitely sensitive to simultaneous onsets in different frequency regions, and to coherent modulation in both frequency and amplitude. Objects in the world change slowly relative to the rapid vibrations of sound waves, so two sound components proximate in time and related in some aspect (pitch, loudness, spectral content, etc.) are likely to have been produced by the same source. By this mechanism, a sequence of phonemes may be heard as a sentence unit, or a sequence of notes produced by a musical instrument may be heard as a melodic phrase.

The proximity constraint leads to the *old-plus-new heuristic*: "If you can possibly interpret any part of a current group of acoustic components as a continuation of a sound that just occurred, do so" (Bregman, 1990, p. 222). After portions of the auditory scene have been accounted for by "old" sounds, whatever is left can be interpreted as belonging to a “new” sound or sounds. This happens at two levels, including both short-term prediction based on a sound’s local properties and longer-term building of auditory *streams*.

2.1.2 The importance of knowledge

The constraints and strategies described so far do not depend on the particular contents of the auditory scene or on the listener’s world knowledge, but the kinds of sounds in a mixture and the listener’s past experience do greatly affect his perception. To account for this, Bregman introduces the concept of *schemata*, or learned patterns, which interact with the more general strategies to explain the auditory scene.

Perhaps the most compelling illustrations of the importance of world knowledge are Warren’s *phonemic restoration* examples (Warren, 1970; 1999). When a brief portion of speech sound from a recorded sentence is completely erased and replaced by an extraneous sound (e.g., a cough), listeners earnestly believe that they have heard the missing sound—indeed, they do not realize that anything is amiss. The effect applies not only to speech sounds, but also to any sound with which the listener has experience (one musical example is the restoration of notes from a melody played on a piano (Sasaki, 1980)). The effect depends on the ability of the extraneous sound to *mask*, or obscure, the neural representation of the expected but missing sound:

> “If there is contextual evidence that a sound may be present at a given time, and if the peripheral units stimulated by a louder sound include those which would be stimulated by the anticipated fainter sound, then the fainter sound may be heard as present. […] But the truly masked signal is no more, and any restoration must be considered a recreation or perceptual synthesis of the contextually appropriate sound.” (Warren et al., 1972)

The ability to infer the presence of masked sounds can be partly explained by short-term prediction based on properties of the preceding sound components, or by interpolation between components preceding and following the interrupting sound. This, however, does not explain the ability to infer entire speech phonemes as demonstrated by Warren’s examples. Clearly, high-level contextual knowledge—even, in the case of phonemic restoration, *semantic* knowledge—is used, in what Helmholtz would have called “unconscious inference” (Helmholtz, 1954). It is not clear how important these effects are to everday listening situations, but we must be careful not to underestimate their significance.

2.1.3 Computational auditory scene analysis

Over the last decade, several researchers have attempted to build computational frameworks that perform auditory scene analysis; the resulting field has been called *computational auditory scene analysis* (CASA). Typically, CASA research projects have involved implementation of some small subset of the strategies suggested by Bregman, often in a manner functionally consistent with the early stages of the human auditory periphery (as they are currently understood).

Ellis (1996) describes several of these systems, with references to their original presentation in the dissertations of Cooke (1993), Brown (1992), and Mellinger (1991), and a paper by Ellis (1994), as instances of a single structural framework. According to his analysis, the overall structure can be broken into four main sections that proceed in sequence (illustrated in Figure 2):

1. **Front-end:** All of the systems employ a filter-bank to break the acoustic signal into different frequency bands. In the human auditory periphery, this function is performed in the cochlea, and this organization by frequency region is preserved at higher levels of the auditory system. Each system includes further processing intended to reveal particular acoustic properties, or "cues." For example, an "onset map" might be generated to facilitate later grouping by common onset.
2. **Basic representation:** In this second stage, the output of the front-end, including the cue detectors, is organized into discrete elements, the "atoms" which make up auditory objects. Typical elements include "tracks," representing stable sinusoids that may correspond to harmonic partials, and "onsets," representing abrupt rises in energy that may correspond to the start of a new sound.
3. **Grouping algorithm:** In the third stage, a subset of Bregman's strategies is employed to group elements (from the basic representation) that correspond to coherent auditory objects. For example, "tracks" with simple frequency relationships may form a group corresponding to a harmonic sound.
4. **Output assessment / resynthesis:** In the final stage, the group representations from the third stage are converted into a form suitable to the goals of the system. In some cases, these are acoustic waveforms corresponding to the "separated" auditory objects.

These early CASA systems suffer from several critical limitations, attributed (by the respective authors, as well as by Ellis) to many factors, including: inadequate cues, inextensible algorithms, rigid evidence integration, and inability to handle obscured (masked) data.

Ellis attempted to address these limitations by introducing short-term prediction, based on the statistical properties of low-level sound objects (noise clouds, transients, and quasi-periodic tonal *wefts*), to infer masked or obscured information (Ellis, 1996). His approach, called *prediction-driven computational auditory scene analysis* (PDCASA), is remarkably successful at grouping low-level time-frequency energy into perceptually salient objects—for example, car horns and slamming doors in a complex, noisy street scene. In a limited test on a few examples, the PDCASA system exhibited good correspondence to human responses regarding the number of objects in the scene (but not their identities). The PDCASA system infers the properties of masked sounds to a small degree, but it is a long way from solving problems like phonemic restoration.

These CASA systems have been constructed with little concern for the actual contents of the auditory scene. However, the *kinds* of sounds in the mixture, and the listener's past experience with similar sounds, can have an enormous effect on how an auditory scene is perceived by a human listener. The research presented in this dissertation is a framework for representing and exploiting knowledge about sound sources. Although the framework presented here does not constitute a CASA system, it is intended to be integrated with one. By recognizing the source of a sound, a CASA system would be better equipped to infer the sound's masked properties. Sound-source recognition is an essential yet largely overlooked component of auditory scene analysis.

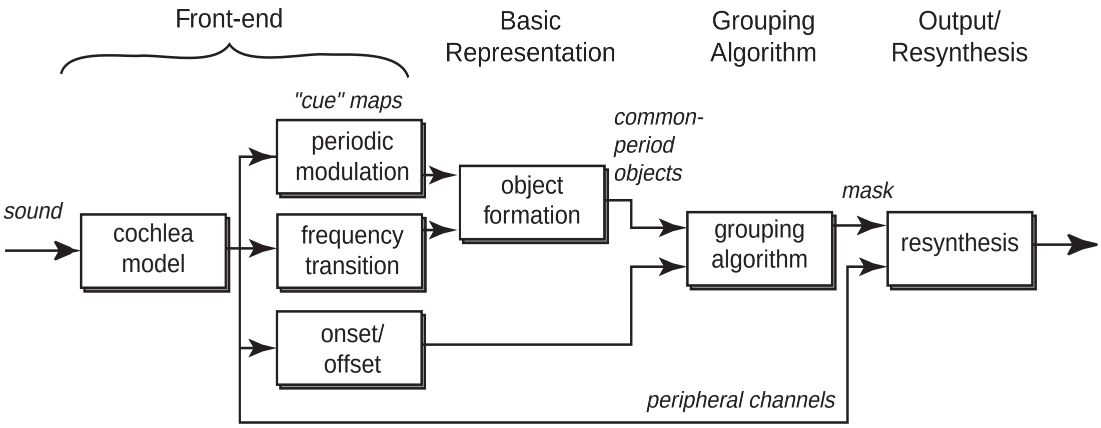

FIGURE 2. Overview of processing flow in CASA architectures, after Ellis (1996).

2.2 Evaluating sound-source recognition systems

Although many sound-source recognition systems have been constructed, it is often very difficult to be objective in evaluating the success of a computational system at recognizing sounds. The performance of an individual system is often quantified by its creator as a percentage of "correct" responses in some kind of test scenario, but the scope of the test—and, indeed, the scope of the system—is not often expressed.

There are several dimensions along which systems differ in competence, and although they are not easy to quantify, they should be considered carefully when comparing the abilities of different systems. The following criteria are presented in roughly descending order of importance.

A sound-source recognition system should:

1. **Exhibit generalization.** Different instances of the same kind of sound should be recognized as similar. For example, a system that learns to recognize musical instruments should be able to do so in a way that does not depend on the particular human performer or the particular acoustic environment. Though they may differ in quality, a clarinet played by a student in a dormitory is as much a clarinet as one played by Richard Stoltzman in Bos-

ton Symphony Hall. The ideal degree of perceived similarity may be context-dependent. For example, in an office it may be important to recognize which *particular* telephone is ringing, whereas in a home it is sufficient to recognize that it is the telephone, and not the doorbell, that is producing sound. In the first situation, a system trained to recognize one particular telephone may suffice, but in the second, it would be more useful for the system to recognize the *class* of sounds telephones make, independent of any particular telephone.

2. **Handle real-world complexity.** Too often, psychoacoustic experiments employ simple stimuli that have little relation to sounds that occur in the environment. As a result, many computational listening systems are tested only with simple stimuli, such as additive synthesis tones, sine waves, bursts of white noise, sounds recorded in an anechoic chamber, and so forth. If these systems are tested on ecological signals—signals that actually occur in the real world—it is quickly discovered that the system cannot handle the additional complexity, noise, temporal characteristics, etc. Many theories can be made to work on "laboratory sounds" or in thought experiments, but most fail if tested in real-world scenarios. In a real-world environment, sounds are rarely heard in isolation, and acoustic reflections and reverberation nearly always affect the sound waves arriving at a microphone or eardrum. Systems limited to recognizing isolated sources or sound with very little reverberation can be useful—as are, for example, current commercial speech recognition systems—but these limitations must be considered when comparing systems.
3. **Be scalable.** The world contains a vast array of sound-producing objects, and it is hard to pin down even the order of magnitude of the number of different sounds mature human listeners can recognize. In contrast, a typical sound-recognition system may be trained on only a few kinds of sounds—perhaps a few *tens* of sound classes. To evaluate a system with such limited knowledge, it is necessary to consider the *competence of the approach*—is the system *capable* of learning to recognize additional sounds, and how would such expansion affect its performance? Different sounds may have different salient characteristics. It may be important to consider whether a system's repertoire of *feature detectors* can be expanded when additional acoustic properties become important.
4. **Exhibit graceful degradation.** As the level of ambient noise, the degree of reverberation, or the number of competing sound sources increases, human sound-source recognition performance gradually worsens. In contrast, many machine systems stop working altogether when a certain level of degradation, abiguity, or obscurity is reached. In realistic scenarios, these complicating factors obscure portions of the "target" sound. In order to continue functioning successfully, a system must have a strategy for handling what has been called the "missing feature problem;" it must be able to recognize the whole from a portion.
5. **Employ a flexible learning strategy.** Machine systems that learn are often classified by whether their learning is *supervised* or *unsupervised.* In the former case, an omniscient trainer specifies the *category* of each training

example at the time of training; in the latter, the system must discover the categories itself. This division is in itself artificial, however; human listeners make use of both “labeled” and “unlabeled” data as they learn. Many machine systems do all of their learning in a large batch, and then remain fixed as they operate. In contrast, human listeners learn continually, introducing new categories as necessary and refining classification criteria over time as new examples of previously learned categories are encountered.

6. **Operate in real-time (in principle).** One of the defining characteristics of biological listeners is that they interact with their environment on the same time scale as the sounds they attend to. In contrast, many computational systems rely on having pre-selected segments of sound presented for analysis. For example, some “music analysis” systems require that the entire piece of music be presented at once. One of the essential aspects of music, however, is that it takes place over time, setting up expectations in the listener and then either satisfying them or invoking surprise. Requiring an artificial system to operate in real-time is too limiting, yet systems should not require human intervention in the form of segmentation into chunks to be processed. To be considered a *listener,* a system should be *real-time in principle.* It should analyze the acoustic waveform sequentially through time, as it would arrive at a microphone or eardrum.

These six criteria must be considered before comparing the quantitative performance of different systems. Other criteria, which do not bear directly on performance, may also be worth considering. For example, two systems that perform equally well on some task and have similar ratings on the foregoing criteria may still be compared on the basis of complexity: all other things being equal, the *simpler* system is better. This simplicity can be in the form of reduced memory size or processing requirements, or in how easy it is to understand how the system works.

Further, if the goal of building a machine listening system is not to achieve a particular level of competence on a given task, but rather to gain insight into the workings of a human or animal listener, it is important to consider the similarity between the biological system and the model. In his influential work on models of the human visual system, David Marr identified three conceptual levels at which information-processing systems can be understood (Marr, 1982). The first, and most abstract, is the *computational theory,* where questions of *what* the system does and *why* are considered. At the second level, the *representation* and *algorithm* are considered, and the forms of the system’s input and output, along with a method of proceeding from one to the other, are detailed. At the third and most concrete level, the particular *hardware implementation* is considered. The three levels are loosely related, and systems may be compared at any or all of them. The approach presented here, and its relation to the human auditory system, will be considered in Chapter 4.

A final criteria, one that in many cases should *not* be used to evaluate machine listening systems, is an ability to *reproduce* the sounds it hears. Recognition often requires much less information than reproduction, and although high-fidelity reproduction may be a useful feature, it is in no way a requirement for good recognition performance. However, if sound-source recognition is to be used as part of a CASA system, it is important to be able to use source identity to infer the masked acoustic properties of the sound at some level of representation (though most likely not at the waveform level).

As a final note, it is important to keep in mind the semantic differences among the words *classification*, *identification*, and *recognition*. In this thesis, *recognition* describes a process of gathering information and making inferences, and *classification* involves the assigment of a category label. *Identification* is used to describe recognition tasks in which the “allowed” category choices are not prespecified. In Chapters 4 and 5, a *recognition* model will be described. In Chapter 6, it will be tested on *classification* tasks. Readers interested in the subtle differences between the words may find Sayre’s (1965) account to be of interest.

2.3 Human sound-source recognition

Humans can identify many events and objects by sound alone. Our sound-recognition abilities are either innate or learned very early in development, and we are unable to introspect about how they work. This is an example of what Minsky calls the *amnesia of infancy*: “In general, we’re least aware of what our minds do best” (Minsky, 1986, p. 29). Recognizing objects in the environment is an essential survival skill, and nearly all vertebrates recognize sounds (Popper & Fay, 1997). In spite of this, and perhaps because of their introspective opacity, the processes underlying sound-source recognition have not been studied in depth. Much of what we know has been learned indirectly, from psychophysical experiments aimed at narrower phenomena. The discussion in this section draws from two recent, complementary, reviews of such research (Handel, 1995; McAdams, 1993).

If a particular sound source generated the same sound wave every time, recognition would be easy—we could simply (at least in principle) memorize every sound and match incoming sound waves to stored patterns in memory. In reality, there is enormous variability in the acoustic waves produced by any given sound source at different times. This variation is due in part to the complexity of the environment—for example, a room’s detailed acoustic response changes with the movement of any object, with changes in air circulation, and even with shifts in humidity! Natural sounds—that is, sounds not produced by human artifacts—vary even more from instance to instance because the physical process of sound production is never the same twice.

The listener must abstract away from the raw acoustic signal in order to discover the identity of a sound event. Although there is much variability in the acoustic signal, there are often *invariants*—things that do not change from instance to instance—in the sound-production process. For example, the kind of excitation—the way that energy is injected into the physical system, for example by banging, blowing, or scraping—affects the acoustic signal in many ways, both subtle and obvious. The material properties and geometry of the vibrating body impose constraints in a similar but complementary way; for example, they affect the frequency spectrum, onset and offset transients, and transitions between sounds. By using features that are influenced by these production invariants, it should be possible to work backward to the invariants themselves, and from there to sound event identity. Both Handel and McAdams suggest that inference based on the detection of invariants is the most likely basis for human sound-source recognition. It is important, however, to look more deeply than "trivial" invariants, such as sound-source identity, that entirely beg the question.

Because excitation and resonance properties simultaneously influence the properties of the sound wave, there are many potential acoustic features to be used for recognition. As a consequence, there is no one predominant cue, separate cues are not entirely independent, and the cues a listener actually uses are highly dependent on the context. Particularly when multiple sounds overlap, it will be difficult to know in advance which cues will be available—therefore, the listener's recognition strategy must be flexible.

McAdams describes recognition as a range of phenomena:

> "Recognition means that what is currently being heard corresponds in some way to something that has been heard in the past…. Recognition may be accompanied by a more or less strong sense of familiarity, by realizing the identity of the source (e.g., a car horn), and often by an understanding of what the source being heard signifies to the listener in his or her current situation, thereby leading to some appropriate action." (McAdams, 1993, p. 147)

His conception of the recognition process, in abstract form, is shown in Figure 3. (Note the similarities with Figure 2.) McAdams suggests that the process is largely sequential: the sound wave is changed, by transduction, into a representation where auditory grouping can take place. Grouped elements are analyzed in terms of some set of features, which are then used as the basis of the recognition process. Although McAdams suggests that recognition is subsequent to the grouping processes of auditory scene analysis, he leaves room for the possibility of feedback from higher, post-recognition processes—this feedback loop is clearly necessary to account for phenomena such as phonemic restoration.

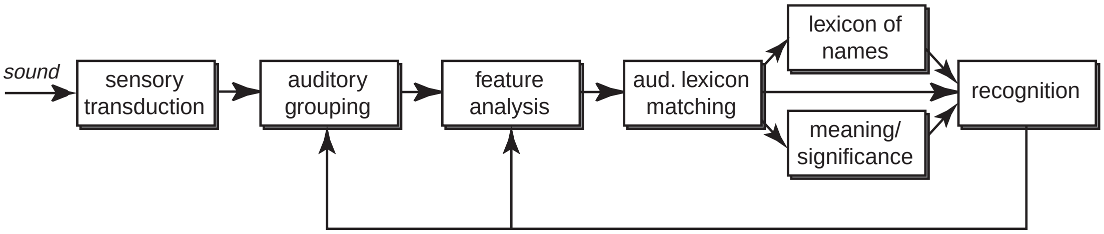

FIGURE 3. Overview of the stages of auditory processing for sound source recognition, after McAdams (1993).

Psychological studies have shown that human object recognition—in all sensory modalities—occurs at multiple levels of abstraction. Minsky terms these *level bands*, and suggests that one or more intermediate levels of abstraction are privileged in recognition (Minsky, 1986). To paraphrase his words, beyond a certain level of detail, increasingly detailed memories of previously observed objects are increasingly difficult to match to new situations. Above a certain degree of abstraction, descriptions are not detailed enough to be useful—they do not provide any discriminating information.

This idea is similar to Rosch's *basic level* (Rosch, 1978; Rosch et al., 1976). Her research suggests that the kinds of objects in the world form hierarchies in the mind and that there is a privileged level—the "basic" level—where recognition initially takes place. The basic level is where the most information can be gained (the best predictions or inferences can be made) with the least effort. Basic objects can be shown "to be the first categorization made during perception of the environment, to be the earliest categories sorted and earliest named by children, and to be the categories most codable, most coded, and most necessary in language" (Rosch et al., 1976). To take an example from audition, a sound heard while driving a car might be recognized as a "bad engine noise" before being classified as a misfiring spark plug.

Minsky suggests that objects may be organized into multiple hierarchies that classify them in different ways. The particular hierarchy used in a given situation may depend on the context, as may the particular level that is privileged. These may depend on the set of features currently available from the sensory input and on the current goals of the perceiver. These shifts of level and of hierarchy happen very quickly and are mostly inaccessible to introspection.

We should not neglect the feedback mechanisms suggested by McAdams's proposed architecture and their importance in thinking about auditory scene analysis. Some high-level influences are obvious. Every human listener is exquisitely sensitive to hearing his or her name, even in complex, noisy environments. There is a great deal of anecdotal evidence that multilingual speakers can understand speech in their over-learned native language relatively easily in adverse environments—they need a higher signal-to-noise ratio to understand speech in their secondary languages.

More subtly, we use what we know about a particular sound source to fill gaps in the available sensory data. As in Warren's auditory restoration phenomena, we fill in details with default assumptions based on our expectations. This process is entirely inaccessible to our consciousness; we are not aware that we are doing it, and we believe that we are hearing more detail than is actually there to be heard. Our perception is a blending of information from sensations and expectations. Indeed, the feedback loops in McAdams's architecture are essential.

Human listeners outpace machine systems on every criterion considered in Section 2.2. We are able to recognize instances from a very large number of general classes, in real-world acoustic conditions and under wide ranges of complexity arising from mixtures of simultaneous sounds. Human recognition degrades gracefully as conditions worsen. Our learning is extremely flexible—we can find structure in the world without being given a label for every object, and we learn continually, adding new object classes throughout our lives. In addition to such "unsupervised" learning, we can learn new classes by instruction—"Can you hear that unusual sound in the mix? It's a digeridoo." And in many cases, we need only a few examples—sometimes only one—to learn a new category (Sayre, 1965). To top it off, our brains work in real-time, and not just in principle.

2.4 Machine sound-source recognition

Many systems have been built to recognize sounds in different domains. To name a few, systems have been constructed to keep track of when particular advertisements are played on a television or radio station, to discriminate speech sounds from music, to identify talkers on a telephone, and to recognize musical instruments in a recording. In this section, sound-source recognition systems from several domains will be considered and evaluated in light of the criteria proposed in Section 2.2. Only those machine-listening systems whose goal is to recognize sound sources from airborne sound waves will be presented. Automatic speech recognition systems, where the goal is to recover the *message* rather than the identity of the talker will not be considered.

2.4.1 Recognition within micro-domains

Several systems have been constructed to recognize examples from very small classes of sounds. A typical example of such a system is one constructed to recognize different kinds of motor vehicles from the engine and road noise they produce (Nooralahiyan et al., 1998). First, a human-selected segment of sound waveform is coded by a linear prediction algorithm (LPC). Then, the LPC coefficients are presented to a time delay neural network (TDNN) that classifies the source of the sound waveform as belonging to one of four categories (roughly, trucks, sedans, motorcycles, and vans).

The authors performed two studies: one with sounds recorded carefully in isolated conditions, to evaluate the propriety of the feature set; and one with sounds recorded on a city street, to evaluate the system in more realistic conditions. In both cases, supervised learning was used. For the city street case, the system was trained with 450 sounds and tested with 150 independent sounds. The system's performance was substantially above chance, with correct classification of 96% of the training samples and 84% of the test samples. The TDNN has apparently found some kind of regularity in the features that enables classification, but as is typical of much connectionist research, no attempt was made to discover exactly which aspects of the features were salient.

Examples of systems with similar scopes include non-connectionist approaches to recognition of songs in movie soundtracks (Hawley, 1993; Pfeiffer et al., 1996).

There are a few examples of "implicit" recognition systems constructed by researchers who were investigating the sound-recognition abilities of humans. For example, while attempting to understand how people recognize the sex of a person by listening to his/her footsteps, Li et al. (1991) identified a set of acoustic properties that correlate with human judgments of walker sex. They used principal-components analysis (PCA) to reduce the dimensionality of the feature space and constructed a discriminator that correlated strongly with human judgments (r=0.82, p<0.05). Another example is a study on human judgments of mallet hardness from the sounds of struck metal pans (Freed, 1990).

Micro-domain recognition systems vary greatly in their ability to generalize from training samples. This variability can stem from a choice of analysis features that does not adequately capture the structure of the sound class, or from a too-narrow range of training examples. Some systems are limited to recognizing pristine recordings of isolated sounds, but others adapt well to real-world noise. None, however, are equipped to deal with mixtures of sounds.

Most micro-domain systems employ techniques from statistical pattern-recognition (e.g., neural networks or maximum-likelihood classifiers) within a supervised learning framework. As with nearly all artificial sound source recognition systems, the sound samples used to train and test these systems are pre-selected (and even pre-segmented, thereby eliminating real-time applications) by human operators. Most often, the systems are not given a "don't know" option for cases when a sound sample falls outside their domain of knowledge. It is uncertain whether micro-domain approaches can scale to larger numbers of classes, not only because their range of feature-detectors may be too small, but also because their recognition frameworks are relatively inflexible.

2.4.2 Recognition of broad sound classes

A typical example of recognizing examples from broad sound classes is speech/music discrimination, which has applications in automatic speech recognition and soundtrack segmentation, for example. There are many examples of such systems (e.g., Spina & Zue, 1996; Scheirer & Slaney, 1997; Foote, 1997; Han et al., 1998; Minami et al., 1998), but the system described by Scheirer and Slaney appears to be the most general and the best able to handle real-world complexity.

Scheirer and Slaney considered 13 features and extensively tested four different multidimensional classification frameworks with various feature combinations. An extensive corpus of training and test data was recorded from FM radio stations in the San Francisco Bay area, covering a variety of content styles and noise levels. Several twenty-minute sets of data were recorded, each consisting of 80 hand-labeled, fifteen-second samples.

In each classifier, learning was supervised, using 90% of the samples in a set for training, and reserving 10% for testing (never splitting a 15-second sample). The best classifier, which used only 3 of the 13 features, had 5.8% classification error on a frame-by-frame basis, and the error rate dropped to 1.4% by integrating several frames (over 2.4 seconds). All of the classifiers tested were capable of real-time performance in principle, and the best classifier was able to run in real-time in software on a workstation. As is true in most domains where appropriate features are selected, the particular classification technique did not affect performance—several different algorithms gave rise to similar performance levels.

At least one system has been built that expands the range of allowable categories beyond music and speech in a sound-retrieval application (Wold et al., 1996). It allows a human user to specify an arbitrary class of sounds by providing a small number of examples. The system uses a feature vector made up of perceptually motivated acoustic properties (for example, correlates of loudness, pitch, brightness, bandwidth, and harmonicity, as well as their variation over time) to form a Gaussian model for the sound class. It then uses the Mahalanobis distance (which takes into account the relative ranges of the various features, and also inter-feature correlation) to retrieve similar sound examples from a database of recordings.

It is difficult to evaluate the performance of a system on such a subjective task, but the authors give several examples of intuitively reasonable classification based on categories such as laughter, female speech, and telephone touch-tones. The approach seems appropriate for general, high-level classes, but because it uses only gross statistical sound properties, it may not be able to make fine class distinctions (e.g., particular human talkers or musical instruments) without considerable additional front-end complexity.

Like the micro-domain examples, broad-class systems such as these employ statistical pattern-recognition techniques within a supervised learning paradigm. In some cases, they have demonstrably generalized from their training examples and can recognize new examples drawn from the classes they have learned. The systems described above operate on real-world recordings, using surface properties of sound mixtures rather than features of isolated sounds—indeed, ignoring the fact that the sounds are typically mixtures. It is difficult to judge the scalability of these systems. The features used in the speech/music discrimination systems are specifically tuned to the particular task; Scheirer and Slaney even point out that the features do not seem to be good for classifying musical genre. The sound-retrieval system seems to be more flexible, but quantitative test results have not been published. This is emblematic of the vast quality differences between evaluation processes. Extensive, quantitative cross-validation, as performed by Scheirer and Slaney, is necessary for honest system evaluation, but too often it is sidestepped.

2.4.3 Recognition of human talkers

Many systems have been built to identify human talkers (Mammone et al., 1996 gives an overview of several different approaches). Most employ statistical pattern-recognition techniques within a supervised-learning framework, using input features motivated by consideration of human perception. The research described by Reynolds (1995) is typical of the scope of such systems.

Reynolds's system, like many others, uses mel-frequency cepstral coefficients (MFCC) as input features. These coefficients, in this case based on 20 ms windows of the acoustic signal, are thought to represent perceptually salient aspects of human vocal-tract resonances (*formants*); the frequencies and bandwidths of these resonances are known to be important for talker identification by humans (Brown, 1981). Given a recorded utterance, the system forms a probabilistic model based on a mixture of Gaussian distributions. During training, these models are stored in memory. To recognize a novel utterance, the system finds the model that is most likely to have produced the observed features.

The performance of the system depends on the noise characteristics of the signal, and on the number of learned models (the *population size*). With pristine recordings, performance is nearly perfect on population sizes up to at least 630 talkers (based on experiments with the TIMIT database). Under varying acoustic conditions (for example, using telephone handsets during testing that differ from those used in training), performance smoothly degrades as the population size increases; on the Switchboard database, correct classification rates decreased from 94% to 83% as the population size grew from 10 to 113 talkers.

Systems constructed to date have relied on only a subset of the acoustic properties human listeners use for talker identification. Approaches that use only low-order cepstral coefficients do not have access to information about the fundamental frequency of the speaker's voice, which is known to be an important cue for human listeners (Brown, 1981; van Dommelen, 1990). Speech rhythm, which is also a salient cue for humans (van Dommelen, 1990), has not been used in systems built to date.

Talker identification systems suffer from lack of generality—they do not work well when acoustic conditions vary from those used in training. From that perspective, they do not handle real-world complexity adequately. Also, they recognize only utterances from isolated talkers; they can not deal with mixtures of sounds. The approaches used in these systems scale reasonably, to much larger numbers of sound classes than systems in the other domains considered so far, but performance suffers as the population size grows.

2.4.4 Recognition of environmental sounds

Although few systems have been built to recognize specific sound sources other than human talkers or musical instruments, two such systems are worthy of mention. The Sound Understanding Testbed (SUT) recognizes instances of specific household and environmental sounds (Klassner, 1996), and Saint-Arnaud's system recognizes sound textures (Saint-Arnaud, 1995).

SUT was constructed as a trial application for the Integrated Processing and Understanding of Signals (IPUS) blackboard architecture, which implements a simultaneous search for an explanation of a signal and for an appropriate front-end configuration for analyzing it (Klassner, 1996). SUT operates in an audio analog of the "blocks world" of vision AI. Whereas early AI systems performed visual scene analysis in highly constrained environments, SUT performs auditory scene analysis on mixtures of sounds from a small library of sources. The IPUS architecture and the knowledge base in SUT are constructed to be very clever about applying signal-processing domain knowledge to identify distortions arising from particular settings of the front-end signal-processing network and to adapt to them.

SUT employs several levels of feature abstraction, based in large part on sinusoidal-analysis techniques. Representations begin with the spectrogram and intensity envelope, and continue through “peaks” representing narrow-band portions of the spectrum, “contours” made up of groups of peaks with similar frequencies, to “micro-streams” made up of sequences of contours, and finally to “streams” and “sources.”

SUT has a library of 40 sounds that it can recognize. Each sound model (consisting, for example of several “streams”) was derived by hand from at least five instances of each sound. Each model represents a particular instance of a sound source rather than a general class (e.g., the sound of one viola note rather than the class of all viola sounds). The collection of models is eclectic, including two particular alarm clocks (one analog bell-and-ringer style and one electronic), a bell, a bicycle bell, a bugle call, a burglar alarm, a car engine, a car horn, a chicken cluck, a “chime,” a clap, a clock chime, a clock tick, a cuckoo clock, a doorbell chime, a door creak, a fire engine bell, a firehouse alarm, a foghorn, a set of footsteps, a glass clink, a gong, a hairdryer, a door knock, an oven buzzer, an owl hoot, a pistol shot, a police siren, an electric razor, two smoke alarms, a telephone dial, a telephone ring, a telephone dial tone, a triangle strike, a truck motor, a vending machine hum, a viola note, and the wind.

SUT was tested on mixtures constructed by placing four independent sounds from the library randomly in a five second recording. Two conditions were tested. In one, SUT was given a minimal library consisting of just the sound sources actually present in the recording; in the second, all 40 models were provided. The system’s task was to identify which sounds occurred and when. A correct identification was credited when SUT chose the right model and estimated a time range that overlapped the actual time of sounding. In the first scenario, the system identified 61% of the sounds correctly; in the second, the recognition rate dropped slightly to 59%. No information has been reported about the kinds of mistakes that were made (for example, whether one telephone was confused with the other).

Because of the limited available information, it is difficult to evaluate SUT’s performance as a recognition system. Based on the simplicity of the sound models and the limited range of training data, it is likely that SUT can only recognize the particular sound instances it was trained with, rather than the general classes those sounds represent. In the evaluation process, real world complexity was limited to artificially-produced mixtures of sounds. Although SUT’s success on such mixtures is praiseworthy, it should not be taken as a prediction of performance on naturally occurring sound mixtures. Learning in SUT takes place only in the form of hand-coded source models, and it is not clear whether the range of models could be expanded while maintaining the current performance level. On the other hand, SUT is the first system to attack the auditory scene analysis problem with extensive world-knowledge, and as such, it is a step in the right direction.

Saint-Arnaud explored a range of little-studied sounds that he termed *textures* (Saint-Arnaud, 1995). He likens sound textures to wallpaper: they may have local structure and randomness, but on a large scale the structure and randomness must be consistent. Examples of sound textures include bubbling water, the noise of a photocopier, and a large number of whispering voices. Saint-Arnaud collected a set of sample textures and performed a psychophysical experiment to determine whether humans perceived textures as members of high-level classes. Indeed, he found that people classify sound textures by the kind of source, such as water, voices, or machines, and by acoustic characteristics, such as periodicity or noisiness.

After studying human responses, Saint-Arnaud attempted to build a computer classifier that might match them. He used a cluster-based probability model on the patterns of energy outputs of a 21-band constant-Q filter bank to form models for segments from 12 recordings of different sound textures. Using a custom “dissimilarity” metric, the system compared models derived from test samples to stored models from the training samples, assigning the test sample the high-level class of the closest training sample. Fifteen samples were tested, including additional segments from the 12 training sounds. Three of the test samples were misclassified. Saint-Arnaud warns against drawing any general conclusions from this small example, but suggests that the results are encouraging.

2.4.5 Recognition of musical instruments

Several musical instrument recognition systems have been constructed during the last thirty years, with varying approaches, scopes, and levels of performance. Most of these have operated on recordings of single, isolated tones (either synthesized or natural), but the most recent have employed musical phrases recorded from commercial compact discs.

De Poli and his colleagues constructed a series of Kohonen Self-Organizing-Map (SOM) neural networks using inputs based on isolated tones (Cosi et al., 1994a,b,c; De Poli & Prandoni, 1997; De Poli & Tonella, 1993). In each case, one tone per instrument was used (with up to 40 instruments in a given experiment), with all tones performed at the same pitch. Various features of the tones (most often MFCC coefficients) were used as inputs to the SOM, in some cases after the dimensionality of the feature space was reduced with principal components analysis. The authors claim that the neural networks can be used for classification, but in no case do they demonstrate classification of independent test data.

In a project of similar scope, Feiten and Günzel (1994) trained a Kohonen SOM with spectral features from 98 tones produced by a Roland SoundCanvas synthesizer. They authors claim that the network can be used for retrieval applications, but no evaluable results are provided.

Kaminskyj and Materka (1995) compared the classification abilities of a feed-forward neural network and a k-nearest neighbor classifier, both trained with features of the amplitude envelopes of isolated instrument tones. Both classifiers achieved nearly 98% correct classification of tones produced by four instruments (guitar, piano, marimba, and accordion) over a one-octave pitch range. Although this performance appears to be excellent, both the training and test data were recorded from the same instruments, performed by the same players in the same acoustic environment. Also, the four instruments chosen have very distinctive acoustic properties, so it is unlikely that the demonstrated performance would carry over to additional instruments or even to independent test data.

Langmead (Langmead, 1995a,b) trained a neural network using several instrument-tone features based on sinusoidal analysis. He writes "the trained network has shown success in timbre recognition" (Langmead, 1995a), however, no details are provided.

At least two authors have applied traditional pattern-recognition techniques to the isolated-tone classification problem. Bourne (1972) trained a Bayesian classifier with perceptually-motivated features, including the overall spectrum and the relative onset times of different harmonics, extracted from 60 clarinet, French horn, and trumpet tones. Fifteen tones were used to test the system (8 of which were not used in training), and the system correctly classified all but one (approximately 93% correct classification). More recently, Fujinaga (1998) trained a k-nearest neighbor classifier with features extracted from 1338 spectral slices representing 23 instruments playing a range of pitches. Using leave-one-out cross-validation with a genetic algorithm to identify good feature combinations, the system reached a recognition rate of approximately 50%.

In an unpublished report, Casey (1996) describes a novel recognition framework based on a "distal learning" technique. Using a commercial waveguide synthesizer to produce isolated tones, he trained a neural network to distinguish between two synthesized instruments (brass and single-reed) and to recover their synthesizer control parameters. His approach can be viewed as modeling the dynamics of the sound source, and as such may be thought of as a variant of the motor theory of speech perception. Although "recognition" results were not quantified as such, the low "outcome error" reported by Casey demonstrates the success of the approach in the limited tests.

Several authors working on CASA research have built systems that can be considered as instrument recognizers. Brown and Cooke (1994) built a system that used similarity of "brightness" and onset asynchrony to group sequences of notes from synthesized brass/clarinet duets. Segregation was successful on 9 out of 10 notes in a short example, but the instruments were not recognized *per se*.

Kashino and his colleagues have constructed a series of systems to perform polyphonic pitch tracking on simple music. Their earliest system, using harmonic mistuning and onset asynchrony, correctly recognized the source of 42 flute and cembalo notes played by a sampling synthesizer (Kashino & Tanaka, 1992).

Later systems, using more features, were able to identify the sources of notes produced by clarinet, flute, piano, trumpet, and violin in “random chords” (Kashino et al., 1995; Kashino & Tanaka, 1993). The authors used an unusual evaluation metric, but reported intriguing results. Their most recent systems, using adaptive templates and contextual information, transcribed recordings of a trio made up of violin, flute, and piano (Kashino & Murase, 1997; 1998). When the pitch of each tone was provided, the system identified the source of 88.5% of the tones in a test recording. An auxiliary report suggested the use of a hierarchy of sound models—a “sound ontology”—to enable recognition of a larger range of sound sources, but no new recognition results were reported (Nakatani et al., 1997).

Until very recently, there were no published reports of musical instrument recognition systems that could operate on realistic musical recordings, but three such systems have been described in the last two years. In all three cases, the authors applied techniques commonly used in talker-identification and speech recognition.

Brown (1997a,b; 1999) has described a two-way classifier that distinguishes oboe from saxophone recordings. A Gaussian mixture model based on constant-Q cepstral coefficients was trained for each instrument, using approximately one minute of music each. On independent, noisy samples from commercial recordings, the system classified 94% of test samples correctly. Brown has extended this work with a four-way classifier that distinguishes among oboe, saxophone, flute, and clarinet (Brown, 1998b,c), getting “roughly 84%” correct classification on independent test data (Brown, 1998a, personal communication).

Dubnov and Rodet (1998) used a vector-quantizer based on MFCC features as a front-end to a statistical clustering algorithm. The system was trained with 18 short excerpts from as many instruments. No classification results were reported, but the vector-quantizer does appear to have captured something about the “space” of instrument sounds. Although there is not enough detail in the paper to evaluate the results, the approach seems promising.

Marques (1999) constructed a set of 9-way classifiers (categories were bagpipes, clarinet, flute, harpsichord, organ, piano, trombone, violin, and “other”) using several different feature sets and classifier architectures. The classifiers were trained with recordings of solo instruments from commercial compact discs and “non-professional” studio recordings, and were tested with independent material taken from additional compact discs. The best classifiers used MFCC features, correctly classifying approximately 72% of the test data. Performance dropped to approximately 45% when the system was tested with “non-professional” recordings,[1] suggesting that the classifier has not generalized in the same way as human

[1] The “non-professional” recordings were a subset of the student recordings described in Chapter 6. They were made in a non-reverberant space (the control room of a recording studio) with a high-quality cardioid microphone placed approximately one meter in front of the musician. (I traded recordings with Marques on one occasion.)

listeners (who do not have difficulty recognizing the instruments in the “non-professional” recordings, as will be demonstrated in Chapter 6).

Perhaps the biggest problem in evaluating musical-instrument recognition systems is that very few systems have been extensively evaluated with independent test data. Until such testing is done, one must not assume that these systems have demonstrated any meaningful generality of performance.

2.5 Conclusions and challenges for the future

Human listeners outpace machine systems on every criterion considered in Section 2.2. The recognition machinery in the human brain is well suited—much more so than any artificial machinery we know how to build—to the complex acoustic environments we inhabit. Currently, we can build artificial systems that can recognize many different sound sources under laboratory conditions or a very small set of sources under more relaxed conditions. Figure 4 (next page) positions recognition systems from the domains considered in Section 2.4 on these two critical axes. The challenge that faces us is to build systems that can recognize more classes of sound sources with increased generality and under conditions of real-world complexity. The framework described in the following chapters extends the range of artificial systems, reducing the gap between humans and machines.

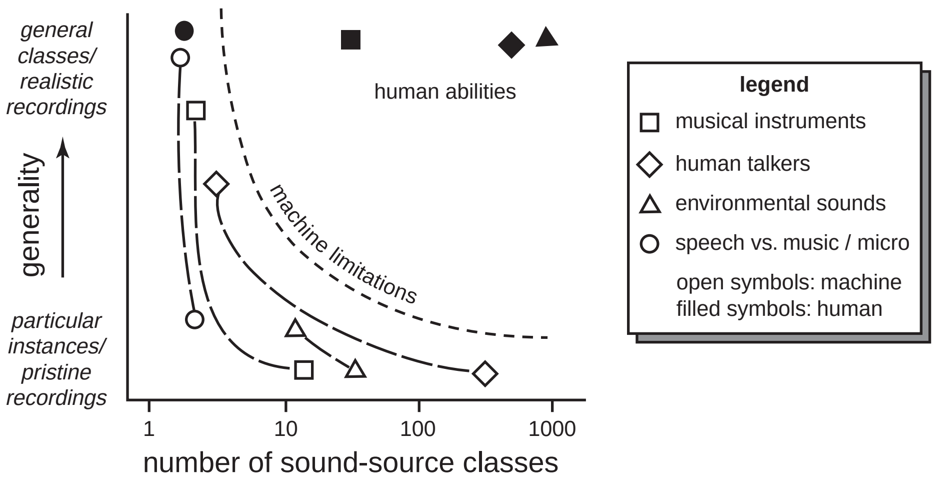

FIGURE 4. Comparison of human and machine abilities. Humans are much better able to recognize—across the board—general classes of sounds than are the current state-of-the-art in machine systems, particularly as the number of sound-source classes under consideration grows beyond three or four.

CHAPTER 3 Recognizing musical instruments

The most difficult tasks in building a successful information-processing system are discovering the constraints underlying the problem domain and determining which features arising from the constraints are best adapted to the task at hand. As David Marr writes:

> “[F]inding such constraints is a true discovery—the knowledge is of permanent value, it can be accumulated and built upon, and it is in a deep sense what makes this field of investigation into a science” (Marr, 1982, p. 104)

For this thesis, I chose the goal of recognizing musical instruments in large part because so much prior research had been done to uncover the constraints and features exploited by human listeners. In no other area of hearing research—with the possible exception of speech—have the relevant acoustics and psychoacoustics been studied in such depth. Much is known about musical instrument sounds, particularly those sounds produced by traditional Western orchestral instruments, so it is with these sound sources that the rest of this dissertation is primarily concerned.

This chapter has four sections. First, we will consider human abilities on the task of recognizing Western orchestral instruments. Second, relevant research in musical acoustics, psychophysics, and analysis-by-synthesis will be considered. Much of this research can be unified within a framework based on the excitation and resonance structures of the instruments. In light of the unified structural framework, a summary of Chapters 2 and 3 will be presented, culminating with a partial list of acoustic features relevant to musical instrument recognition.

3.1 Human recognition abilities

Recognizing musical instruments is a basic component of listening to many kinds of music, and it is considered to be a natural and easy task for many people. For example, Robert Erickson writes:

> "Anyone can recognize familiar instruments, even without conscious thought, and people are able to do it with much less effort than they require for recognizing intervals, harmonies, or scales." (Erickson, 1975, p. 9)

Unfortunately, this common perception is not entirely accurate. In spite of the wide range of research effort in musical acoustics (which will be considered in Section 3.2), very few researchers have tested how reliably people can identify musical instruments. And nearly all of the published research has used rather unnatural testing conditions, asking subjects to identify instruments from single, isolated tones with little or no contextual information. This contrasts starkly with natural listening situations, where melodic phrases consisting of multiple notes are typically heard. Although the studies provide only limited information about natural listening contexts, several general results have been suggested.

It is easier to identify the source of an isolated tone when the attack transient—the tone's onset—is present. According to Kendall (1986), Stumpf noted this as early as 1910 (Stumpf, 1926). This result has been confirmed many times (e.g., Eagleson & Eagleson, 1947; Berger, 1964; Saldanha & Corso, 1964; Thayer, 1972; Volodin, 1972; Elliott, 1975; Dillon, 1981) but has been rejected by Kendall (1986), who did not find such an effect.

Some instruments are more easily identified than others, although different studies have revealed different orderings, and the results appear to be strongly dependent on the context provided by the experiment. In a study with tones from nine instruments (violin, alto horn, trumpet, piccolo, flute, clarinet, saxophone, bells, and cymbals) playing isolated tones at middle-C (approximately 261 Hz), violin, trumpet, and bells were easiest to identify, and alto horn, piccolo, and flute were most difficult (Eagleson & Eagleson, 1947). Saldanha and Corso (1964) tested 20 trained musicians with isolated tones from ten instruments (clarinet, oboe, flute, alto saxophone, French horn, trumpet, trombone, violin, cello, and bassoon) at three pitches (C4, F4, and A4; approximately 261 Hz, 349 Hz, and 440 Hz respectively). They found that the clarinet was easiest to identify (84% correct identifications), followed by oboe (75%) and flute (61%). Violin (19%), cello (9%), and bassoon (9%) were the most difficult. Berger (1964) tested university band performers with tones from ten instruments (flute, oboe, clarinet, tenor saxophone, alto saxophone, cornet, trumpet, French horn, trombone, and baritone) playing at 349 Hz (F4). He found that the oboe was easiest to identify and that the flute and trumpet were the most difficult.

Several authors noticed particular patterns in the mistakes made by subjects. Saldanha and Corso (1964) found that subjects commonly confused bassoon with saxophone; oboe with English horn; trumpet with cornet, saxophone, and English horn; and trombone with French horn, saxophone, and trumpet. Berger (1964) noted confusions between alto and tenor saxophone; cornet and trumpet; and French horn, baritone, and trombone. A series of experiments in Melville Clark's laboratory at MIT provided compelling evidence that the most common confusions occur between instruments in the same *family*, and often in tight family subgroups. For example, Robertson (1961) found evidence for a coherent brass family and sub-families for violin and viola (strings); cello and double bass (strings); and oboe and English horn (double reeds). Schlossberg (1960) additionally found sub-families for trombone and trumpet (brass); and French horn and trombone (brass). Milner (1963) found that musicians make fewer between-family confusions than do non-musicians.

Most studies have found that some people are much better than others at identifying musical instruments. As just stated, Milner (1963) found that musicians make fewer between-family confusions than do non-musicians. Kendall (1986) found that university music-majors performed better than non-majors. However, the "superiority" of trained musicians is not absolute. Eagleson and Eagleson (1947) found that musicians did not perform statistically better than non-musicians in their experiment. Indeed, their best-performing subject had never played a musical instrument. However, instrument identification is a skill that must be developed. In agreement with this view, Saldanha and Corso (1964) noted that their subjects performed significantly better with practice at their identification task.

Several other results, with only limited supporting evidence, are also of interest. Saldanha and Corso (1964) found that identification performance depends on the pitch of the isolated tone in question; their subjects performed better at F4 (349 Hz) than at C4 (261 Hz) or A4 (440 Hz). The presence of vibrato (roughly, sinusoidal pitch modulation with a frequency near 6 Hz and a depth on the order of 1%) makes identification easier (Robertson, 1961; Saldanha & Corso, 1964). Several authors have suggested that note-to-note transitions may be important cues for identification (e.g., Milner, 1963; Saldanha & Corso, 1964). According to Kendall (1986, p. 189):

> "Campbell and Heller (1979; 1978) identified a third category of transient, the legato transient, existent between two sounding tones. Using six instruments playing a major third (F to A), they found that signals containing transients allowed more accurate identification of instrument type than those without, except for 20-msec attack transients."

Actual performance levels vary a great deal between studies. Eagleson and Eagleson (1947) report correct-identification percentages between 35-57% on a free-response task. As mentioned above, Saldanha and Corso's (1964) results depended strongly on the instrument tested, from 9% for cello and bassoon (near chance on their 10-way forced-choice task) to 84% for clarinet. Strong's (1967) subjects correctly identified 85% of the test samples on an 8-way forced-choice task (94% when within-family confusions were tolerated). Berger's (1964) subjects correctly identified 59% of the test samples (88%, tolerating within-family confusions) on a 10-way forced-choice task. Kendall's (1986) subjects, on a 3-way forced-choice task, correctly recognized 84% of the test samples.

In a groundbreaking study, Kendall (1986) questioned the applicability of these isolated-tone studies to realistic listening situations. Because isolated tones are such unusual, unnatural sounds, experiments using them do not necessarily lead to any useful conclusions about sound-source recognition. To test his ideas, Kendall compared his subjects' ability to recognize musical-instrument sounds in several situations, ranging from rather unnatural isolated tones with truncated onsets and offsets to phrases recorded from performances of folk songs (intended to represent "natural" musical signals). Intermediate conditions tested recognition of phrases with attack- and note-to-note transients removed, and with steady-state components removed (leaving *only* the transients).

The results showed that transients are neither sufficient nor necessary for recognizing instruments from musical phrases. In contrast, the "steady-state" is both necessary and sufficient for recognizing trumpet and violin from phrases, and sufficient but not necessary for recognizing clarinet from phrases. In isolated-tone conditions, "transient-only" stimuli were equally recognizable as "normal" and "steady-state only" stimuli. Kendall's subjects performed significantly better in whole-phrase contexts than with isolated tones. Music majors correctly categorized 95% of the phrase stimuli (non-majors scored 74%). On unaltered isolated tones, music majors scored 58% (non-majors scored 50%).

My interpretation of Kendall's results is cautious. His test recordings included examples from only three instruments (clarinet, violin, and trumpet), each from a different family, and his experiments used a 3-way forced-choice paradigm. It is clear, however, that instrument identification is easier in whole-phrase contexts than with isolated tones, and it is likely that transients, both in the attack and in note-to-note transitions, convey less information than the quasi-steady-state in whole-phrase contexts.

Two recent studies are worthy of mention. Crummer (1994) measured event-related potentials (a gross electrical measurement of brain activity) in subjects performing a musical recognition task. His results demonstrate that expert musicians perform such tasks with less effort than do non-musicians. This highlights the importance of learning in sound-source recognition. In a series of recent experiments, Sandell and his colleagues (e.g., Sandell & Chronopoulos, 1996; 1997) have demonstrated that listeners learn to distinguish similar musical instruments (for example, oboe and English horn) better when trained with multiple notes—at different pitches—than when trained with one note at a time. When trained with notes from a limited pitch range, listeners trained in multiple-note contexts generalize better to new, out-of-register, notes than do listeners trained with single tones.

3.2 Musical instrument sound: acoustics and perception

Over the last century-and-a-half, the sounds produced by Western musical instruments, and their perception by human listeners, have been studied in great depth, beginning with the pioneering work of Helmholtz and Seebeck and leading to the latest issue of the *Music Perception*. Musical instrument sounds have been studied from three complementary perspectives: through musical acoustics, through psychophysical experimentation, and through *analysis-by-synthesis*. There are no clear-cut boundaries between these perspectives—researchers often work in more than one area—so the following discussion draws liberally from all three.

Readers interested in more material on these subjects are in luck—there are thousands of relevant journal articles and books. Books by Fletcher and Rossing (1998) and Benade (1990) summarize the acoustics of musical instruments rather well, and often in great depth. Classic texts on the human perception of musical sound include those by Helmholtz (1954) and Plomp (1976); a book and chapter (1995) by Handel bring the early work up to date. Publications in analysis-by-synthesis are more scattered. Risset and Wessel (1982) is a classic. Road's tome, *The Computer Music Tutorial* (Roads, 1996), is an extensive annotated bibliographical history of musical synthesis techniques.

3.2.1 An aside on "timbre"

Much of the psychophysical research on musical sound falls under the rubric "timbre." Timbre is a nebulous word for a perceptual quality (as opposed to a physical quantity) in addition to loudness, pitch, and duration. Debate over the term continues even today (e.g., Houtsma, 1997), though the closest thing to an accepted definition has not changed in decades:

> "[Timbre is] that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar….[T]imbre depends primarily upon the spectrum of the stimulus, but it also depends upon the waveform, the sound pressure, the frequency location of the spectrum, and the temporal characteristics of the stimulus" (American Standards Association, 1960, p. 45)

Unfortunately the word has no useful scientific meaning. It is, as Bregman (1990) notes, a wastebasket category—a holistic word, analogous to *appearance* in vision. It means different things to different people in different contexts, and it encompasses many different features and qualities—indeed, as early as 1890, Seebeck listed at least 20 semantic scales relevant to it (Plomp, 1976), and Helmholtz's translator, A. J. Ellis, hated the way the word had come to be used. He wrote:

> "*Timbre*, properly a kettledrum, then a helmet, then the coat of arms surmounted with a helmet, then the official stamp bearing that coat of arms (now used in France for a postage label), and then the mark which declared a thing to be what it pretends to be, Burns's 'guinea's stamp,' is a foreign word, often odiously mispronounced, and not worth preserving." (Helmholtz, 1954, p. 24)

Although the word timbre appears in the abstract of this dissertation, in the previous two paragraphs, and briefly in the conclusions of Chapter 7, it is not used anywhere else in this dissertation. It is empty of scientific meaning, and should be expunged from the vocabulary of hearing science.

3.2.2 The magnitude spectrum

The modern history of musical-sound research begins in the early 19th century with Fourier's theorem (Fourier, 1822), which proved—among other things—that any periodic signal can be expressed as a sum of sinusoids whose frequencies are integer multiples of a fundamental (whose frequency is the inverse of the signal's period). Ohm, better known for his contributions to the theory of electricity, observed that the human ear performs a kind of frequency analysis and concluded that it analyzes sound waves in terms of sinusoids—a Fourier spectrum (Helmholtz, 1954). Helmholtz, the great German scientist (and an endless source of quotations for hearing researchers), expressed *Ohm's law* as an analysis of sound "into a sum of simple pendular vibrations" (Helmholtz, 1954, p. 33). He proposed a high-level sound taxonomy, dividing sounds into "noises" and "musical tones" (which were defined to be periodic). According to his theory, musical tones are perceived in terms of the magnitudes of their Fourier spectrum components—as opposed to their phases, which he believed to be irrelevant:

> "The quality of the musical portion of a compound tone depends solely on the number and relative strengths of its partial simple tones, and in no respect to their differences of phase." (Helmholtz, 1954, p. 126)

Since Helmholtz, there has been a figurative tug-of-war between proponents of his "spectral theory" of musical sound and researchers who recognized the importance of sound's temporal properties. *Analysis-by-synthesis* research, by trying to discover methods for synthesizing realistic sounds, has revealed several critical limitations of purely spectral theories. Clark demonstrated that recordings played in reverse—which have the same magnitude spectra as their normal counterparts—make sound-source identification very difficult. Synthesis based on Fourier spectra, with no account of phase, does not produce realistic sounds, in part because the onset properties of the sound are not captured (Clark et al., 1963). Although most musical instruments produce spectra that are nearly harmonic—that is, the frequencies of their components (measured in small time windows) are accurately modeled by integer multiples of a fundamental—deviations from strict harmonicity are critical to the sounds produced by some instruments. For example, components of piano tones below middle-C (261 Hz) must be inharmonic to sound piano-like (Fletcher et al., 1962). In fact, all freely vibrating strings (e.g., plucked, struck, or released from bowing) and bells produce inharmonic spectra, and inharmonicity is important to the attack of many instrument sounds (Freedman, 1967; Grey & Moorer, 1977). Without erratic frequency behavior during a note's attack, synthesized pianos sound as if they have hammers made of putty (Moorer & Grey, 1977).

So Helmholtz's theory is correct as far as it goes: the relative phases of the components of a purely periodic sound matter little to perception. However, as soon as musical tone varies over time—for example, by turning on or off—temporal properties become relevant. In the real world, there are no purely periodic sounds, and an instrument's magnitude spectrum is but one of its facets.

A further amendment to Helmholtz's theory is that not all frequency components of complex sounds are created equal. The mammalian ear is constructed in such a way that, for quasi-periodic sounds, only the components with the lowest frequencies—up to about 6 or 7 times the fundamental frequency—are represented separately by the auditory periphery (Plomp, 1976). Components with higher frequencies are represented in tandem with neighboring components. It has been demonstrated that these high-frequency components are perceived as groups—with group rather than individual properties (Charbonneau, 1981). In addition, some aspects of the magnitude spectrum of a quasi-periodic sound may be more salient than are others. The spectral centroid, for example, appears to be more salient than the high-frequency roll-off rate and the overall smoothness of the spectral shape, at least by dint of the number of studies that have revealed it to be important. And the non-periodic, noisy, portions of the sound may also be perceptually salient, though they have not been studied in nearly as much depth (however, see Serra (1989) for an influential attempt to model them for synthesis purposes).

3.2.3 The dimensions of sound

A great deal of research effort has been devoted to revealing the underlying perceptual dimensions of sound. The primary dimensions—pitch, loudness, and duration—are relatively obvious, but their perceptual complexity is not. Additional dimensions are less obvious, and the tacit assumption that it even makes sense to talk about perceptual "dimensions," as if they could be varied independently, is questionable if not outright incorrect.

Pitch is an essential property of many kinds of musical sound and a salient perceptual attribute of many non-musical sounds, including talking human voices and animal vocalizations. It is defined by the American National Standards Institute (ANSI, 1973, as cited by Houtsma, 1997, p. 105) as "that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from high to low". The pitch of a sound can be defined operationally as the frequency of the sinusoid that it "best matches." As a scale, pitch is monotonically related to scale of sinusoid frequencies. This aspect of pitch is related to the *periodicity* of the sound waveform, and in this limited sense, the *pitch frequency* is just the inverse of the waveform's repetition period.

Pitch is not, however, a unidimensional scale (Shepard, 1982). There are at least three *pitch attributes* that complicate a simple definition by periodicity. First, not all pitches with the same fundamental period are equivalent; sounds may differ in degree of "pitchiness," from harmonic complexes that evoke a strong, rather unambiguous pitch sensation to bands of noise whose pitch strength varies inversely with bandwidth. A second kind of variation, termed *pitch height, sharpness,* or as I will refer to it, *brightness*, is related to the spectral content—a periodic sound becomes *brighter* as its high-frequency partials become stronger relative to its low-frequency partials—rather than the fundamental period. A third aspect, *pitch chroma,* complicates matters further. Traditional Western music divides the octave (a doubling in pitch period) into twelve logarithmically spaced steps, which make up the chromatic scale. Pitch periods related by a power-of-two ratio have the same chroma and are functionally equivalent (the musicological term is *octave equivalence*) in many musical settings.

Among these aspects, pitch period and chroma are most important for music-theoretical purposes such as defining melodic fragments. Brightness and its relation to pitch period are crucial for sound-source identification because they encode information about the physical sound source. The pitch period encodes the vibration frequency of the source excitation, and brightness is affected both by the frequency content of the source excitation (its *harmonic richness*) and by the resonant properties of the vibrating body, which may enhance or weaken various portions of the spectrum.

The pitch of real-world sounds is not static; it varies over time, either in the relatively discrete steps between pitch chroma or continuously, with vibrato (periodic modulation) or jitter (random modulation).

Another primary perceptual dimension of sound is *loudness*, defined by ANSI as "that intensive attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from soft to loud" (quoted by Houtsma, 1997, p. 105). Although loudness is not as complex as pitch, it is by no means simple, and many models have been proposed for estimating the loudness of a sound based on its waveform. The loudness of a sound source depends on the acoustic energy it produces at the position of the listener, on the duration of the sound (for relatively short sounds, loudness increases with duration), and on the frequency content of the sound's spectrum (Moore, 1989). A simple but effective first-order model relates loudness to the sum of the energy in the frequency regions termed *critical bands* (Moore, 1989).

The third primary dimension of sound, duration, has not been studied as extensively as pitch and loudness. Humans are better at comparing the durations of short sounds (on the order of 0.5-10 seconds) than of longer sounds. Although sound duration may play a role in sound-source recognition, to my knowledge such an influence has not been explored experimentally, except to note how much of a signal is required for recognition of various qualities. For example, humans require 2-3 cycles of a periodic sound to identify its octave, and several more to recognize its pitch chroma (Robinson & Patterson, 1995).

Researchers have long been interested in identifying perceptual dimensions of sound in addition to pitch, loudness, and duration. Multidimensional scaling (MDS) is a method for finding underlying perceptual/conceptual dimensions of a collection of stimuli, if such structure exists. MDS techniques have been extensively applied to the perception of isolated musical times (a partial list includes: Plomp et al., 1967; Plomp, 1970; Wedin & Goude, 1972; Grey, 1975; 1977; 1978; Gordon & Grey, 1978; Grey & Gordon, 1978; Wessel, 1983; Krumhansl, 1989; Kendall & Carterette, 1991; Krumhansl & Iverson, 1992; McAdams & Cunible, 1992; Iverson & Krumhansl, 1993; Hajda et al., 1994; Kendall et al., 1994; McAdams et al., 1995). A recent review chapter (Hajda et al., 1997) provides an excellent critical overview of these and other related investigations.

A typical musical MDS study begins with a collection of 8-25 isolated tones, with differences of pitch, loudness, and duration minimized. Subjects are asked to rate either the similarity or dissimilarity of each pair of tones. These judgments are collected, and a computer program finds a low-dimensional arrangement of the stimuli (each stimulus occupies a point in the space) that best accommodates the dissimilarity ratings (viewing dissimilarity as analogous to distance in the space). If the set of stimuli has an underlying dimensional structure, the dimensions of the arrangement uncovered by MDS can often be interpreted in terms of acoustic/perceptual/conceptual attributes or categories.

Hajda et al. (1997) interpret the results of the various musical MDS studies as highly inconsistent. The “space” recovered by the MDS algorithms depends strongly on the particular set of stimuli used, which implies that the subjects’ criteria for similarity are context-dependent. Only the average spectral centroid, which correlates strongly with subjects’ ratings of brightness, is consistently found to be a principal dimension. Other dimensions have been interpreted as related to the attack rise-time, spectral irregularity, and instrument family.

Even if studies of the similarity of pairs of tones led to a consistent set of dimensions that could be interpreted in terms of simple acoustic/perceptual properties, the “space” implied by such dimensions would be of questionable relevance to sound-source recognition. The assumption that sounds occupy positions in a perceptual space with a uniform distance metric has not been justified, and the interpretations of MDS results often beg the question.

Rather than seek a set of dimensions to describe sounds, my approach is to find a set of perceptually relevant acoustic attributes that yield information about source identity. In particular, these attributes are indicative of the production invariants of the source, and it is these invariants that underly sound-source recognition. Such attributes may be continuous- or discrete-valued, and there is no reason to expect that any two attributes will be independent, statistically or otherwise.

As mentioned above, brightness—as estimated by the spectral centroid—is consistently found to be a salient sound attribute, one that strongly mediates the perceived similarity between pairs of sounds. Beauchamp found that many musical instruments exhibit a nearly monotonic relationship between intensity (indicating loudness) and spectral centroid (Beauchamp, 1982; Beauchamp, 1993). In most cases, louder sounds have a higher concentration of high-frequency energy and are thereby brighter. Beauchamp suggests that matching the intensity and spectral centroid of a synthesized sound—as a function of time—to a recorded original sound, goes a long way toward creating a convincing resynthesis (i.e., one that is judged by listeners to be similar to the original).

3.2.4 Resonances

In Section 2.3, I stated that the geometry and material properties of a source’s vibrating body impose constraints on the acoustic waveform produced by the source. The vibrating body can be viewed as a *resonator* coupled to the source’s means of excitation. In this section, a simple physical resonator will be considered, and some of its properties will be developed (the discussion is adapted from the presentation of Fletcher and Rossing (1998)). The intuitions gained by this exercise are necessary to appreciate the discussion of the (often more complex) resonant properties of orchestral instruments presented in the next section.

Consider a mass, *M*, connected by an ideal spring and damper to a fixed surface, as shown in Figure 5. The forces acting on the mass, arising from the restoring force of the spring (*F* = -*Kx*, where *K* is the spring constant and *x* is the mass's position), from the damper (*F* = -*Rv*, where *R* is the damping constant and *v* is the mass's velocity), and from an external force *f(t)*, impose an acceleration on the mass (from Newton's second law of motion, *F* = *Ma*, where *F* is the force and *a* is the mass's acceleration). In combination they yield the equation

$$M\frac{d^2x}{dt^2} + R\frac{dx}{dt} + Kx = f(t)\ . \qquad \textbf{(1)}$$

Substituting

$$\alpha = \frac{R}{2M} \text{ and } \omega_0 = \sqrt{\frac{K}{M}} \qquad \textbf{(2)}$$

we have

$$\frac{d^2x}{dt^2} + 2\alpha\frac{dx}{dt} + \omega_0^2 x = \frac{f(t)}{M}\ . \qquad \textbf{(3)}$$

If there is no external force (i.e., *f(t)* = 0), the equation has solutions of the form

$$Ae^{-\alpha t}\cos(\omega_d t + \phi) \qquad \textbf{(4)}$$

where

$$\omega_d = \sqrt{\omega_0^2 - \alpha^2} \qquad \textbf{(5)}$$

is the natural, or free-vibration, frequency of the system.

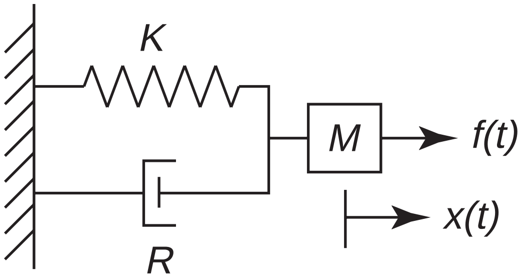

FIGURE 5. A simple physical resonator, consisting of a mass *M* attached to a fixed surface by a spring (spring constant *K*) and a damper (damping coefficient *R*). An external force, *f(t)*, acts on the mass, whose position is notated *x(t)*. The resonator's properties are discussed in the text.

When an external driving force of frequency ω is applied, the steady-state response of the system (which is linear) will be at the same frequency, so we can replace *x(t)* in Equation 3 with *A*exp(*j*ω*t*). Taking the appropriate derivatives and rearranging slightly, we have

$$Ae^{j\omega t}(-\omega^2 + j\omega 2\alpha + \omega_0^2) = \frac{Fe^{j\omega t}}{M} . \qquad \textbf{(6)}$$

This equation has a solution given by

$$x(t) = Ae^{j\omega t} = \frac{Fe^{j\omega t}/M}{\omega_0^2 - \omega^2 + j\omega 2\alpha} . \qquad \textbf{(7)}$$

Thus, the amplitude of vibration depends of the driving frequency (ω), the natural frequency of the undamped system (ω_o) and the damping (α).

Figure 6 illustrates the frequency response of the system for various values of α. Defining the value *Q* as the ratio of the system's natural frequency to the –3 dB bandwidth of the frequency response (or, equivalently, $\omega_o/2\alpha$), we see that as the damping decreases, the frequency response narrows, increasing the *Q* of the resonator.

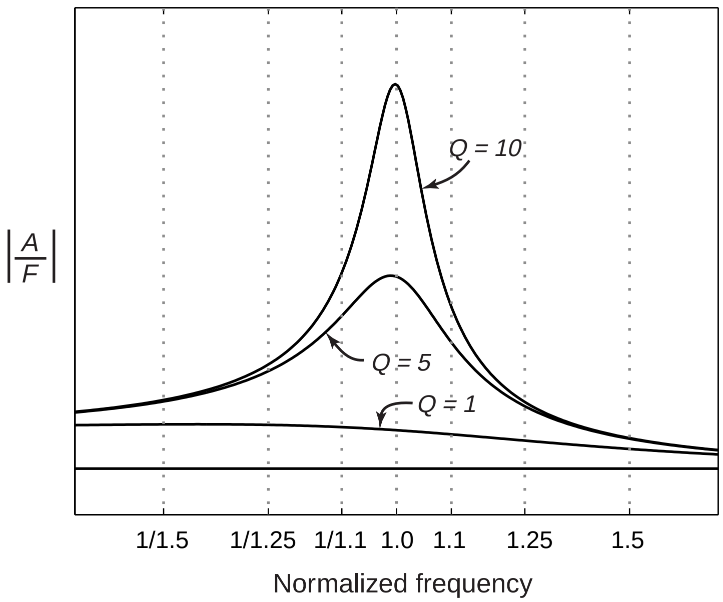

FIGURE 6. The effect of damping (hence, *Q*) on the transfer function of the resonator. As the damping decreases, the *Q* increases, and the frequency response narrows.

Damping also plays an important role in the time-evolution of the resonator's response to real-world external forces, which are not always easy to express as sums of infinite-extent sinusoids. The response to a transient can be characterized as a *ringing* at the system's natural frequency, which decays at a rate that depends on the *Q* of the resonator (in fact, *Q* can be equivalently defined, for the simple example used here, as $\omega_o\tau/2$, where τ is the time required for the impulse-response of the resonator to decay by a factor of $1/e$). The response to a gated sinusoid (e.g., turned on at $t = 0$) is a combination of the transient response and the steady-state response, which may *beat* against each other, causing various degrees of apparent complexity as the driving frequency varies. When the driving frequency is exactly equal to the resonator's natural frequency, the system's output will grow from zero, approaching the steady-state amplitude with the time constant τ used to define the *Q* of the system. Figure 7 illustrates the response of the simple resonator to gated sinuoids of different frequencies, for different *Q* values.

The behavior of real-world resonant systems is generally more complicated than that of the simple oscillator presented above, but the intuitions developed by its consideration are useful for understanding more complicated systems. In the next section, the effects of resonances on the sounds of orchestral instruments will be considered, on a family-by-family basis.

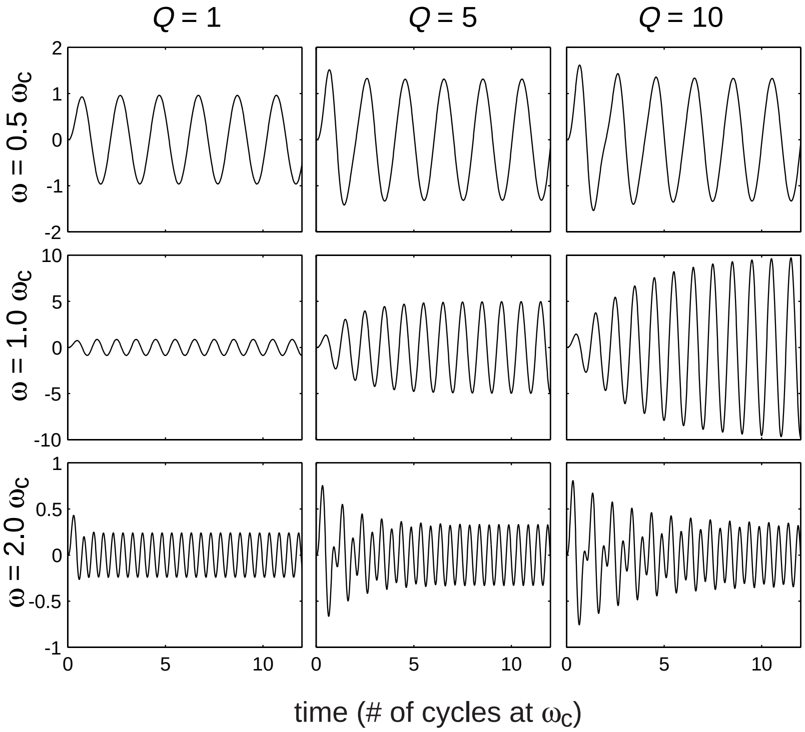

FIGURE 7. The response of the simple resonator to gated sinusoids at three different frequencies (relative to the resonant frequency), and for three different resonator *Q* values.

3.3 Instrument families

The non-percussive orchestral instruments are commonly divided into three *families*: the brass, the strings, and the woodwinds. Although this division is largely due to the historical development of the instruments (e.g., flutes, now made of metal, were originally made of wood and are still considered members of the woodwind family), commonly confused instrument pairs (e.g., violin and viola; oboe and English horn; trombone and French horn) nearly always occur within a particular family (see Section 3.1). It is possible to use instrument geometry, materials of construction, and playing method to construct a single taxonomy of musical instruments (a good example is given by von Hornbostel & Sachs, 1961), and commonly-confused instruments will usually occupy neighboring taxonomic positions. In this section, the traditional families will be considered in turn. Within each family, the acoustic and perceptual properties of the perceptual “confusion groups” will be presented.

3.3.1 The brass instruments

Of the three broad families, the brass family has the simplest acoustic structure. The family includes the cornet, trumpet, fluegel horn, trombone, French horn, baritone, euphonium, and tuba. Each instrument consists, in its barest essence, of a long hard-walled tube (often made of brass) with a flaring bell at one end.

The player makes sound by blowing into a mouthpiece affixed in the narrow end of the tube. The player's tensed lips allow puffs of air into the tube, which travel the tube's length and partly reflect off the impedance mismatch caused by the bell. This reflection allows standing waves to build at near-integer multiples of a frequency corresponding to the speed of sound divided by twice the tube's length. The modes do not occur exactly at integer multiples because the bell reflects low frequencies sooner than high, making the effective length of the tube frequency-dependent (Benade, 1990). The player can vary the pitch by changing his lip tension, which changes the particular vibration mode that is excited (trumpet players commonly excite one of the first eight modes; French horn players can excite modes as high as the 16th), or by changing the length of the tube (either by depressing valves or moving a slide) (Roederer, 1973).

The instrument provides feedback to the player in the form of the bell reflection arriving back at the mouthpiece, but it can take several round trips for the standing waves to build up (Benade, 1990; Rossing, 1990). During this time—which depends on the tube length, not the pitch—the instrument is not stable, and for the high modes, many pitch periods can elapse before a stable oscillation is set up. This effect can cause the instrument's pitch to wander during the attack; the pitch has been observed to scoop up from below and to oscillate around the target value (Luce, 1963; Risset, 1966). The very best players minimize this effect through extremely precise control of lip tension. Instability at onset may also be the cause of "blips"—a term used by Luce (1963) to describe small, inharmonic bursts of energy—preceding the tonal part of a note. Luce observed blips in tones produced by all of the brass instruments, most commonly at their lowest pitches.

The internal spectrum of a brass instrument varies with the air pressure at the player's lips. At very low amplitudes, the pressure wave is nearly sinusoidal, but at increasing amplitudes, it becomes more pulse-like. Figure 8 shows, in schematic form, how the spectrum broadens with increasing pressure. The standing-wave modes are indicated by filled circles, and they are connected by lines for clarity. In the steady-state, the puffs of air are injected periodically, so the internal spectrum is harmonic even though the resonance modes of the tube are not.

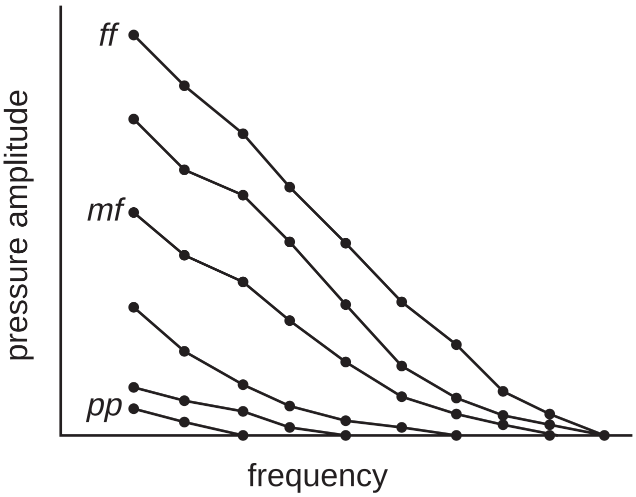

FIGURE 8. The internal spectrum of a brass instrument, for a range of air-pressure levels, after Benade (1990).

The bell reflects low-frequency energy more effectively than high, and this has three important effects. During the onset of a note, the low-frequency modes build up more rapidly than the high-frequency modes. This explains Luce's (1963) observation that the onsets of the partials are skewed, with low frequency partials building up energy quickly, in close synchrony, and high-frequency partials entering later. The second effect of the bell reflection is that the instrument's external spectrum—what is actually heard by a listener—is a high-pass version of the internal spectrum. The transformation function is sketched in Figure 9, and the resulting external spectrum is shown in Figure 10 (again, with the harmonic modes indicated by filled circles). The final effect is that, because the bell's radiation pattern is more directional at high frequencies, the actual projected spectrum varies with the angle between the bell's axis and the listener's position. The general result, however, is a single broad resonance, whose center frequency is more-or-less fixed by the bell's lowpass cutoff frequency.

As described above, the instruments of the brass family have much in common. The differences are primarily of scale: the large instruments have lower cutoff frequencies and pitch ranges. Measured values for the center-frequency, low- and high-frequency rolloff slopes (from Strong & Clark, 1967), and approximate onset times (from Luce, 1963), for four particular brass instruments are shown in Table 1. These values may vary from instrument to instrument and from player to player, but are representative according the authors.

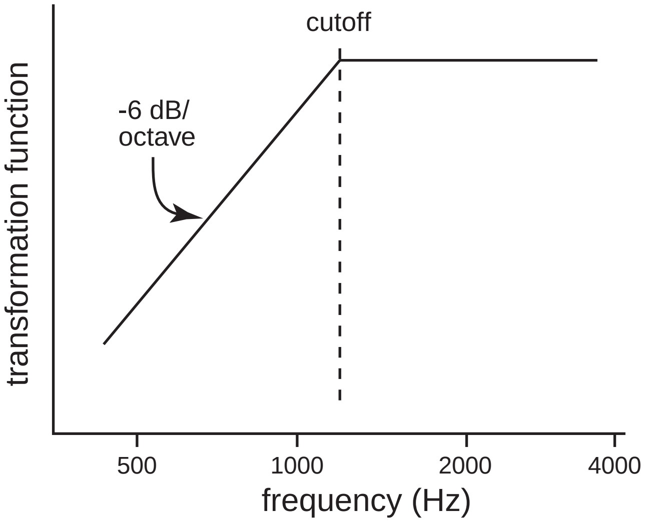

FIGURE 9. Schematic of the bell transformation function for a trumpet, after Benade (1990).

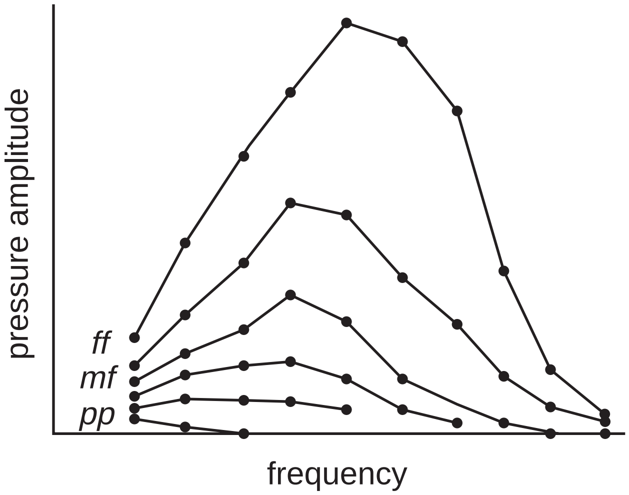

FIGURE 10. The external spectrum of a brass instrument, for a range of air-pressure levels, after Benade (1990).

Instrument	cutoff frequency (Hz)	low-frequency rolloff (dB/octave)	high-frequency rolloff (dB/octave)	Amplitude onset (ms)	Waveform onset (ms)
Trumpet	1150	6	10-20	100	25
French horn	500	10	20	40	30
Trombone	475	5	8-18	50	35
Tuba	275	?	10-20	75	90

TABLE 1. Characteristics of several brass instruments. Spectral data are from Strong & Clark (1967); onset data are from Luce (1963).

A final complication in the analysis of brass instruments is that some are commonly played with devices called *mutes* inserted into the bell. Several varieties of mutes are used with the cornet, trumpet, fluegel horn, and trombone. Each introduces a set of acoustic resonances and anti-resonances, generally above 1 kHz, which give the instrument's tone unique qualities (Fletcher & Rossing, 1998). In addition, French horn players often insert a hand into the bell to mute high-frequency components (Rossing, 1990).

3.3.2 The string instruments

The common bowed-string instruments, in order of increasing size, are the violin, viola, cello, and double bass. Each string instrument consists of an ornate wooden body with an extended neck. The strings (usually numbering four) are stretched along the neck, over a fingerboard, attached at one end to the body (by way of the *bridge*), and at the other to tuning pegs (which control the string tension). When the strings vibrate, coupling through the bridge causes the body—and the air mass contained within—to vibrate, which in turn projects sound into the air. The performer sets a string in motion by plucking it or by dragging a bow (usually consisting of stretched horse hair on a wooden frame) across it.

When bowed, the string "sticks" to the bow for brief periods, moving in synchrony with the bow's motion and then suddenly snapping back. This causes the string's motion to resemble a sawtooth pattern (Benade, 1990; Mathews et al., 1966). In the steady-state, the waveform is approximately periodic (the period depends on the length between the bridge and the player's finger on the fingerboard, along with the tension and mass of the string) and thus has a harmonic spectrum. The exact shape of the waveform—hence the frequency content of the spectrum—depends on the pressure of the bow against the string and on the bow's position relative to the bridge (bowing nearer the bridge or with increased pressure increases the proportion of high frequencies in the spectrum, making the sound brighter). To a first approximation, the strength of the *n*th partial relative to the first is $1/n$ (Benade, 1990; Rossing, 1990). There may, however, be partials with near-zero strength if the bow position mutes them.

It is, however, somewhat misleading to speak of a steady-state for a bowed string. The complexity of the interaction between the bow and string causes the length of each "sawtooth" to vary from cycle to cycle, creating a great deal of frequency *jitter* (Benade, 1990), which is coherent among the various partials (Brown, 1996). The attack and release of a bowed tone are particularly complex. The bow may scrape the string during the attack, creating substantial noise, and the spectrum is generally not quite harmonic (Beauchamp, 1974; Luce, 1963). For example, the low partials start very sharp when the string is excited vigorously (Benade, 1990).

The spectrum of a plucked string is never harmonic. Because of dispersion in the string (that is, waves of different frequencies travel at different speeds along the string), the high-frequency partials are somewhat sharp relative to the low-frequency partials (Fletcher, 1964; Roederer, 1973). As in the case of bowing, the spectrum of the plucked string depends on the plucking position (Roederer, 1973); the spectrum will be brighter for positions nearer the bridge, and some partials may be muted, having near-zero strengths.

The bridge is the main connection between the vibrating string (which does not move enough air by itself to be audible in the context of an orchestra) and the instrument's body (which does). The bridge introduces broad resonances to the instrument's spectrum; for the violin these occur near 3 kHz and 6 kHz (Rossing, 1990). Players sometimes attach a *mute* to the bridge, which increases the bridge's effective mass and lowers the resonance frequencies, creating a somewhat darker tone.

A string instrument's body—with its ornate geometry—has many different modes of vibration, both of the air inside and of the body's wood plates. These vibration modes introduce a large number of narrow (high Q) resonances, at different frequencies, between the vibration spectrum of the strings and that of the air around the instrument. The low-frequency resonances (e.g., the first "air" and "wood" resonances) are tuned carefully in high-quality instruments, but details of the high-frequency resonances vary tremendously from instrument to instrument (and even change over time as the instrument is played and the wood ages or is strained (Hutchins, 1998)). Analysis-by-synthesis research (e.g., Mathews et al., 1966; Risset & Wessel, 1982) has demonstrated that convincing bowed-string sounds can be synthesized by passing a $1/n$ spectrum (with some zeroed partials) through a filter with a large number of narrow resonances in roughly the correct frequency regions, without paying attention to the details of resonance placement.

To a first approximation, the complex resonance structure of a string-instrument's body causes the spectrum of any particular note to be less regular than the simple $1/n$ rolloff of the bowed string. With frequency jitter, or the commonly used frequency modulation called *vibrato* (in which the player modulates the effective string length—hence the pitch—by rocking a finger back and forth on the fingerboard), the position of each harmonic partial in relationship to the body resonances changes over time. This interaction creates complex patterns of amplitude modulation (Risset & Wessel, 1982). The amplitude modulation of each partial varies at the same rate as the frequency modulation, but can be in different directions for different partials, depending on their particular relationships to nearby resonances (Fletcher & Sanders, 1967) and different depths (as much as 15 dB, according to Fletcher & Rossing, 1998).

The body resonances also affect the attack and release of each note. The rate of energy buildup or decay of a particular partial is related to the effective *Q* of nearby resonances, and this causes the attack and release rates of the different partials to vary with partial number and pitch (Beauchamp, 1974). The attack rates of isolated string tones are generally much slower than those of the other orchestral instruments. Indeed, it can take a large fraction of a second for a string tone to reach "steady state;" in contrast, brass tones generally reach steady state in less than 100 ms. The overall attack time appears to vary greatly from instrument to instrument, possibly from player to player, and perhaps even from note to note. Some representative values, measured by Luce (1963), are shown in Table 2.

Instrument	Time required to reach steady state (ms)	Time required to reach full amplitude (ms)
Violin	100	200
Viola	40	100
Cello	120	350
Double bass	80	100

TABLE 2. Attack times for the bowed string instruments, as measured by Luce (1963).

The violin is, perhaps, the "king" of the orchestra; it is the most-engineered, most-studied, and most-uniformly-constructed member of the string family. The open strings of a violin are typically tuned in fifths, to the pitches G3, D4, A4, and E5 (196 Hz, 290 Hz, 440 Hz, and 660 Hz), and the first air and wood resonances of a high-quality violin's body are tuned to correspond to the pitches of the open middle strings (approximately 290 Hz and 440 Hz respectively) (Benade, 1990). As stated above, the upper body resonances vary greatly from instrument to instrument, but there is usually a broad maximum near 3 kHz that is due to the bridge resonance. Figure 11 depicts the resonance modes of the violin.

The viola is somewhat larger than the violin, but the change in body size is not in scale with the change in pitch range; the open strings of a viola are tuned a musical fifth below those of a violin (C3, G3, D4, A4, or 130 Hz, 196 Hz, 290 Hz, and 440 Hz), but the first air and wood resonances are relatively more flat (230 Hz and 350 Hz, or D-flat-3 and F4), falling slightly above the frequencies of the lowest two strings (Benade, 1990). Violas are not made as uniformly as violins, so the string-to-resonance relationships vary more (Benade, 1990). The viola's principal bridge resonance is close to 2 kHz, causing the upper body resonances to form a maximum there.

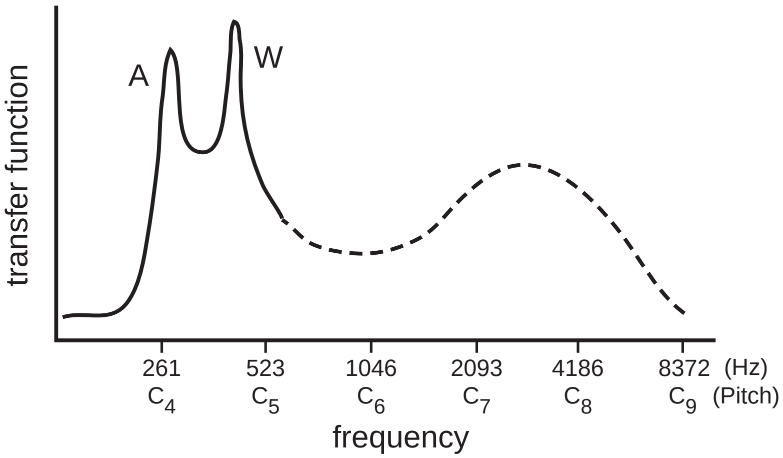

FIGURE 11. The resonance modes of a violin, after Benade (1990). The first air and wood modes (indicated by A and W) are typically tuned to particular frequencies. The upper modes are much more complex and vary greately from instrument to instrument (hence are indicated by a dashed line showing the general trend). The broad maximum near 3 kHz is due to the bridge resonance.

The dimensions of the cello are about twice those of the viola. Its strings are tuned one octave below the viola, to C2, G2, D3, and A3 (65 Hz, 98 Hz, 145 Hz, and 220 Hz), and its first air and wood resonances are typically near 125 Hz and 175 Hz respectively. Benade reports that the cello often exhibits a deep notch in its resonance structure near 1500 Hz.

The dimensions of the double bass are about twice those of the cello. The strings are tuned in fourths, to E1, A1, D2, and G2 (41 Hz, 55 Hz, 73 Hz, and 98 Hz), and some instruments have a fifth string. The first air and wood resonances of the bass occur at approximately 60 Hz and 98 Hz respectively, and the bridge resonance frequency is approximately 400 Hz.

The string instruments form a very tight perceptual family. Several of the experiments reviewed in Section 3.1 demonstrated listeners' inability to reliably distinguish the four instruments by sound alone; each is commonly confused with its neighbors in scale. The violin and viola, because they are closest in pitch and scale, are the most difficult to distinguish. The limited available evidence suggests that listeners are very good at determining whether or not an instrument is a member of the string family, but that once that determination is made, they use relatively unreliable criteria such as pitch range or overall brightness to, in effect, guess the particular instrument. Experienced musicians make use of highly cognitive cues—such as recognizing particular pieces or playing techniques—to make much better decisions when given access to an entire phrase.

3.3.3 The woodwind instruments

The woodwind family is much less homogenous than the brass or strings. It is made up of several distinct subgroups, both acoustically and perceptually: the double-reeds, the single-reed clarinets, the flutes (or "air" reeds), and the remaining single-reeds, the saxophones.

Although the various sub-families have distinct properties, each woodwind instrument has several properties common to the family as a whole. Woodwinds produce sound by creating standing waves in a tube, whose effective length is altered by selectively opening or closing tone-holes. As with the brass instruments, the player can *overblow* to change the pitch, by selecting a set of vibration modes with higher frequencies (Roederer, 1973); in contrast to the brass instruments, woodwinds often have *register keys*, which when depressed open small tone-holes that diminish the strength of the tube's lowest vibration mode, easing register-to-register transitions (Fletcher & Rossing, 1998). The open tone-holes of a woodwind instrument impose a low-pass characteristic on the instrument's spectrum, and the cutoff frequency—which varies surprisingly little across the pitch range of the instrument—is essential to the tone of the particular instrument (it alone can determine whether an instrument is suitable for a soloist or for an ensemble performer). As Benade (1990) writes:

> "[S]pecifying the cutoff frequency for a woodwind instrument is tantamount to describing almost the whole of its musical personality."

Finally, the woodwinds—with the exception of the flutes—tend to have the most rapid attack transients of the three major families. In the rest of this section, the perceptual/acoustic subdivisions of the woodwind family will be considered in turn.

The double-reed subfamily consists of, in order of increasing size, the oboe, English horn, bassoon, and contrabassoon. Each instrument's body consists of a conical tube, and the performer creates sound by forcing air through a sandwich of two reeds, which is attached to the tube at one end. The conical tube supports vibration modes at integer multiples of the frequency corresponding to the tube's effective length, which is altered by opening or closing tone holes. The double-reeds are commonly played with vibrato.

The oboe typically has two resonances—a strong one near 1 kHz and a weaker, more variable one near 3 kHz (Rossing, 1990; Strong, 1963)—separated by an anti-resonance near 2 kHz (Strong, 1963). Luce (1963) measured one oboe, finding that it takes very little time for the attack transient waveform to stabilize in shape (15 ms) and amplitude (20 ms), and noting that the fundamental (the first partial) appears first.

The English horn is, to a first approximation, a larger oboe, but its properties are not as consistent as those of its smaller sibling (Luce, 1963). It typically has a prominent resonance near 600 Hz and a weaker resonance near 1900 Hz (Strong, 1963), separated by an anti-resonance between 1300 Hz (Strong, 1963) and 1600 Hz (Luce, 1963). Above the resonances, the instrument's spectrum rolls off abruptly, at approximately 20 dB per octave (Strong, 1963). Luce's (1963) measurements suggest that the instrument's waveform stabilizes in 30 ms and reaches a stable amplitude in 50 ms during the attack.

The bassoon is much larger than the oboe and English horn. It has a tone-hole cutoff frequency near 375 Hz (Benade, 1990) and a prominent resonance between 440-494 Hz (Rossing, 1990). The bassoon's spectrum rolls off rapidly above the primary resonances, and there may be a minor anti-resonance near 850 Hz (Luce, 1963). Luce's attack measurements suggest waveshape and amplitude stabilization times of 30 ms and 40 ms respectively. As a final note, the bassoon is unique among the members of the orchestra in that the first partial of its tones (the fundamental frequency) is very weak—perhaps because its tube is so long that it must be folded to be playable. The contrabassoon is—to a first approximation—similar to a bassoon whose dimensions are doubled.

The clarinets are a singular sub-class of the orchestral instruments. A clarinet has a single-reed mouthpiece attached to a cylindrical tube that, to a first approximation, supports vibration modes only at *odd* multiples of the fundamental corresponding to *twice* the tube's length. There are several different sized members of the clarinet group; the B-flat and A (tenor) clarinets and the bass clarinet are most commonly used in the orchestra.

The B-flat and A clarinets are nearly identical. Players alternate between them for ease of playing particular musical keys rather than for reasons of tone quality. The clarinet's spectrum is limited by the tone-hole cutoff, which varies from 1200-1600 Hz depending on the instrument (Benade, 1990) and the 5 kHz limitation of reed vibration (Luce, 1963). Two registers separated by a musical twelfth (again because of the cylindrical tube closed at one end—the first two registers of other woodwinds are separated by an octave) cover most of the clarinet's range. The relative strengths of the odd and even partials depend on their frequencies and the playing register. They are shown in schematic in Figure 12. Above the cutoff frequency (approximately 3 kHz for a B-flat clarinet), the odd and even partials are of similar strength; below the cutoff, the odd-numbered partials are stronger (the difference is exaggerated in the upper register). Luce observed waveform and amplitude attack times of 40 ms and 60 ms, and noted that the fundamental partial appears first; the upper partials are delayed by 5-10 cycles and then rise very rapidly (Luce, 1963).

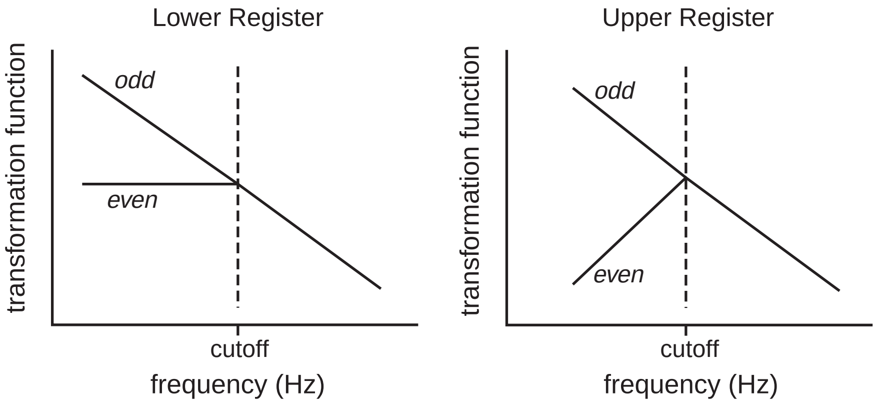

FIGURE 12. Schematic of the clarinet's spectrum, after Strong (1963). Above the cutoff frequency, the odd and even partials behave similarly. Below the cutoff, the even-numbered partials are suppressed relative to the odd-numbered partials. The effect depends on the instrument's playing register.

The flute family, or "air reeds," consist of (in order of increasing size) the piccolo, flute, alto flute, and bass flute. Of these, only the piccolo and flute are commonly used in orchestras. The flute player excites the instrument's tube by projecting a flow of air across a metal edge at one end. The resulting turbulent, noisy signal excites the tube at dips in its acoustic impedance (Benade, 1990). The common flute has an overall resonant maximum near 600 Hz, with a high-frequency rolloff from 10-30 dB per octave (Strong, 1963). It has a very slow, smooth attack (Luce observed rise times in the neighborhood of 160 ms), commonly followed by strong periodic amplitude modulation—called *tremolo*—at frequencies like those used in double-reed or string vibrato. At pitches above 500 Hz, the flute's spectrum is dominated by the fundamental frequency, and above 880 Hz, the waveform is nearly sinusoidal (Luce, 1963).

The piccolo is essentially a very small flute, and it shares many of the flute's properties. Luce measured waveform and amplitude attack times of 25 ms and 100 ms respectively, but observed that the attack gets much longer at high pitches (Luce, 1963). At pitches above 1 kHz, the piccolo's waveform is nearly sinusoidal.

The last sub-class of woodwind instruments is the saxophones, which are used only in modern orchestral music and have been studied in less depth than the other orchestral instruments. There are several different sized saxophones, including the soprano, alto, tenor, and baritone. The saxophone is a single-reed instrument with a conical bore. Rossing (1990) notes that the saxophone spectrum has few high harmonics, and Freedman (1967) observed that inharmonicity is important for the bite of its attack, but further details are hard to come by.

3.4 Summary

This chapter examined musical instrument recognition from two perspectives. First, human instrument-recognition abilities were considered and quantified for the common orchestral instruments. Second, the sounds produced by the instruments of the orchestra were examined in order to discover the features upon which the human recognition process might operate. This section summarizes the relevant findings within a unified framework.

There is a common belief that people can become very good at identifying musical instruments from sound alone, but this conventional wisdom is flawed. The evidence presented in Section 3.1 suggests that people can become very good at recognizing *classes* of instruments with similar excitation and resonance properties. These classes correspond closely to the traditional instrument families, with the exception of the woodwind family, which comprises several distinct subgroups. Distinctions between members of the same class—e.g., violin and viola, oboe and English horn, or trombone and French horn—are made much less reliably.

Based on this evidence, it is plausible that the process of musical instrument recognition in humans is taxonomic—that classification occurs first at a level corresponding to instrument sub-families (perhaps: strings, brass, double-reeds, clarinets, flutes, and saxophones) and progresses to the level of particular instrument classes (e.g., trombone, violin, etc.). Although I have not presented objective proof of this structure, it is highly consistent with the structure of human perception in other domains, as demonstrated by Rosch and her colleagues (Rosch, 1978; Rosch et al., 1976) and summarized in Section 2.3. In the next two chapters, a system based on this taxonomic structure will be described, and its performance will be demonstrated to be similar in many aspects to that of humans.

One of the core theses of this dissertation is that many sound sources—and, in particular, the orchestral instruments—are identified through recognition of their resonant properties. The construction of musical instruments and the physics of sound production place strong constraints on musical sounds—constraints that measurably and perceptually affect the acoustic signal. This viewpoint illuminates many of the experimental results. For example, in an isolated-tone context, the attack transient may be a more salient cue for identification than the steady-state spectrum precisely because the rise-times of the various partials reveal more about the resonance structure (in particular, the effective Q of resonances in different frequency regions) than do their asymptotic values. If, however, the steady-state portion is performed with vibrato, the amplitude modulations of the partials (induced by the frequency modulation as they interact with the resonances of the vibrating body) reveal the resonant structure, and human recognition performance improves.

There are many properties of the waveforms produced by musical instruments that reveal information about the excitation and resonance structure of the instruments. As suggested above, different properties are salient in different contexts. To date, machine systems have not taken advantage of this; as described in Section 2.4, nearly all “instrument recognition” systems have operated on isolated tones (and, crucially, have not demonstrated any kind of performer-independent generalization). The most intriguing systems are those that operate on musical phrases rather than isolated tones. Such systems have had good success at distinguishing among a small number of instrument categories by using cepstral coefficients calculated on small time windows. The cepstral data are used in such a way that they capture information about the short-term spectral shape of the sound wave, while discarding information about its variation over time. Many of the cues known to be important for humans are not represented, including pitch, vibrato, FM induced AM, and the rise times of the harmonic partials.

The sound of a musical instrument is often thought of as multidimensional. Although there are several sound properties that apply to many sounds (e.g., pitch, loudness, brightness), there is no evidence that there is a simple, multidimensional space underlying perception or recognition. In contrast, the myriad cues used by listeners vary from source to source and are better described as collections of features—some discrete, some continuous.

The perceptually salient features of sounds produced by orchestral instruments include:

- **Pitch**: The periodicity pitch of a sound yields information about the size of the sound source. Typically, smaller sources produce higher-pitched sound; larger sources produce lower pitches. Variations in pitch are also sources of information. The degree of random variation reveals information about the stability of the source excitation and the strength of its coupling to the resonant body. For example, brass instruments, which have relatively weak excitation-resonance coupling, exhibit wide pitch “wobble” at onset; similarly, the unstable interaction between bow and string causes the tones of string instruments to have a high degree of pitch *jitter*. The relationships of pitch to other sound properties are also important. For example, the wide pitch variations of *vibrato* cause an instrument’s harmonic partials to interact with the resonant modes of the instrument, producing amplitude modulations, and these provide a wealth of information about the instrument’s resonant structure.
- **Loudness**: The intensity of an instrument’s sound interacts with other sound properties, producing salient cues. *Tremolo* (that is, sinusoidal variation of loudness) often accompanies *vibrato*, and the relative strengths of pitch *and* loudness variation may be salient. For example, flutes typically produce much stronger tremolo than strings or double-reeds.
- **Attack transient**: When listening to an isolated musical tone, listeners use information contained in the attack transient to identify the tone’s source. The rise-times—both absolute and relative—of the harmonic partials reveal information about the center-frequency and *Q* of resonances in the sound source. The low-amplitude “blips” preceding the tonal portions of some tones—particularly those produced by brass instruments—may also contain

useful information. Finally, it is possible that the non-harmonic, noisy, portions of the attack contain information that may be used to aid identification, but I am not aware of any demonstration of their use by human listeners.

- **Spectral envelope**: Several features of the relative strengths of a musical tone's harmonic partials reveal information about the identity of the tone's source. For example, the spectrum can reveal the center-frequencies of prominent resonances and the presence of zeros in the source-excitation. The relative strength of the odd and even partials can be indicative of the cylindrical tube (closed at one end) used in clarinets, and the irregularity of the spectrum can indicate a complex resonant structure as found in string instruments.
- **Inharmonicity**: Deviations from strictly integer-related partial frequencies are common in freely-vibrating strings, bells, and in the attacks of some instruments (saxophones, for instance).

The relative importance of these various features has not been studied in much depth, and typically, little is known about the ways in which they are extracted and represented by the human auditory system. The next chapter describes a set of signal-processing techniques and a series of representations at various levels of abstraction for many of the features described above, along with demonstrations of their extraction from recordings of orchestral instruments.

CHAPTER 4 Representation

Chapter 2 examined human sound-source recognition and compared several kinds of artificial recognition systems to the human system, highlighting their many limitations. Chapter 3 examined human abilities on a particular recognition task—identifying orchestral musical instruments from the sounds they produce—and described a set of acoustic features that could form the substrate of human recognition abilities in this small domain. This chapter builds on the insights gained from the previous two chapters. A series of signal-processing transformations are described, which convert an audio recording through a series of *representations* intended to highlight the salient features of orchestral instrument sounds.

4.1 Overview

A classic example of an artificial perceptual system is David Marr's model of early vision.[1] He used a series of increasingly abstract representations to describe visual scenes, starting with a raw image and culminating in an object description that could be matched against templates stored in memory. In his words:

> "A representation is a formal system for making explicit certain entities or types of information, together with a specification of how the system does this. And I

[1] The analogy between Marr's work and the system described in this dissertation is loose. I subscribe to the broad aspects of his modeling philosophy, but the system described here is not intended to be an auditory analog of his vision system. Marr explicitly decries the importance of high-level knowledge in perception, and I view this as a critical limitation of his work.

> shall call the result of using a representation to describe a given entity a *description* of the entity in that representation….[T]here is a tradeoff; any particular representation makes certain information explicit at the expense of information that is pushed into the background and may be quite hard to recover." (Marr, 1982, p. 20-21)

At each successive level in Marr's representation, the perceptually salient aspects of the image are more explicitly represented. At the first level, the raw image is transformed into a so-called "primal sketch," which makes intensity changes (some of which correspond to edges of objects) explicit, noting their geographical distribution and organization. At the second level, called the "2 ½-D sketch," the orientation and rough depth of *surfaces* are represented, making particular note of contours and discontinuities. Finally, the 2 ½-D sketch is transformed into a 3-D model representation that describes the shapes and spatial organization of objects in the scene from an object-centered viewpoint (because recognition demands a representation that does not depend much on the perceiver's viewpoint). These transformations are performed as a sequence of relatively simple stages because "it is almost certainly impossible in only one step" (Marr, 1982, p. 36).

4.1.1 Mid-level representation

Marr's intermediate representations are examples of what have been termed *mid-level representations* in the artificial intelligence literature. Referencing Marr's work, Ellis and Rosenthal (1995) provide a set of desiderata for auditory mid-level representations:

1. **Sound source separation**: As a signal is transformed through a set of representations, representational elements should correspond more and more to single sound sources or events. This feature is necessary to enable reasoning about individual components in an auditory scene.
2. **Invertibility**: The series of representational transformations should be invertible. Ellis and Rosenthal make too strong a demand in this case, requiring that "the regenerated sound be perceptually equivalent to the original". Although such a property may be desirable from a practical engineering standpoint, it is not necessary for many applications. As long as all perceptually equivalent sounds map into the same description in the representation, an ability to regenerate an acoustic signal is not necessary. However, it should be possible to use information contained in a particular representation to reason about the contents of lower-level representations (and this requirement, which bears little relation to resynthesis *per se*, may be necessary for disentangling mixtures of sounds).
3. **Component reduction**: At each successive level of representation, the number of objects in the representation should diminish and the meaningfulness of each should grow.
4. **Abstract salience of attributes**: At each re-representation, the features made explicit should grow closer to the desired end result, which in many cases will be the perceptually salient aspects of the signal.

5. **Physiological plausibility**: Given the goal of understanding the operation of the human auditory system, it is desirable for representational transformations to match those used by the brain. Of course, this is only important insofar as it serves the goals of the research, as discussed in Section 2.2.

Of these desiderata, the third and fourth are the most relevant to the current work. In addition, I would add that it is important for the representation to be robust with respect to sound-scene complexity (e.g., noise, the presence of multiple simultaneous sound sources, etc.). Although it is unreasonable to expect that the descriptions of the independent sources in an auditory scene be identical to their descriptions when heard in isolation, the system as a whole should be able to reason about noisy or obscured observations and their effect in the representation. Ellis's prediction-driven architecture does this well for its relatively low-level descriptions of noise beds, transients, and quasi-periodic signals, but it is not obvious how to identify and specify appropriate constraints for higher level descriptions.

Marr's low-level representations are symbolic, beginning at the level of the primal sketch, and this has some desirable effects. Transformation into symbols can ease some of the difficulty associated with noisy, incomplete data (Dawant & Jansen, 1991) and can be used to suppress unnecessary detail (Milios & Nawab, 1992). These features can lead to more robust analysis and decreased storage requirements, but it is important not to discard information that will be needed to resolve discrepancies arising at higher levels.

4.1.2 Features and classification

As was pointed out in Section 2.3, recognition systems cannot operate by memorizing every instance of every object that is to be recognized. Object identification is a pattern-recognition problem, and it is worthwhile to consider some of the general properties of pattern-recognition systems. Pattern-recognition systems (see, for example, Duda et al., 1997) operate by measuring a set of features from a representation of an object and then employing a classification function (usually learned during a training period) to make a classification decision. The classification function operates in a multidimensional space formed by the features. With an infinite number of "training" examples (i.e., for which the system is told the correct classification of each object), the classification function improves as additional features are added to the system. In realistic contexts, however, the number of training examples is limited, and the increased number of feature dimensions makes it increasingly difficult to find a good classification function because the "classification space" grows exponentially with the number of dimensions. This creates a kind a paradox, where it is both better and worse to have a large number of features.

One solution to the number-of-features paradox is to employ *meta*-features. By computing a large number of first-order features directly from the representation and then combining them intelligently into a smaller group of second-order features, the recognition engine can employ a small number of features that contain information equivalent to the larger set. If the problem is structured well, it may be possible to construct the system so that it does not matter if a particular subset of first-order features is missing (or is too noisy) in a particular sample—and that is an important feature because the particular set of features that is available will depend on the context. With this approach, the goal of the representational engine should be to generate a feature set that is as small as possible, yet still enables robust recognition.

The approach taken here is to avoid using classification algorithms that employ a large number of features at once. Instead, by using multiple classifiers, each operating on a small number of features, with some kind of voting scheme to combine their classifications, the curse of dimensionality can be alleviated. However, this approach may not take full advantage of the statistical relationships (e.g., correlations) between features, which given enough training data could be better exploited in the full-dimensional classification space.

The representational scheme used here is constructed of several different levels, as depicted in Figure 13, and is structurally similar to the one proposed by McAdams (Figure 3 on page 27). The following briefly describes the various components, each of which is described in more detail in the remainder of this chapter:

- **Raw signal**: The acoustic pressure waveform measured by a microphone represents the acoustic signal reaching the eardrum of the listener. For ease of analysis, it is stored in a data file.
- **Front-end**: The first stage of signal processing consists of a filterbank whose outputs are half-wave rectified, lightly smoothed, and then analyzed by short-term autocorrelation to make periodicity—the primary basis of pitch—explicit.
- **Weft**: The second stage of processing identifies stable periodicities in the signal that are likely to correspond to musical tones. Each periodicity is represented as a pitch-track and a corresponding time-varying spectral envelope.
- **Note properties**: A large number of features are extracted from the weft representation, corresponding to the properties we know affect human perception.
- **Source model**: The note properties are accumulated over time to form a model of the sound source’s excitation and resonance structure.
- **Model hierarchy**: The sound’s excitation/resonance model is compared to members of a hierarchically arranged set of stored reference models. The sound is “recognized” as an instance of the class represented by the model that matches most closely. (The recognition process is described in Chapter 5).

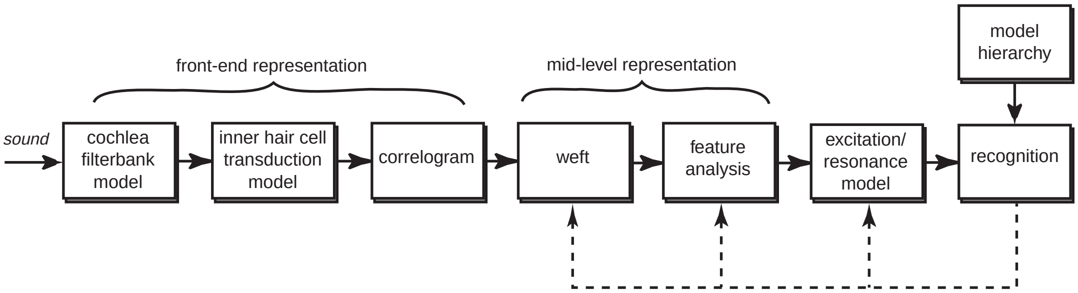

FIGURE 13. The representational/signal-processing scheme used here. The front-end consists of a fixed signal-processing network implemented in three stages. The mid-level representation makes explicit the many acoustic features known to be perceptually salient for human listeners. Recognition is based on a compact excitation/resonance model that integrates the many acoustic features into a simplified, abstract form. The feedback loops have not yet been integrated into the model.

4.2 The front end

The first representational transformation is implemented by a fixed signal-processing network called the front-end. It consists of three sub-stages that culminate in a three-dimensional representation called the correlogram, as shown in Figure 14. The implementation described here is modeled after the one described by Ellis (1996); differences between the two implementations are minor and will be described as they arise.

The sound-pressure wave itself is represented by a sequence of 16-bit fixed-point samples, recorded at 32,000 samples per second of sound (Ellis used a 22.05 kHz sampling rate). This representation is capable of coding vibration frequencies up to 16 kHz (the Nyquist rate, or “folding” frequency), so the sound wave is filtered before sampling to remove any higher frequencies. This bandwidth is sufficient to recreate a high-quality audio signal (better than FM radio broadcasts but not as good as compact discs). Many orchestral musical instruments produce frequency spectra that continue beyond 16 kHz (indeed, above 80 kHz in some cases!), though the spectra of most non-percussive instruments roll off well below 16 kHz (Boyk, 1997). The signals sampled at 32 kHz are quite sufficient for humans to recognize the instruments, as demonstrated by the experiment described in Chapter 6.

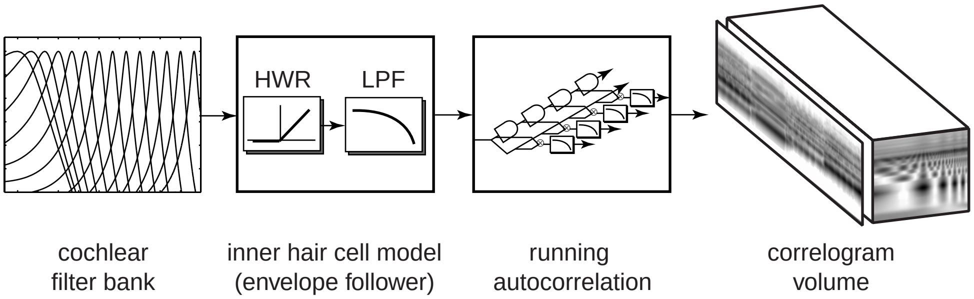

FIGURE 14. Detail of the front-end processing chain. Processing occurs in three discrete stages, modeling the frequency analysis performed by the cochlea, the nonlinear transduction of the inner hair cells, and a higher-level periodicity-based representation.

4.2.1 Bandpass filterbank

The first stage of signal processing consists of a fixed array of linear bandpass filters that model the frequency analysis performed by the *cochlea*. The cochlea is a bony, coiled, fluid-filled structure with two small openings covered by flexible membranes. At one end, a chain of tiny bones (the *ossicles*) attaches one of the flexible membranes (called the *oval window*) to the eardrum (called the *tympanic membrane*). When pressure variations (sound waves) reach the ear, they travel down the ear canal and cause the tympanic membrane to vibrate; the vibrations are transmitted across the ossicles to the oval window, where vibrations are transferred to the cochlear fluid. The cochlea's interior is separated into two main compartments by a set of flexible tissues that includes the *basilar membrane*. Vibrations travel the length of the basilar membrane, with high frequencies traveling further than low.

Any small region of the basilar membrane can be modeled as a bandpass filter (von Békésy, 1960), and although there are nonlinearities involved in the physiological chain to this point, they appear to be of secondary importance in relation to the bandpass frequency analysis, which is preserved at higher levels of the neural processing chain (Pickles, 1988). It is assumed that the breakup of the acoustic signal into various frequency bands is the primary function of the cochlea; at the very least, it is fair to say that we do not yet understand how the nonlinearities at the level of the cochlea help the hearing process.

The bandpass filter model I use is based on the one proposed by Patterson and his colleagues, which in turn is modeled after neurophysiological and psychophysical data (Patterson & Holdsworth, 1990; Patterson & Moore, 1986). The software implementation is modeled after Slaney's (1993). Each bandpass filter is implemented by four cascaded second-order filter sections, which realize an 8th order filter with a "gammatone" impulse response (an example, for a filter with a 1 kHz center frequency, is shown in Figure 15). The bandwidth of each filter is set to match the *equivalent rectangular bandwidth* (ERB) of the cochlear tuning curve at the corresponding frequency, as modeled by Moore and Glasberg (1983). The filter *Q*s, as a function of center frequency, are nearly constant (approximately 9.3) over much of the relevant frequency range. At low frequencies, the filters are somewhat broader (they have smaller *Q* values). For ease of implementation, the center frequencies are spaced evenly on a logarithmic scale, with six filters per octave, ranging from 31.25 Hz to nearly 16 kHz (in Ellis's implementation, center frequencies covered a smaller range, from 100 Hz to just over 10 kHz). This provides a significant overlap between adjacent filters (particularly at the lowest center frequencies), as shown in Figure 16.

Figures 17 and 18 depict the impulse responses themselves, illustrating their similarity on a logarithmic time scale. This similarity, which implies that the impulse responses are approximately time-scaled versions of a single function (in this case, the gammatone), is characteristic of *wavelet* transformations. The time-scale approximation is most accurate in the upper octaves. As is evident from Figure 17 the "center of mass" of the impulse responses varies with center frequency over a range of approximately 20 ms. This variation, called *group delay*, is compensated in the current implementation by the introduction of a pure delay element at the output of each filter. This compensation has no physiological (or even computational) justification, and it has no effect on recognition performance; it merely makes the representations at this and higher levels easier to "read" by a human observer.

To better illustrate the effect of this first representational transformation, consider a simple sawtooth waveform, beginning at (t = 10 ms) and repeating at 125 cycles/second (see top panel of Figure 19). An infinitely repeating sawtooth wave has a discrete Fourier spectrum with each component proportional to the inverse of its component number. When played through a loudspeaker, the waveform generates a buzzing sound with a pitch corresponding to the fundamental frequency of 125 Hz.

Figure 19 illustrates the response of the cochlear filter bank to the sawtooth waveform without group-delay compensation; Figure 20 shows the same response *with* compensation. In the main panels, the output of every second filter channel is depicted as a function of time and amplitude. The left panels illustrate the root-mean-squared (RMS) energy in each channel, as a function of center frequency, in alignment with the waveforms in the main panel. The upper panels display the waveform to illustrate the mis-alignment of the amplitude-modulation peaks across frequency (highlighted by an overlaid dotted line showing the variation of group-delay as a function of center frequency).

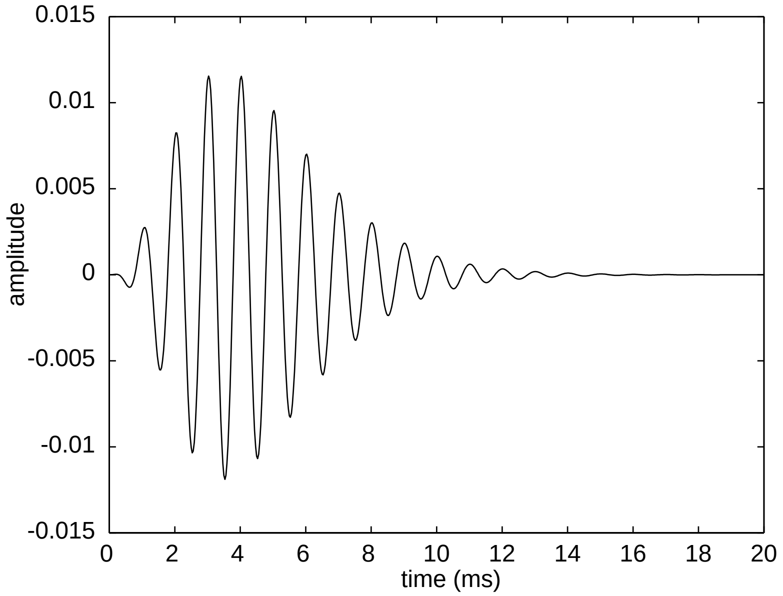

FIGURE 15. Impulse response of the cochlear bandpass filter centered at 1 kHz.

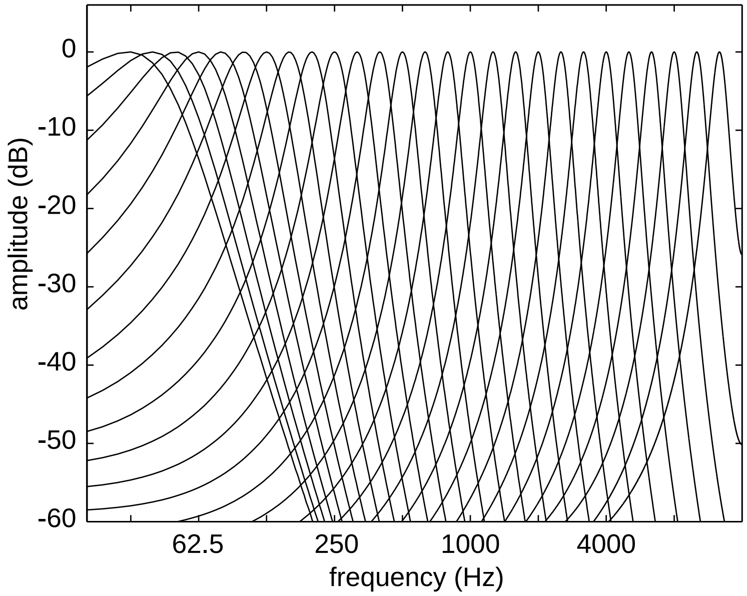

FIGURE 16. Overall frequency response of the cochlear filterbank, plotted on a logarithmic frequency scale (every second filter is shown).

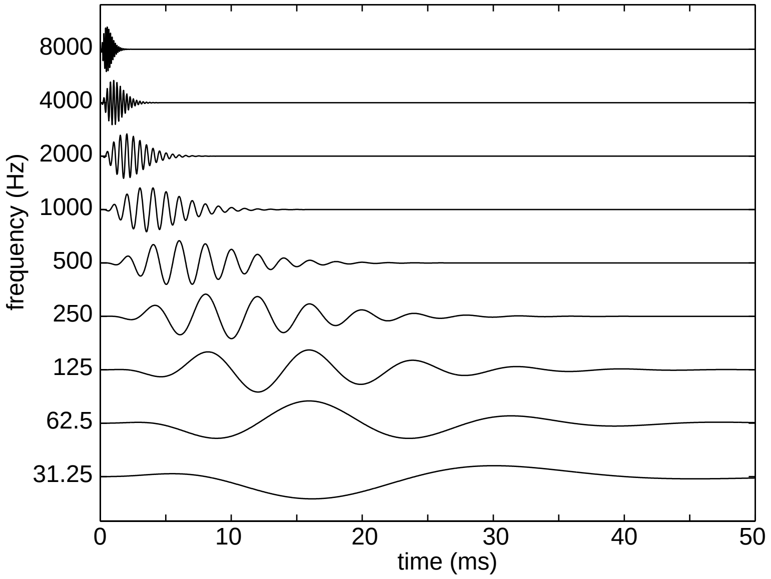

FIGURE 17. Impulse responses of nine cochlea bandpass filters (one filter is shown per octave). Their amplitudes have been normalized to a uniform scale for display purposes.

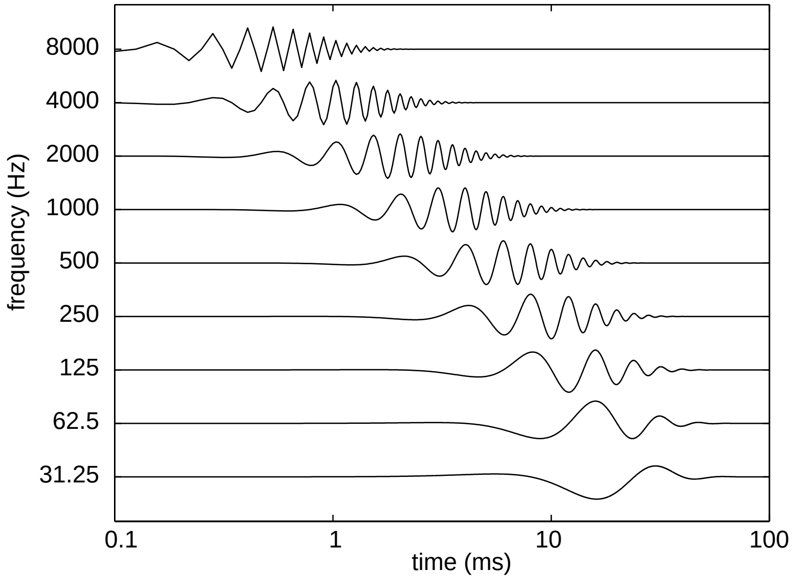

FIGURE 18. Impulse responses of nine cochlea bandpass filters (one filter is shown per octave) on a logarithmic scale. Their amplitudes have been normalized to a uniform scale for display purposes. Note the similarity of structure that is characteristic of a *wavelet* filterbank.

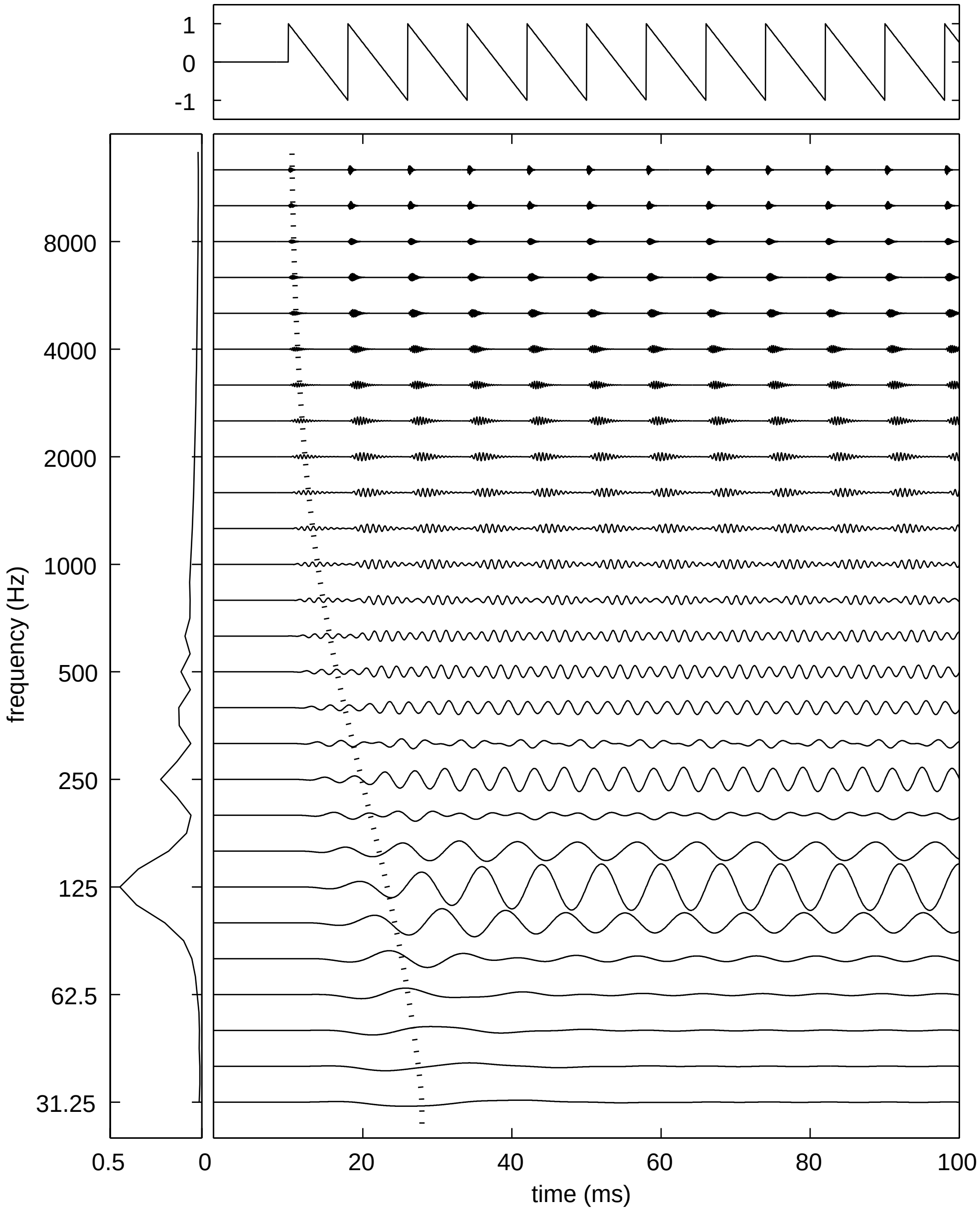

FIGURE 19. Response of the cochlear filter bank *without* group-delay compensation. The output of every second filter channel is shown. The left panel shows the RMS amplitude of the filters as a function of center frequency. The top panel shows the sawtooth waveform to illustrate the alignment of the amplitude-modulation peaks.

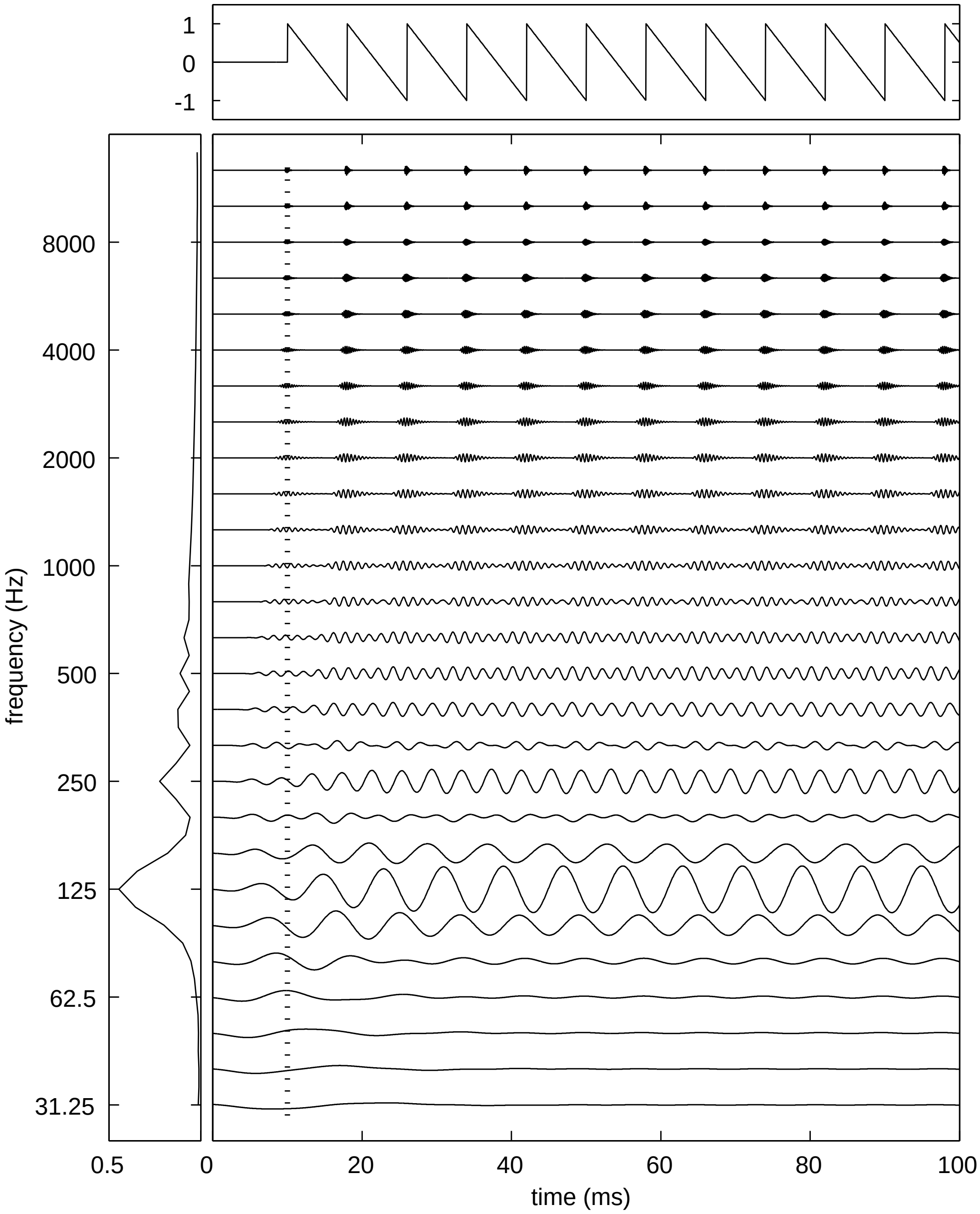

FIGURE 20. Response of the cochlear filter bank *with* group-delay compensation. The output of every second filter channel is shown. The left panel shows the RMS amplitude of the filters as a function of center frequency. The top panel shows the sawtooth waveform to illustrate the alignment of the amplitude-modulation peaks. Note the improved vertical alignment in comparison with Figure 19.

4.2.2 Inner hair cell transduction

The basilar membrane contains the *inner hair cells*, which act as transducers, converting the motion of the membrane in the cochlear fluid into electrical impulses. The inner hair cells have tiny embedded hairs (*cilia*) that bend when the basilar membrane moves relative to the cochlear fluid, and the cells emit electrical spikes with a probability that depends on the degree of deflection.

There are two properties of the inner hair cells that have particularly important effects on the signals transmitted to higher levels. First, the cells respond only to cilia deflection in one direction, and this introduces a half-wave rectification stage to the signal-processing chain. Second, at low frequencies, the hair cells tend to fire at a particular phase of the signal—a process called *phase locking*. As the frequency of the input signal increases, phase locking begins to run out at about 1.5 kHz and disappears by 5 kHz, but in the absence of locking to the *fine structure* of the waveform, the hair cells lock to the signal's amplitude envelope. This effect is simulated in the current implementation by a light smoothing operation (convolution with a 0.25 ms raised-cosine function[1]), which has little effect at low frequencies, but, in combination with the half-wave rectification, produces a reasonable envelope function at high frequencies.

Figure 21 shows the response of several cochlear filter channels after half-wave rectification and light smoothing. Several much more complex models of inner hair cell function have been developed (for example, several are compared in Hewitt & Meddis, 1991) that are more faithful to the nonlinear properties of mammalian inner hair cells, but the simple model described here was chosen for two reasons. First, as with the cochlear filters, we do not know what benefit additional nonlinearities bring to the hearing process. Second, the current implementation has the desirable property of preserving the relative energy levels in the various cochlear filters. Because the energy levels in the cochlear channels (and their variation over time) greatly affect human perception, it is desirable for intensity to be easily recoverable from the representation.

4.2.3 Pitch analysis

Pitch is one of the most important attributes of orchestral instrument sounds, and its relations to other acoustic properties form much of the basis of human sound-source recognition. In addition, pitch is thought to be one of the primary cues for auditory scene analysis. It is therefore desirable for pitch to be explicitly represented in any computational auditory scene analysis or sound-source recognition system. The third stage of the front end does exactly that.

[1] Ellis used a 1.0 ms window, but I found that it removed too much fine structure in the 2-5 kHz region. The rather shorter window used here (0.25 ms) may instead be *too* short.

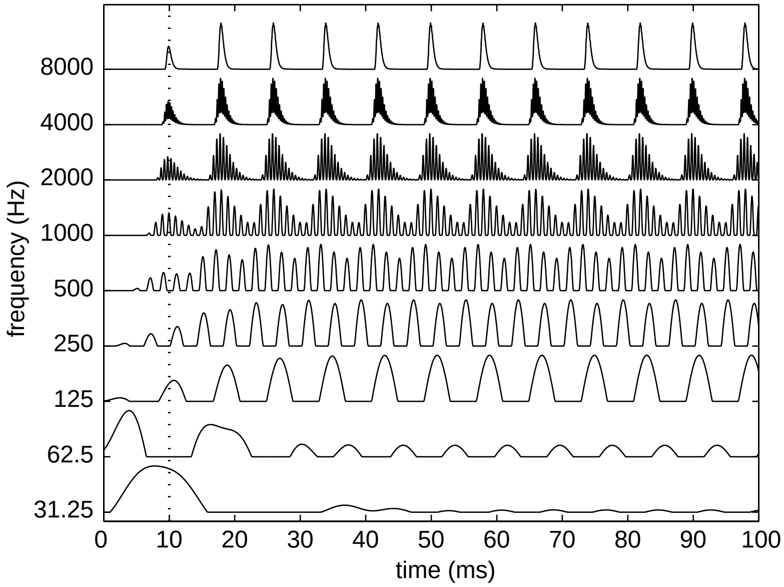

FIGURE 21. Responses of nine cochlea bandpass filters (one filter is shown per octave) to the 125 Hz sawtooth signal after half-wave rectification and light smoothing intended to model inner hair cell transduction. The output amplitudes have been normalized to a uniform scale for display purposes.

An approximately periodic signal will, in each cochlear filter output, produce an approximately periodic signal—with the same period as the full-bandwidth signal. This across-channel similarity of periodicity is the usual basis of human pitch perception. Autocorrelation is one of the conceptually simplest signal-processing techniques for discovering such periodicity in a signal. By multiplying the signal with delayed (time-shifted) versions of itself and measuring the average energy as a function of delay lag, it is possible to identify the underlying period of the signal. J. C. R. Licklider (1951) proposed such a mechanism, operating in parallel on the outputs of cochlear filters, as a possible basis for human pitch perception. Equation 8 is the usual definition of autocorrelation, with the integration ranging over the entire signal.

$$R_{xx}(\tau) = \int_{-\infty}^{\infty} x(t)x(t-\tau)dt \qquad \textbf{(8)}$$

In practice, it is impossible—and undesirable—to integrate over the whole signal. Pitch can vary over time, so the autocorrelation should be a *running*, or *short-time* operation applied to the signal. The representation that results will have three dimensions: cochlear position (corresponding to frequency), autocorrelation lag (corresponding to pitch period), and time, as shown in Figure 22.

A short-time operation implies some sort of averaging window, which can be applied in one of two ways. The usual approach is to apply the window first, before autocorrelation, as shown in Equations 9 and 10. Defining a window function *w(t)*, we have

$$x_w(t, t_0) = x(t)w(t - t_0) \tag{9}$$

$$R_{x_w x_w}(\tau, t_0) = \int_{-\infty}^{\infty} x_w(t, t_0)x_w(t - \tau, t_0)dt \tag{10}$$

These calculations can be implemented efficiently, using FFT operations to perform the autocorrelation in the frequency domain. Such an approach was described by Slaney and his colleagues (Duda et al., 1990; Slaney & Lyon, 1990; Slaney & Lyon, 1993) with reference to Licklider's original proposal. Meddis and Hewitt (1991a; 1991b) used a correlogram of this sort to model human pitch perception. They formed a *summary autocorrelation* by summing the contributions of each cochlear channel at each autocorrelation lag and identified the largest peak, which corresponds to the pitch period. With this model, they successfully demonstrated correlates of "the missing fundamental, ambiguous pitch, the pitch of interrupted noise, the existence region, and the dominance region for pitch" (Meddis & Hewitt, 1991a). A similar approach has been applied to the outputs of actual inner hair cells in a cat, using pooled inter-spike-interval histograms—which are very similar to autocorrelations—with similar results (Cariani & Delgutte, 1996a; 1996b). This style of correlogram processing was also used in two of the first computational auditory scene analysis systems (Brown, 1992; Mellinger, 1991).

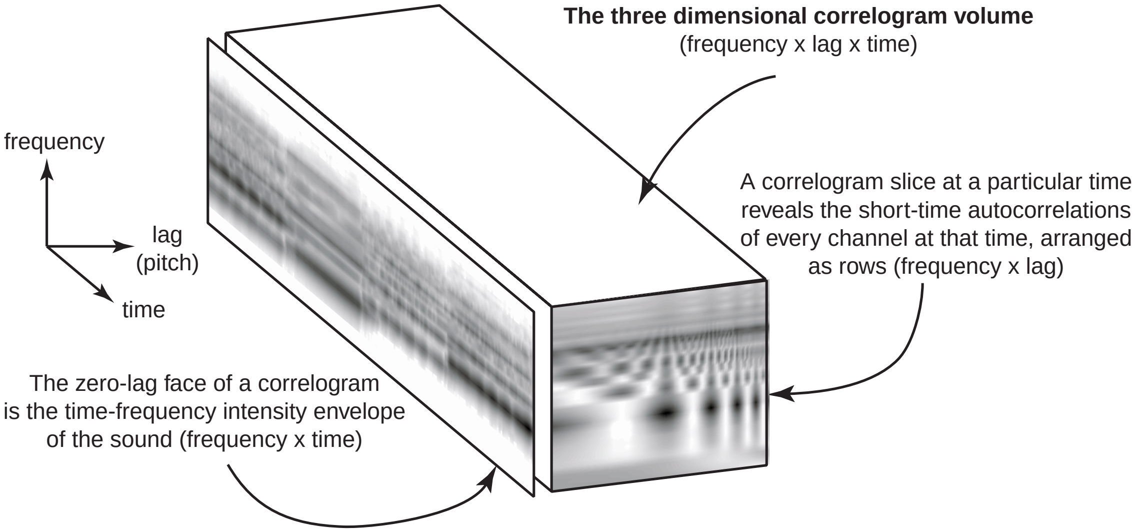
The three dimensional correlogram volume
(frequency x lag x time)
frequency
lag
(pitch)
time
A correlogram slice at a particular time reveals the short-time autocorrelations of every channel at that time, arranged as rows (frequency x lag)
The zero-lag face of a correlogram is the time-frequency intensity envelope of the sound (frequency x time)

FIGURE 22. Illustration of the correlogram volume, after Ellis (1996).

The “window-first” technique, and in particular its implementation with FFT-based autocorrelation, has several drawbacks. First, the length of the window limits the range of lags that can be calculated. Second, FFT-based methods usually sample the lag axis at uniform intervals on a linear scale. A logarithmically sampled lag axis makes more sense from a perceptual standpoint because human sensitivity to fundamental frequency differences is roughly constant on a logarithmic frequency scale (Moore, 1989). Defining a running autocorrelation by separating the window function from the multiplication of the signal with its delayed version, as shown in Equations 11 and 12, it is possible to sample any lag without regard to the window length.

$$R_{wxx}(\tau, t_0) = \int_{-\infty}^{\infty} w^2(t - t_0)x(t)x(t - \tau)dt \quad \textbf{(11)}$$

$$R_{wxx}(\tau) = [x(t)x(t - \tau)]*w^2(-t) \quad \textbf{(12)}$$

The portion of the correlogram corresponding to each cochlear filter can then be calculated using a tapped delay line, multiplying its output by the original signal, and smoothing (windowing) the output. A block diagram of the complete operation is shown in Figure 23. In the current implementation, fractional delay filters (Laakso et al., 1996) are used to calculate the delay line outputs, and the smoothing (window) filter consists of two cascaded one-pole lowpass filters, each with a 10 ms time constant (Ellis used a single, 25 ms, one-pole lowpass).

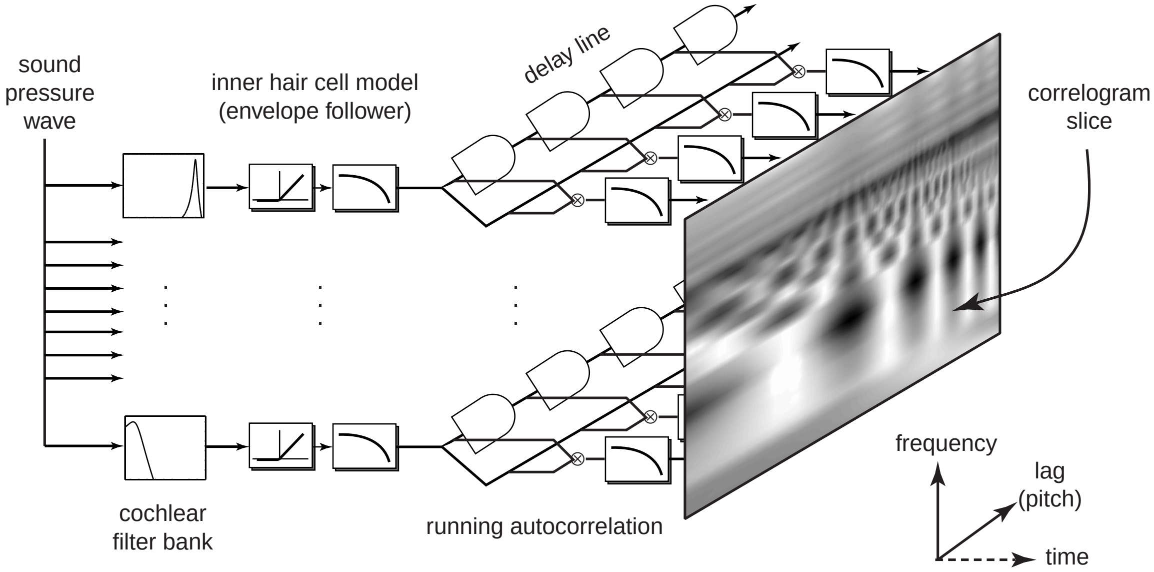

FIGURE 23. Block diagram of the calculation of the correlogram, after Ellis (1996).

There is a minor complication that arises from the logarithmic sampling. The bandwidth of a signal's autocorrelation function is equal to that of the signal itself, and in order to avoid aliasing (under-sampling the signal), it should be sampled at a rate at least greater than twice the highest frequency in the original signal. Therefore we must filter the signal so that it does not contain any frequencies higher than half the local sampling rate of the lag axis. This is accomplished by introducing another lowpass filter, prior to the delay line. In practice, this filter is combined with the "light smoothing" filter in the inner hair cell model, and separate tapped delay lines are used for various regions of the lag axis. This implementation is much more computationally expensive than the FFT-based version; however, it is well suited to parallel processing architectures.

The examples presented in this dissertation sample the lag axis at 150 lags from 0.33 ms to 33 ms, corresponding to fundamental frequencies from 30-3000 Hz (approximately the full range for musical pitch). This spacing includes approximately 23 lags per octave (nearly 2 per musical semitone), in contrast with Ellis's 48 (4 per semitone, for fundamental frequencies from 40-1280 Hz). A denser sampling would be desirable, but the current density was chosen as a compromise favoring computational speed and storage requirements over a more detailed representation. In practice, it is possible to interpolate between neighboring cells of the correlogram, so the limited sample density on the lag axis does not cause problems for higher levels of representation. The time axis is sampled at 2 ms intervals. This is a somewhat finer resolution than Ellis's 5 ms sampling, adopted mainly to improve visualization of instrument-tone onsets during later analysis.

4.3 The weft

The correlogram contains a great deal of information about the acoustic signal that it represents, but it is unwieldy. With 150 lags and 54 filter channels per slice and 500 time slices per second, it is a more than 125-fold expansion of the original sampled acoustic waveform (this calculation assumes 16-bit samples; with 32- or 64-bit floating-point samples, the growth increases). The weft representation addresses this drawback.

The *weft* is a novel representation for quasi-periodic, pitched sounds, which was proposed by Ellis and Rosenthal (1995) (and refined by Ellis (1996), from which this presentation is adapted) to address the limitations of traditional sine-wave models. The name comes from a weaving term for a parallel set of threads running through a woven fabric. A quasi-periodic input waveform creates vertical "spines" in the lag-frequency plane (e.g., Figure 24a) that change slowly as a function of time, and the values measured along the spines correspond to the energy associated with the given lag (here, we may say pitch period). Traced along time, these spines form a weft, as shown in Figure 24b. Because a periodic waveform with period T is also periodic at integer multiples of T, the spine pattern is repeated at multiples of the pitch period (corresponding to *sub-harmonics* of the pitch frequency). Only one weft is needed to represent the entire set of subharmonics; indeed, a single weft—stored as a *period track* and a corresponding

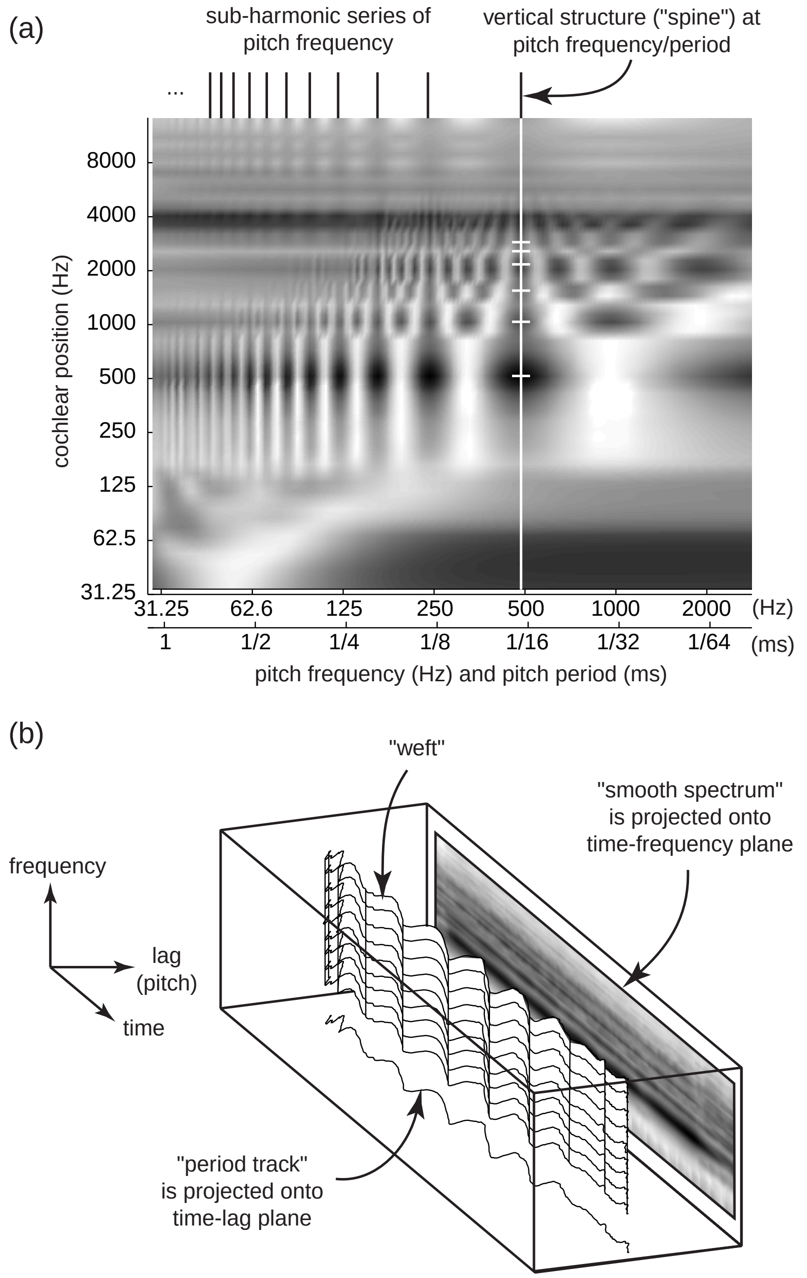

FIGURE 24. The weft calculation. (a) A correlogram slice during a violin tone performed with a pitch near 500 Hz. Note the vertical structure present at the pitch frequency/ period, indicated by the vertical white line, and at its subharmonics/harmonics. The cross-hatch marks indicate the approximate frequency regions of the first six harmonic partials. (b) Spines are accumulated over time. The period track is given by the spine position; the smooth spectrum is given by the energy along the spine as a function of frequency.

smooth spectrum—is sufficient to represent the harmonic portion of any quasi-periodic signal.

The weft can be viewed as a source-filter model, with the period track controlling a quasi-periodic impulse generator whose output is modulated by a time-varying filter controlled by the smooth spectrum, and this model can be used to resynthesize portions of the acoustic waveform. Taking this view, we can express the quasi-periodic impulse excitation as

$$e(t) = \sum_i \delta(t - t_i) \tag{13}$$

with

$$t_i = \arg_t \left\{ \int_0^t \frac{2\pi}{p\tau} d\tau = 2\pi \cdot i \right\} \tag{14}$$

where $p(\tau)$ is the period track. The output signal can be expressed as

$$x_w(t) = [e(\tau) * h_w(\tau;t)](t), \tag{15}$$

where $h_w(\tau;t)$ is the time-varying impulse response of the filter corresponding to the smooth spectrum. The task of weft-analysis is to recover $p(\tau)$ and $h_w(\tau;t)$ (usually thought of in the frequency domain, as $H_w(\omega;t)$). This decomposition is not unique, but it is the simplest to define and is relatively simple to compute.

Ellis (1996) describes a complicated algorithm for recovering the period track and smooth spectrum of multiple, overlapping wefts from the correlogram volume, even when the quasi-periodic portions of the acoustic signal are “corrupted” by wide-band and transient noise. Readers interested in the processing details should consult his excellent presentation. However, most of the details of the weft extraction algorithm are unnecessary for the discussion here. The signals used in this work are simpler than those used by Ellis, and my implementation simplifies Ellis’s algorithm in several ways.

With the assumption that the input signal contains only one source of quasi-periodic vibration, it is relatively simple to recover the period of vibration given a single time-slice of the correlogram volume. The most commonly used method (and the one used by Ellis) is to integrate over the cochlear position dimension to create a *summary autocorrelation*. The pitch of the signal—at that time—is then given by the lag exhibiting the largest peak. As mentioned earlier, this simple method, with minor variations, has been used as a model of human pitch perception with good results on a wide range of examples (Meddis & Hewitt, 1991a; 1991b). The principle weakness of the summary-autocorrelation approach is that it is prone to (sub)harmonic errors—that is, it occasionally generates pitch estimates that differ from human pitch judgments, most often by an octave, because the “wrong” peak is chosen accidentally.

The approach taken here is more complex, but more robust for signals generated by orchestral instruments. Rather than find peaks in a summary of the correlogram slice, the current implementation searches for sets of local maxima at the same pitch period across a range of cochlear channels. Each such collection of maxima is considered in turn, and the one that best explains the entire correlogram image is selected. The selection is based on two simple heuristics, and is an exercise in rule-based programming. First, if the total energy (the sum of the values at the maxima) of a set is much smaller (by some threshold) than that of another set, the weaker set is discarded. Second if the pitch periods of two sets with similar total energy is related by an integer ratio, the set with the larger pitch period is discarded.

After the pitch period is determined, a cross-section is taken of the correlogram slice, recording the energy at the pitch period as a function of cochlear position. This is a first-order approximation to the *smooth spectrum* at the time corresponding to the correlogram slice. Ellis makes several refinements to this measurement. First, he uses the local peak-to-trough energy ratio in the correlogram slice to estimate the energy of locally wide-band noise in the channel. He then subtracts the result from the smooth spectrum. Second, he uses non-negative least-squares (NNLS) inversion to account for the overlap of the filter channels. These refinements are not used in the current implementation, in part because they are computationally expensive (NNLS is an iterative procedure). Also, the wide-band noise components in the recordings used to train and test the system were relatively small, so the refinements would not change the computed representation drastically.

Although the smooth spectrum computed in this way does not correspond exactly to the time-varying *spectral envelope* of the instruments analyzed, it is a reasonable approximation. It has several desirable qualities from a perceptual-modeling viewpoint. For example, the spectral envelope is computed with a local resolution corresponding to the bandwidth of the appropriate cochlear channels. This means that the first 4-6 harmonic partials of a quasi-periodic sound *of any pitch* are *resolved* and can be analyzed individually. Higher harmonics are represented as overlapping groups within so-called *critical bands*. Human listeners perceive only the *group properties* of partials above roughly the 6th, and this limitation is inherent in the representation at this level.

The recordings used to test the current implementation are of solo instruments playing musical phrases and isolated notes. A single weft very naturally represents an isolated note, and as long as a phrase is played one-note-at-a-time, a single weft can represent it. However, since it will be useful to analyze properties of single notes within phrases (e.g., for their attack properties), the period track is segmented into regions corresponding to relatively stable pitch periods. Each segment forms a separate weft, usually corresponding to a single note. This segmentation stage is not strictly necessary, and it may create problems for some musical signals, such as a soprano singing with extremely exaggerated vibrato or a jazz trombone played with pitch glides. It also does not correspond strictly to my view of music perception, in which a rapid sequence of notes may be heard as a single entity rather than as a series of separate entities. Segmentation is adopted here only because it simplifies certain parts of the next stage of representation, at least conceptually.

The weft elements do not contain information about any non-periodic components of the input signal. This means that, for example, bow, breath, and valve noises are not represented at this level. Although such components would be needed to fully explain human sound-source recognition abilities, they are not necessary to account for a great deal of the human experimental data, as will be demonstrated in Chapter 6.

4.4 Note properties / source models

In the next representational stage, perceptually salient features are measured from the weft representation and accumulated over time to form a model of a sound source as it is heard. Because the weft is already made up of perceptually salient components, feature extraction is generally very simple and is accomplished with heuristic signal-processing techniques. In this section, feature extraction is illustrated with example tones produced by six instruments (representing classes with distinct excitation and resonance properties). Short segments of the period tracks and smooth spectra for six sample tones, performed respectively by violin, trumpet, oboe, clarinet, flute, and alto saxophone, are shown in Figure 25. In examples where information is integrated over multiple notes, recordings of chromatic scales are used for illustrative purposes.

The features extracted from the weft representation are of two types. Some are direct measurements on a physically meaningful scale, such as a ratio of energies; others are pseudo-binary indicator features, representing the presence or absence of a particular attribute. Not every feature is applicable to every sound source, and, in particular, some features are hierarchically dependent on others. For example, although it might make sense to define the “vibrato depth” of a non-vibrato note to be zero, the relative strength of amplitude- to frequency modulation (a ratio) has no meaningful definition in the absence of vibrato.

The representation at this level consists of a *frame* (Minsky, 1974) for each sound source, or sound-source category, with each frame containing a *slot* for each feature. Because each sound source may have a different set of applicable features, the set of slots may vary from one frame instantiation to another. In Chapter 5, examples will be given of *methods* attached to particular slots, and of default slot values inherited from parent nodes. For now, the frames may be thought of as feature lists, temporarily ignoring the more powerful attributes of the representation.

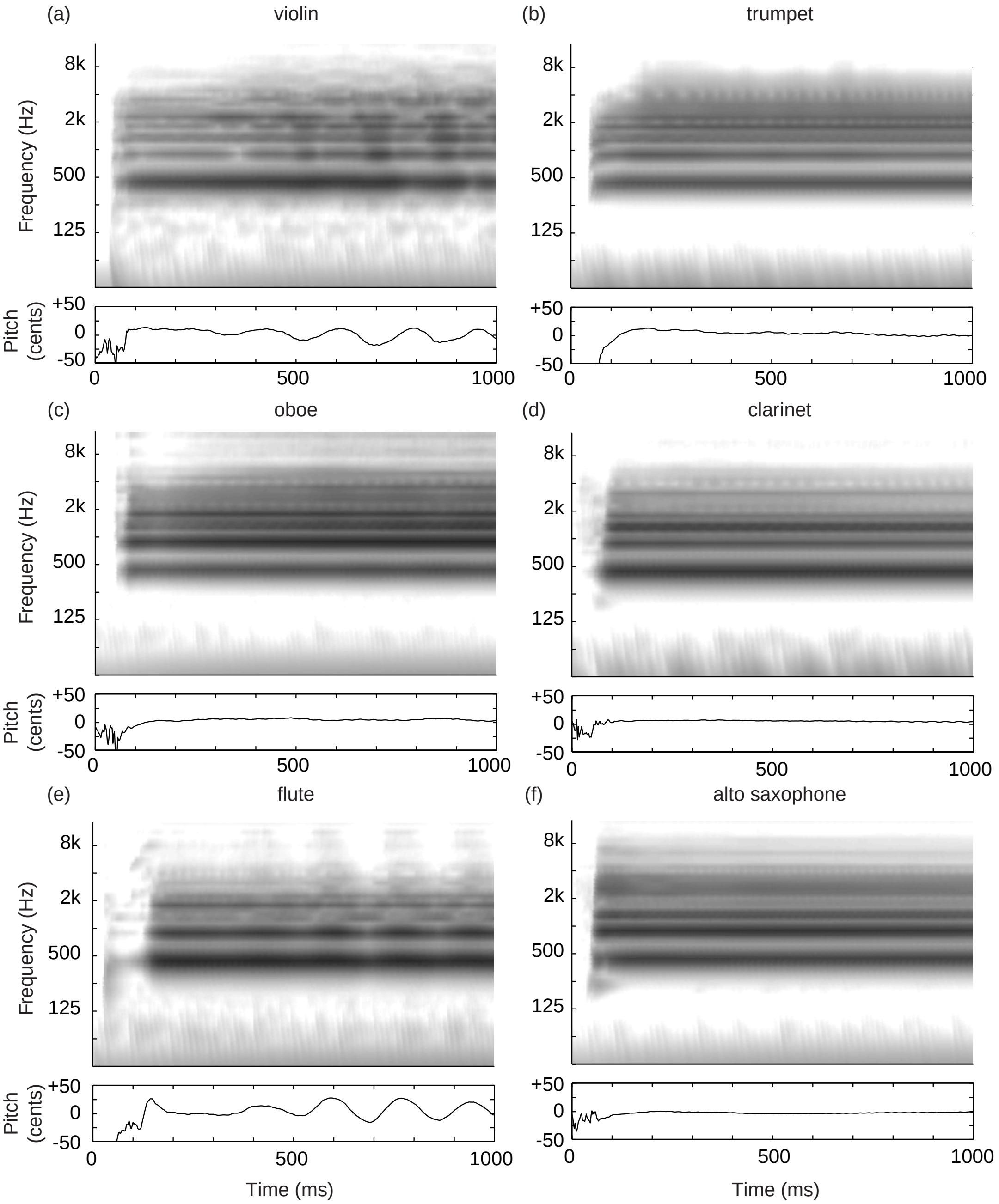

FIGURE 25. *Period tracks and smooth spectra* for example tones produced by (a) violin, (b) trumpet, (c) oboe, (d) clarinet, (e) flute, and (f) alto saxophone. In each case, the main panel shows the smooth spectrum as a function of time (abscissa) and cochlear frequency (ordinate); energy level is indicated by intensity, ranging from low (white) to high (black) over a range of 75 dB. The lower panel displays the period track, expressed deviation in *cents* from 440 Hz (a logarithmic scale, with 100 cents equivalent to a musical semitone, or a frequency/period ratio of $2^{1/12}$).

Rather than storing the feature values themselves, statistical summaries of the feature values are accumulated as sounds produced by a particular source are heard. In general, observations of each feature are assumed to have been generated by a process with a Gaussian distribution, for which the sample mean and standard deviation are *sufficient statistics* (Duda et al., 1997). Along with these values, the number of samples used in their calculation is recorded, so that statistics from multiple source models may be pooled later on.

Because instruments may behave differently in different portions of their pitch range, many of the feature statistics are accumulated separately as a function of pitch or frequency range. It will be obvious from the presentation when this is the case. Table 3 lists the features considered in this chapter.

Spectral Features	Spectral centroid (and relative spectral centroid)
	Average relative spectrum
	Average relative spectrum by partial #
	High-frequency rolloff rate and cutoff frequency
	Spectral irregularity and # of “zeros”
	Relative energy in odd and even partials
Pitch, Vibrato, and Tremolo Features	Pitch range
	Tremolo: absolute strength and relative (to vibrato) strength and phase
	Centroid modulation: absolute strength and relative (to vibrato) strength and phase
	Individual harmonic amplitude modulation: absolute strength and relative (to vibrato) strength and phase
	(pitch “wobble”)
	(pitch jitter)
Attack Features	Relative onset time by partial frequency
	“Rise likelihood” by frequency and post-onset time
	(# of “blips”)
	(Explicit onset skew)
	(Rise rates)
Other Possibilities	(Inharmonicity)
	(Note-to-note transitions)
	(Explicit identification of resonances)
	(“Cognitive” cues)

TABLE 3. Features considered in this chapter. Features in parentheses have not been implemented.

4.4.1 Spectral features

As discussed in Chapter 3, the harmonic spectrum contains a great deal of information about the sound source, possibly including, for example, the center-frequencies of prominent resonances and the presence of zeros in the source-excitation. Relatively weak strength of even partials can be indicative of the cylindrical air column (closed at one end) used in clarinets, and overall irregularity of the spectrum may be indicative of the complexity of a sound source's resonance structure. These features and others are readily computed from the weft representation.

The *spectral centroid* is a simple feature that correlates strongly with the perceived *brightness* of a sound. It is trivially calculated from the smooth spectrum of the weft representation by computing the first moment of the energy as a function of frequency, using the cochlear-channel index, *k*, as a log-frequency axis:

$$C = \frac{\sum_k kE_k}{\sum_k E_k} . \qquad \textbf{(16)}$$

Here, *E(k)* is the energy in cochlear channel *k*. The result may be converted to a frequency scale by the following transformation:

$$f = 1000 \times 2^{\frac{(k-31)}{6}} \qquad \textbf{(17)}$$

This equation is based on the current implementation, for which the center frequency of channel 31 is 1000 Hz (the 1/6 factor arises because there are six cochlear channels per octave). It is worth noting that these measures are not invariant with respect to overall coloration of the audio signal.

Because the relationship between pitch and brightness is important to the perception of musical sounds, the *relative spectral centroid*, calculated as the ratio of the spectral centroid to the pitch, is a useful feature. Using the period track from the weft representation, the relative centroid can be calculated by multiplying the spectral centroid (on a frequency scale) by the pitch period (in seconds). This is equivalent to dividing by the pitch frequency. The calculation can also be performed by converting the pitch frequency into its equivalent filter-channel index and then subtracting the result from the spectral centroid expressed the same way (this is due to the trivial equivalence of subtraction of logarithms to division). The mean spectral centroid and relative spectral centroid—estimated as a function of pitch from recordings of chromatic scales—are shown for the six instruments in Figures 26 and 27.

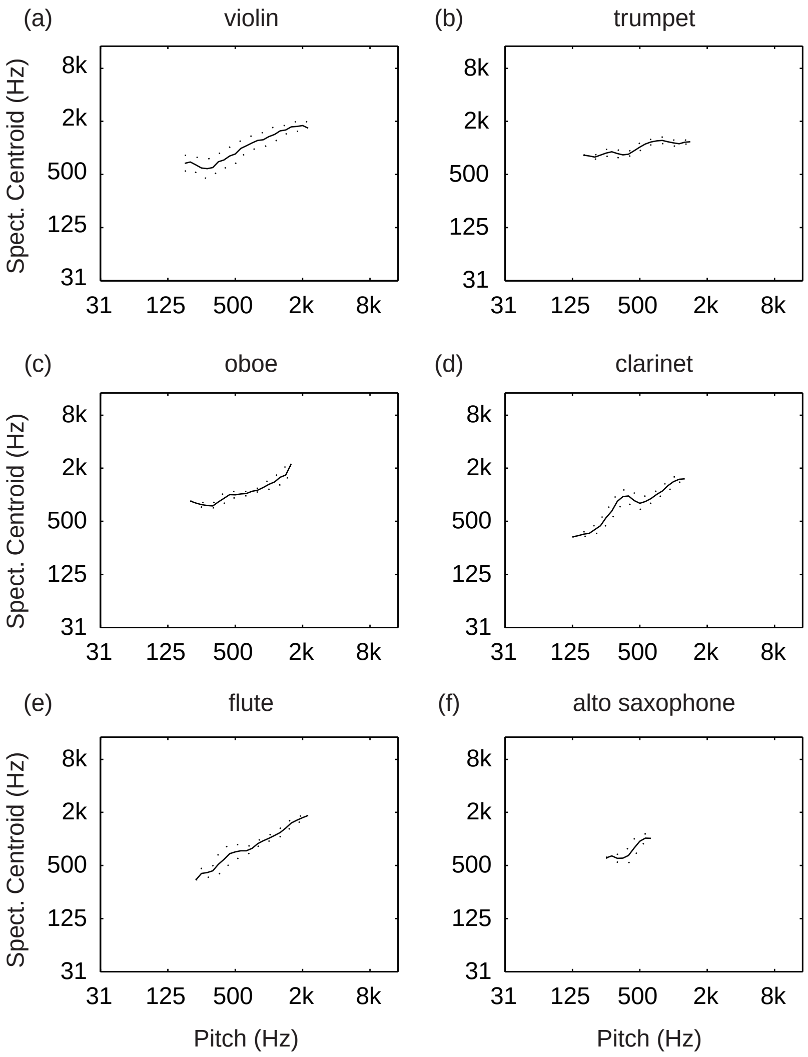

FIGURE 26. *Average spectral centroid as a function of pitch*, estimated from chromatic scales performed by (a) violin, (b) trumpet, (c) oboe, (d) clarinet, (e) flute, and (f) alto saxophone. The abscissa is the pitch frequency, and the ordinate is the spectral centroid, expressed as a frequency. In each case, the solid line indicates the mean value for tones at that pitch, and the dotted lines indicate bounds of one standard deviation.

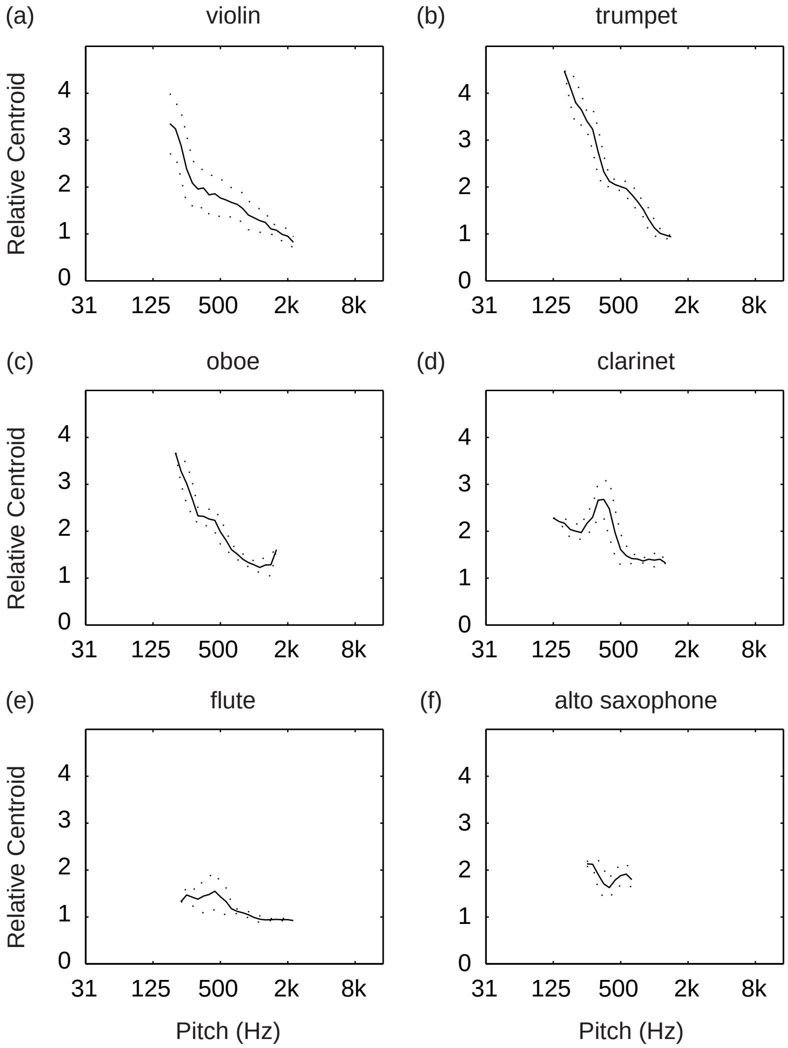

FIGURE 27. *Average relative spectral centroid as a function of pitch*, estimated from chromatic scales performed by (a) violin, (b) trumpet, (c) oboe, (d) clarinet, (e) flute, and (f) alto saxophone. The abscissa is pitch frequency, and the ordinate is the relative spectral centroid, expressed as a ratio of spectral centroid frequency to pitch frequency. In each case, the solid line indicates the mean value for tones at that pitch, and the dotted lines indicate bounds of one standard deviation.

The period track and smooth spectrum of the weft representation can used to estimate the relative strengths of the harmonic partials comprising a musical tone. Given the pitch frequency, it is straightforward to identify the filter-bank channels that are dominated by each of the first six harmonic partials, simply by comparing their center frequencies to the expected frequencies of each partial (which are just integer multiples of the pitch frequency). The energy levels in those channels are taken as estimates of the energy levels of the corresponding partials. Estimates are made in the same way for partials above the 6^{th}, with the caveat that more than one partial influences the energy in any given cochlear channel. For each region of pitch-period stability in the weft's period track (usually corresponding to a single musical tone), the maximum energy for each resolved partial (or channel containing unresolved partials) is determined. The results are shown in Figure 28 for single tones produced by the six instruments. The harmonic spectrum is normalized by its maximum value, and the *average relative spectrum* is accumulated as a function of frequency, with three separate estimates calculated: one from the first three odd-numbered partials, one from the first three even-numbered partials, and one from the entire spectrum. The average relative spectra of the six instruments (based on all partials, except for the clarinet, for which both the odd and even estimates are displayed) are shown Figure 29.

In addition, the relative levels of the first six partials are stored as a function of pitch. This representation highlights the reduced strength of the first partial in double-reed instruments, the reduced even partials in the clarinets, and the simple formant structure of the brass instruments. Figure 30 shows the strength of the first six partials as a function of pitch frequency for the six instruments.

Several subsidiary features are also computed from the harmonic spectra measured from individual notes. For example, the average difference between the energy of a partial and its two neighbors is computed as a local measure of spectral irregularity and accumulated both as a function of frequency and of partial number. Partials with particularly low energy levels relative to their neighbors are noted, as they may correspond to zeros of the excitation spectrum. Alternatively, they may be due to a suppression of “even” harmonics in a cylindrical vibrating air column (as in the clarinet), or to suppression of the first partial (as in the bassoon). In addition, a line is fit to the high-frequency roll-off of the spectrum (in dB energy versus log frequency). The slope of the line (in dB/octave) is recorded as the high-frequency roll-off rate, and the frequency at which the line crosses the maximum energy level of the spectrum is recorded as an estimate of the spectrum's cut-off frequency. Both the roll-off slope and cut-off frequency are accumulated as functions of pitch frequency.

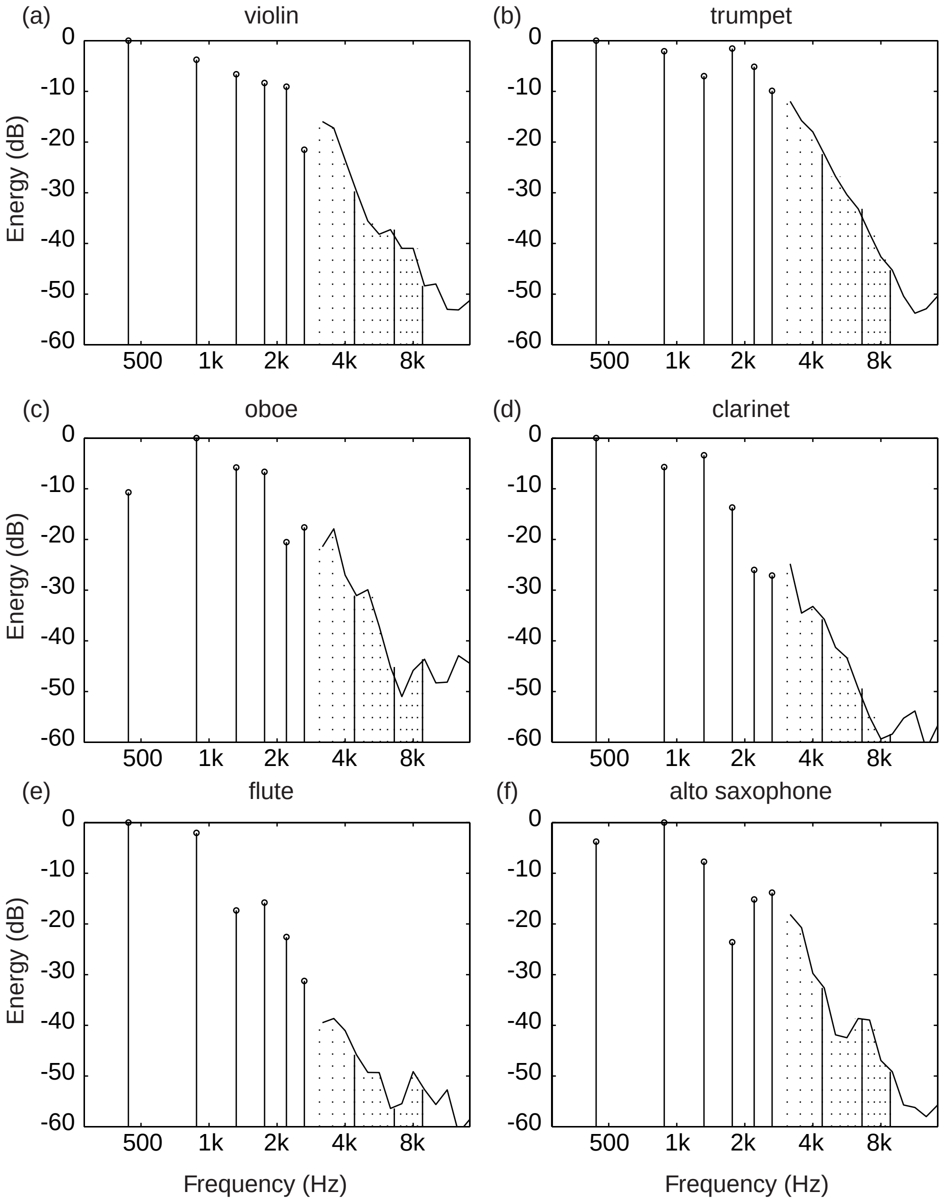

FIGURE 28. The *maximum values of the harmonic spectra* for isolated tones performed by (a) violin, (b) trumpet, (c) oboe, (d) clarinet, (e) flute, and (f) alto saxophone. In each case, the energies of the first six partials are estimated independently. Above the sixth, energy is measured by cochlear channel rather than by partial number because multiple partials mix in each cochlear channel. The abscissa is frequency; the ordinate, relative energy (in dB). The frequencies of the first 20 partials are indicated by vertical lines (dotted lines, above the sixth partial).

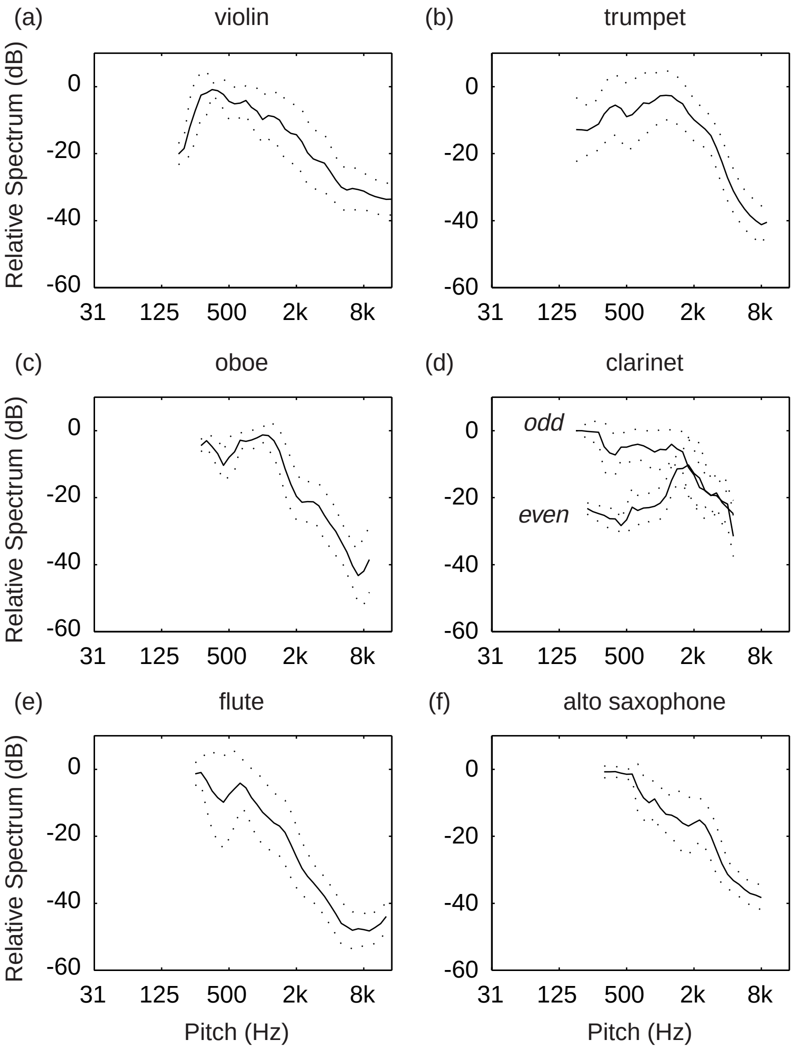

FIGURE 29. The *average relative spectra* measured from chromatic scales performed by (a) violin, (b) trumpet, (c) oboe, (d) clarinet, (e) flute, and (f) alto saxophone. In each case, the solid line results from an average over all harmonics, and the dashed lines indicate bounds of one standard deviation. The abscissa is frequency; the ordinate, relative energy (in dB). In each case, the solid line indicates the mean value for partials at that frequency, and the dotted lines indicate bounds of one standard deviation. In panel (d), the relative spectra computed using the low odd- and even-numbered partials are shown because they differ significantly (compare to Figure 12 on page 61).

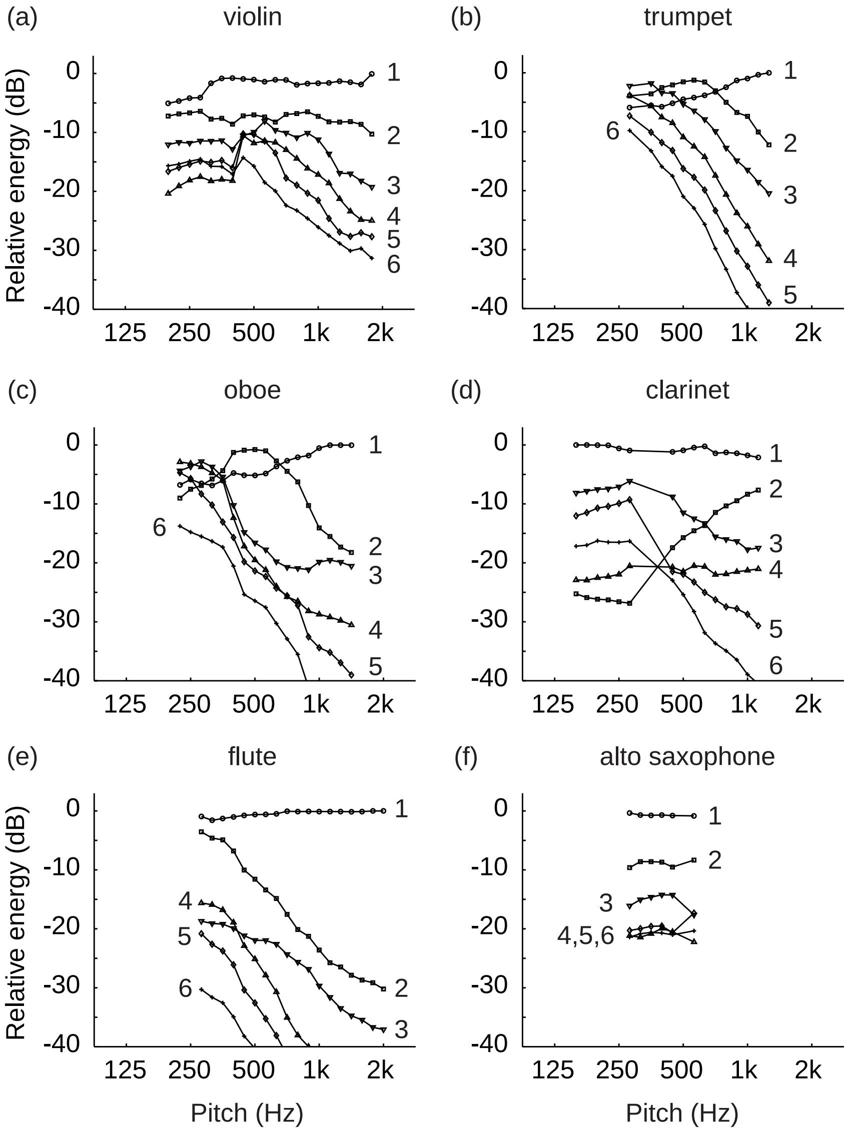

FIGURE 30. *Average strength of the first six partials as a function of pitch frequency*, measured from chromatic scales performed by (a) violin, (b) trumpet, (c) oboe, (d) clarinet, (e) flute, and (f) alto saxophone.

4.4.2 Pitch, vibrato, and tremolo features

As described in Section 4.3, the pitch of a sound is made explicit by the weft, which represents it as a function of time. Pitch is a useful feature on its own for ruling out sound-source hypotheses during the recognition process, but it becomes even more useful when considered in combination with other features. In Section 4.4.1, pitch was used as the abscissa in many of the feature representations. In this section, the pitch range of a sound-source is represented explicitly, along with the effects of the source's resonance structure when the performer applies a periodic pitch variation (vibrato). Other features that may have an affect on human recognition, including pitch "wobble" during the attack of brass tones and random variations, or *jitter*, have not yet been included in the framework described here, although they may readily be computed from the weft representation. For now, they are postponed as obvious future developments to the current system.

The pitch range of a sound source is represented by a histogram of 1/6-octave bands. The value in each histogram bin is simply the period of time for which sounds in the corresponding pitch-frequency range have been observed. The maximum value of a histogram bin is limited to ten seconds, an *ad hoc* threshold representing "sufficient" evidence that the sound source can produce sounds in that pitch range. Histograms accumulated for chromatic scales performed by the six instruments are shown in Figure 31.

As described in Chapter 3, vibrato is a performance technique whereby a player imposes a nearly periodic pitch variation—with a period in the neighborhood of 6 Hz—on the steady-state pitch frequency of the note being played. In order to detect this variation, the period track of the weft representation is converted to pitch frequency, and a short-time discrete Fourier transform is computed over the modulation frequency range from 2-15 Hz, using a 400 ms Hamming window and a 50 ms hop size. If the spectrum exhibits a peak in the 4-8 Hz range, the peak's amplitude (in cents) and phase are recorded, along with the relative time (measured in hops). Using the smooth spectrum component of the weft, the same process is applied to the spectral centroid (expressed in *channels*), the total energy (expressed in dB), and to the energy of each of the first six harmonic partials.

These first-order features are then organized into several second-order features. The modulation strength of the total energy is termed the *tremolo strength*. The mean and variance of its amplitude (in dB) is recorded as a function of pitch, as is its amplitude relative to the vibrato strength (expressed in dB/cent). The phase of the amplitude modulation is compared to that of the frequency modulation, and the probability of the two being out of phase is recorded. Similarly, the absolute and relative strength of the spectral centroid modulation is recorded as a function of pitch, along with the probability of being out of phase with the frequency modulation. Finally, the absolute and relative modulation strengths and phase for each of the first six partials is recorded as a function of partial number and of frequency (compiled across all six).

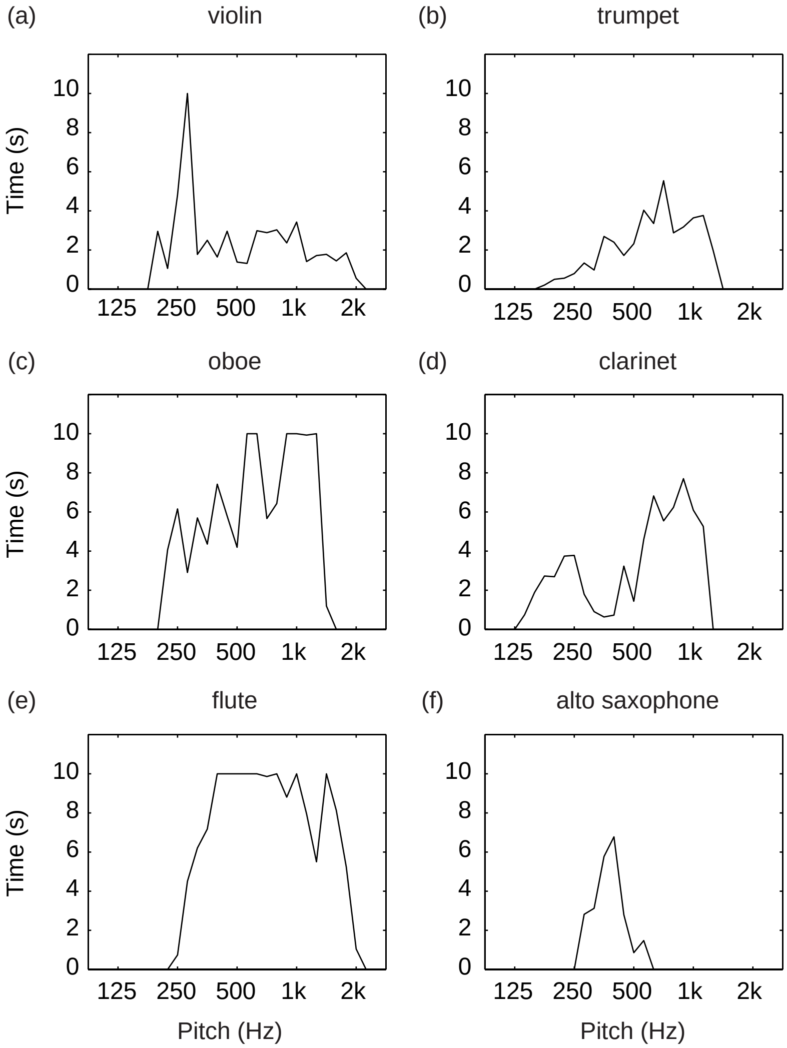

FIGURE 31. *Pitch range histograms*, in 1/6-octave bins, measured from chromatic scales performed by (a) violin, (b) trumpet, (c) oboe, (d) clarinet, (e) flute, and (f) alto saxophone. The abscissa is pitch frequency; the ordinate, time (in seconds; each bin is limited to 10 seconds as described in the text).

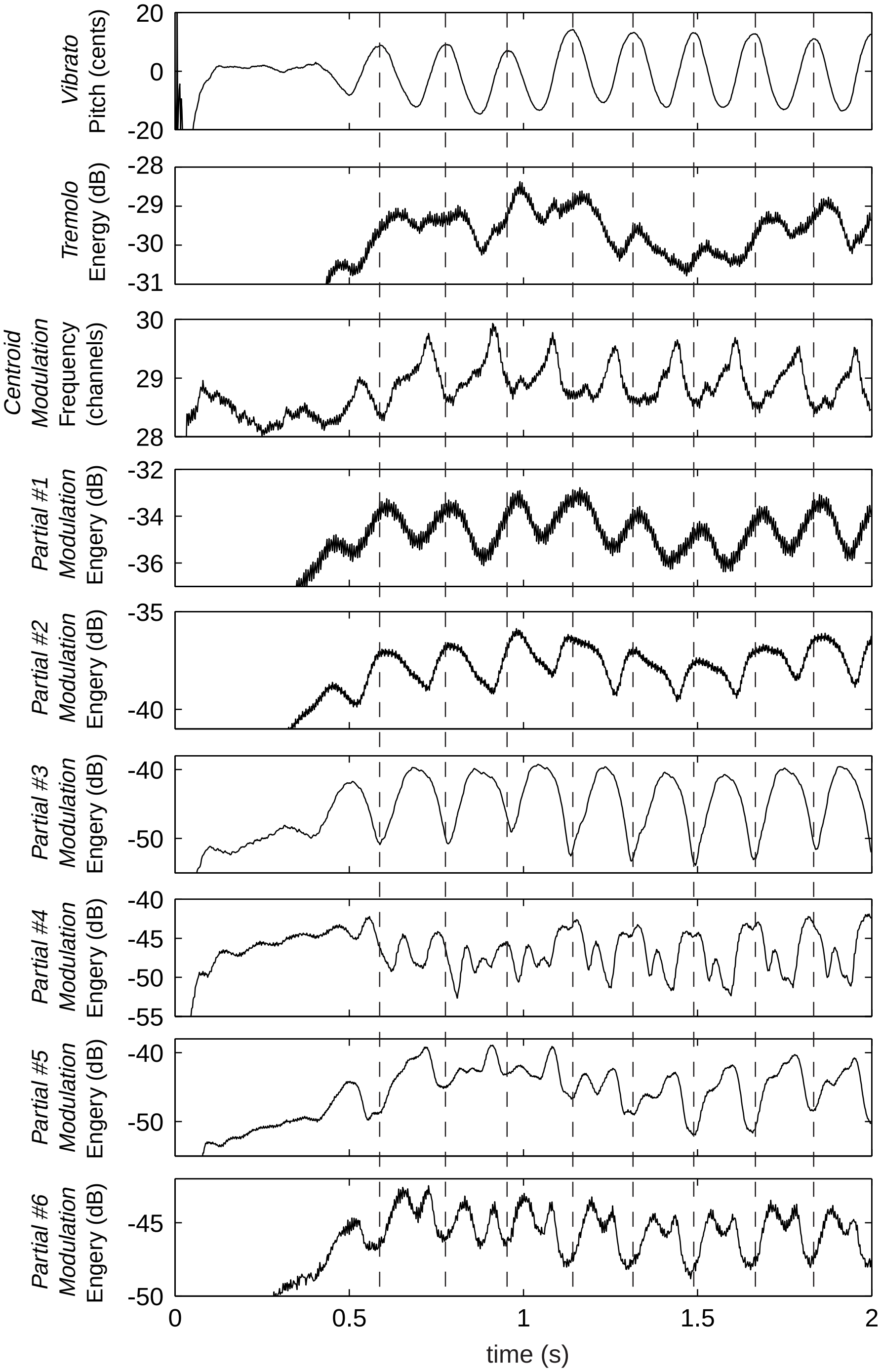

FIGURE 32. *The effect of vibrato on a violin tone*. Each panel shows a separate feature: pitch, total energy, spectral centroid, and the energy of each of the first six partials. The dashed lines are superimposed at the maximum value of the pitch signal for each cyle of vibrato, showing how some of the features vary out of phase with others.

Figure 32 shows the pitch, energy, and centroid waveform for a sample violin tone, along with the amplitude waveforms for the first six partials. (Note that the ordinates have been scaled so that each waveform occupies approximately the same space on the page.) Figures 33-35 show the various vibrato/tremolo features accumulated from chromatic scales played by the example instruments.

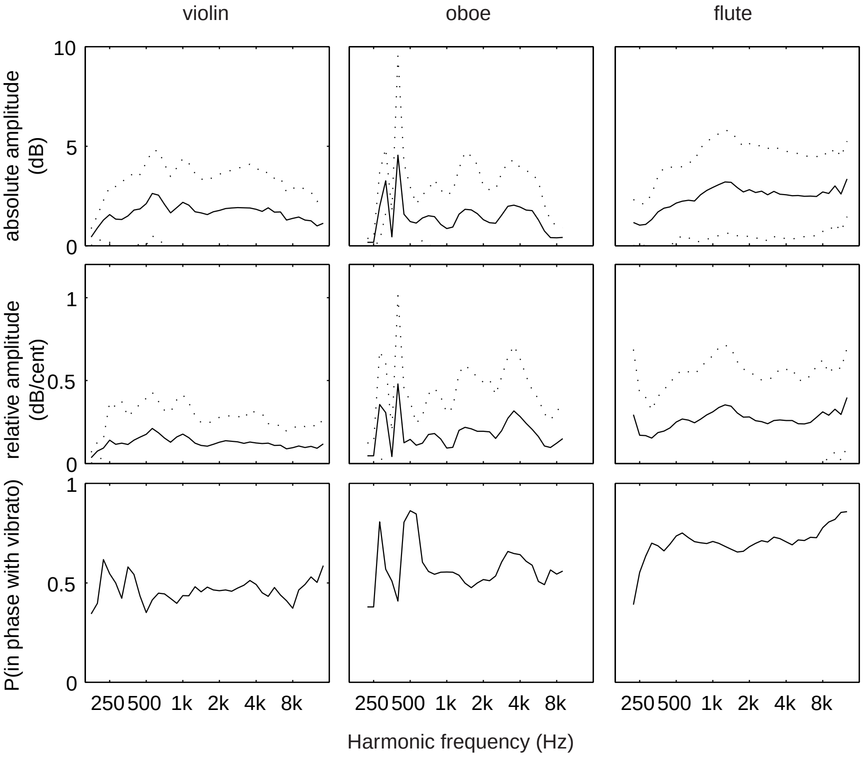

FIGURE 33. *The effect of vibrato on the harmonic partials*, mesaured by amplitude modulation strength as a function of partial frequency. Data for trumpet, clarinet, and alto saxophone have been omitted because their chromatic scales were not performed with vibrato.

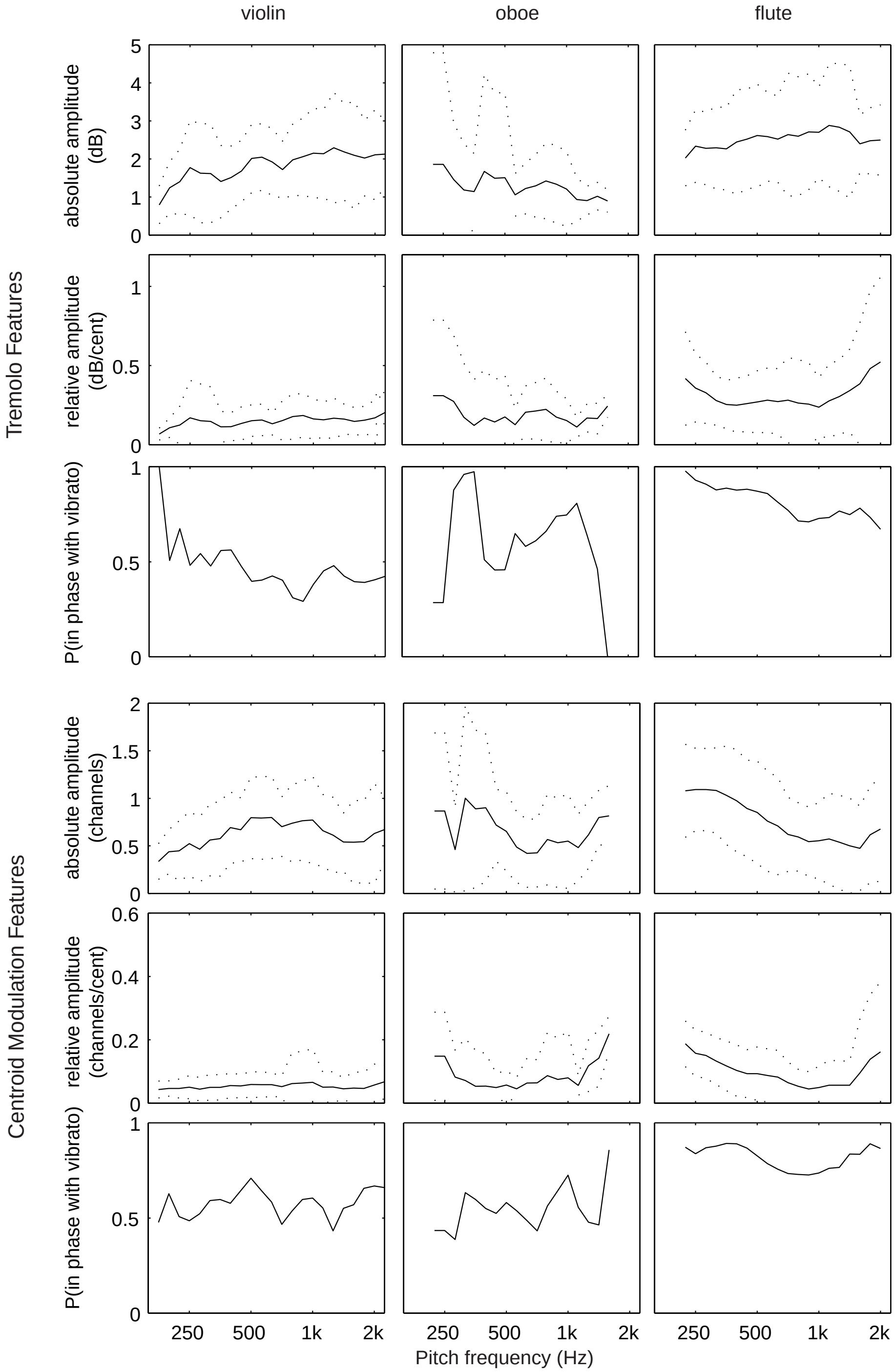

FIGURE 34. *The effect of vibrato on the overall energy and spectral centroid*. Data for trumpet, clarinet, and alto saxophone have been omitted because their chromatic scales were not performed with vibrato.

4.4.3 Attack transient properties

It is evident from the available human perceptual data (see Chapter 3) that the attack transient of an isolated musical tone played on an orchestral instrument contains crucial information for identifying the particular instrument that generated the tone. It is not clear, however, which aspects of the attack transient provide the essential information. Indeed, it is not even clear how to define when the "transient" ends and the "steady-state" begins. The literature is at best equivocal on these issues. It has been suggested that the relative onset times of the harmonic partials are important features, as are their attack rates (perhaps measured in dB/ms). Little has been written, however, about how to measure these properties from recordings of real instruments, and I am aware of no published descriptions of techniques for measuring these properties from recordings made in reverberant environments such as concert halls.

The techniques described here are necessarily tentative. They were inspired by visual inspection of the weft representations of tones from the McGill University Master Samples collection (Opolko & Wapnick, 1987), and they work reasonably well on the very cleanly recorded tones in that collection. The techniques have not, however, been adequately tested on a broad data set. I include them here because they may serve as a useful starting point for other researchers who might replace them with better techniques.

The signal-processing techniques underlying the attack-transient characterization performed here were inspired by methods for visual edge detection (Marr, 1982). The insight is that a sharp rise in acoustic energy corresponding to an attack or onset is analogous to a change in light intensity corresponding to an edge in an image. The algorithm begins by measuring the slope and curvature of an energy signal expressed as a function of time. These are computed using the *surfboard* technique (Schloss, 1985), which fits a regression line to local segments of the signal using a minimum mean-square error criterion, and records its slope as a function of time. This operation is less susceptible to noise than approximations based on simple differences (Schloss, 1985). After the slope is computed, the technique is reapplied to compute the curvature. Estimates are calculated using seven different regression-line lengths (or *scales*), ranging exponentially from 5 ms to 250 ms. The short windows are suitable for characterizing very rapid changes, the long windows for slower changes.

When the slope and curvature estimates are complete, the system identifies neighboring (in scale) zero crossings (of the curvature) that correspond to positive slopes. The positions of these zero crossings correspond to times at which the local energy rise rate is at a maximum. For percussive sounds, these times correspond very closely to the *perceptual attack time* of the sound (Gordon, 1984).

Each set of adjacent zero-crossings is termed a *rise*. The slope curve (at the appropriate scale) is examined at the time of each rise, and the time range surrounding the rise-time for which the slope is greater than 50% of the slope at the rise-time is called the *rise region*. A regression line is fit to the energy signal in this region, and its slope and total energy change is noted along with the time index of the rise. When rises occur in close temporal proximity across a range of cochlear filter channels, their average time index is termed the *attack time*.

The effect of these manipulations is to fit simple linear segments to the energy curve in the regions where the energy level is increasing substantially. The complexity of the algorithm seems to be necessary to make reliable measurements of changes that occur on different time scales (for example, a plucked string may reach full amplitude in 5 ms, whereas a bowed string might require 500 ms—a difference of two orders of magnitude).

Four time windows (0-50 ms, 50-100 ms, 100-200 ms, and 200-400 ms) are examined for additional rises after each attack. The probability of a rise occurring (the "rise likelihood") is estimated for each filter channel and each time window by pooling over all attacks. The motivation for this measurement comes from the observation that, for example, energy in partials produced by bowed-string instruments rises irregularly in both time and frequency, but energy in partials produced by brass instruments rises more predictably (earlier at low frequencies, later at high frequencies).

Finally, the *relative onset time* is computed for each partial by selecting the last-occurring (within the 200 ms window) rise from the appropriate filter channel, calculating the time index at which the regression line reaches within 3 dB of its maximum, and subtracting the attack time. The mean and standard deviation of the relative onset time is estimated for each filter channel by pooling over all attacks.

It is to be stressed that these techniques are tentative. Attack-transient characterization has received frustratingly little attention in the acoustics and synthesis literature. This has the potential to be a fertile area for future research.

4.5 The model hierarchy

The recognition system's knowledge base is a taxonomic hierarchy of source models of the type described in Section 4.4. In the current implementation, the taxonomy is specified in advance, rather than being acquired during training. Figure 35 shows an example taxonomy. The taxonomy has three levels. At the topmost level is a single category, labeled "All instruments." At the lowest level are the individual instrument classes. At the middle level, the instruments are assembled into family groups based on their common excitation and resonance structures. Thus, the pizzicato (plucked) strings are separated from the bowed strings, and the muted brass instruments are separated from the non-muted brass instruments. The woodwinds are divided into the flute, clarinet, double-reed and saxophone subgroups, in accordance with the discussion in Section 3.3.3 on page 58.

In the experiments performed in Chapter 6, each training sample is labeled with the name of the appropriate bottom-level (leaf) node of the taxonomy. During training, feature values are accumulated (as described in Section 4.4) at the appropriate leaf node and *at all of its ancestors*. By this method, the *double-reed* node, for example, accumulates feature data from all *oboe*, *English horn*, *bassoon,* and *contrabassoon* samples.

Alternately, it is possible to train only the leaf nodes and then to combine their accumulated feature distributions appropriately to train the more abstract nodes of the taxonomy. This method can be used to facilitate the comparison of many different taxonomies.

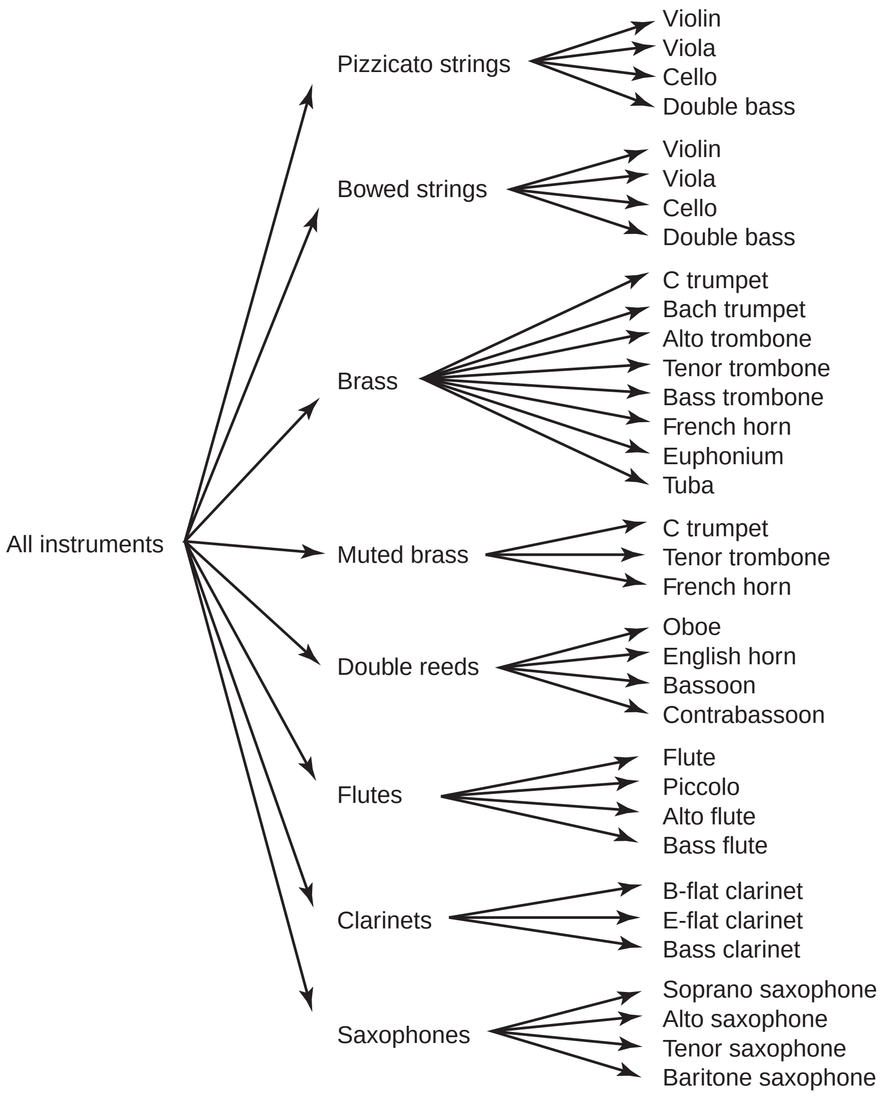

FIGURE 35. Taxonomy used in Computer experiment #3 (Section 6.5) to test the recognition system.

CHAPTER 5 Recognition

As described in Chapter 1, recognition is a process of gathering information about an object in the environment so as to be able to predict or more reliably infer its behavior or properties. Recognition was described as a process of categorization at multiple levels of abstraction, typically beginning at some intermediate level and becoming more specific (or general) according to the needs of the perceiver. Chapter 4 showed how an audio signal could be transformed through a series of representations into a high-level sound-source model. In this chapter, methods for categorization using sound-source models as *prototypes* are developed, and a computational model of the recognition process is presented.

5.1 Overview and goals

The recognition framework described here is an amalgam of several different techniques, with conceptual ties to taxonomic Bayesian belief networks (Pearl, 1988), decision trees (Breiman et al., 1984), spreading activation (Maes, 1989), and traditional search (Winston, 1992). This mélange is the result of an attempt to satisfy a conflicting set of desiderata, derived in part from the evaluation criteria described in Section 2.2:

- **Robustness**: A system based on the framework should perform well on classification and identification tasks, exhibiting generalization and handling real-world complexity. It should be able to classify new examples of any particular class reliably, given sufficient exposure to other sound sources belonging to that class. This performance should degrade gracefully as the available sensory data degrades.

- **Extensibility**: The framework should scale well from small to large sets of object classes; adding new classes should not drastically alter the system’s performance. It should also be possible to add new features to an object or class description, or to add new discrimination functions for classification, and these additions should improve the systems level of performance.
- **Flexibility**: The framework should not be dependent upon features that may not always be available. Its performance on classification tasks should degrade gracefully as features are removed from consideration, or as the quality of feature measurements decreases (the *flexibility* criterion overlaps the *robustness* criterion somewhat). The quality (and specificity) of classification should vary directly with the effort expended by the system. If only a rough classification at an abstract level is needed, then less effort should be required than would be for a more detailed classification.
- **Consistency**: The same basic algorithm should work well for a very brief exposure to a sound source (e.g., a single musical tone produced by an instrument), for extended exposure (an entire cadenza), and for the continuum between the two extremes. Presumably, performance on classification and identification tasks will improve as the degree of exposure increases.

The algorithm developed here is based on a taxonomic hierarchy of sound-source classes. There is a substantial literature on tree-based classification algorithms, but unfortunately there are as yet no deep theorems proving their optimality or competence (Ripley, 1996). There are, however, several justifications for their use. For example, there is evidence from psychology that humans use hierarchical structures during the recognition process (Rosch, 1978; Rosch et al., 1976). Hierarchies are often good models for the structure of the world (Bobick & Richards, 1986; Bobick, 1987), and hierarchical methods can make better use of sparse training data than their non-hierarchical counterparts (McCallum et al., 1998). If during the recognition process, the perceiver can rule out, or prune, branches of the hierarchy, the classes represented by those branches need never be considered directly, and this can provide immense computational savings over non-hierarchical methods; a system with fixed computing power can indirectly consider a much larger set of possibilities than it could consider directly.

Like a decision-tree classifier, the algorithm described here begins at the root node of a tree—at the top of the taxonomy—and makes decisions, traversing from node to node as the classification is performed. There are, however, several critical improvements that distinguish it from traditional decision trees (Breiman et al., 1984). In a decision tree, only the leaf nodes are usually interpretable in terms of coherent object classes, whereas each node of the taxonomic hierarchy used here represents a meaningful grouping. In a decision tree, the process of choosing one child node over another from a particular parent node is usually all-or-nothing, and the decision is usually based on a single feature. Further, the set of features used at each decision node is specified in advance (usually the features are chosen during a training process). In contrast, the algorithm used here is improvisational. It decides on the fly which features to use, based on the current context, and it can be configured to employ a range of behaviors from greedy all-or-nothing decisions to exploring the entire tree and testing every leaf node. The main drawback is that, in the current implementation, the taxonomy must be prespecified rather than learned from training data. In contrast, decision trees are typically learned rather than prespecified.

5.2 Definitions and basic principles

In this section, we will ignore the taxonomy and first consider non-hierarchical classification. To begin, we define a *categorization* as a set of non-overlapping categories that partitions a set of sound sources into non-overlapping groups. Each category has a single *prototype*, consisting of a sound-source model as described in Section 4.5. (In general, each category could have multiple prototypes with only minor extensions to the algorithms described here.) The goal of the categorization process is to determine which category an unlabeled sound source belongs to, based on measurements of its acoustic features.

The category prototype can be viewed as a generative probabilistic model for the features of sounds produced by sound sources in that category. As described in Chapter 4, the prototype is a *frame* with a *slot* for each feature. Most of the features are assumed to arise from Gaussian processes, and each slot contains the mean and variance of feature values observed from sound sources of the appropriate category. In addition, each slot has an associated *comparison method*, which is used to make probabilistic comparisons between models. In general, if we are given a model corresponding to an unlabeled sound source, the comparison methods of each category prototype will calculate the log likelihood that the feature values observed from the unlabeled sound source could have arisen from each category. This is accomplished by using Bayes' rule to invert the probabilistic models, making the *naive Bayes* assumption that each feature is independent of every other feature and of the feature's context given the category identity. Despite the fact that the independence assumption is strongly violated by the actual data, the naive Bayes technique is very flexible and works well in many situations (McCallum et al., 1998). The rest of this section describes the probabilistic basis of the algorithm in more detail.

Consider a set of *N* categories and a single feature measurement from an unlabeled sound source *M*. Each category has its own prototype, consisting of the mean and standard deviation of a normally distributed feature. The probability of observing a particular value f_0 of feature *f*, given that it is observed from a member of category *n*, is given by

$$P(f_j = f_0 | C_n) = \frac{1}{\sqrt{2\pi\sigma_{n,j}^2}} \exp\left\{-\frac{1}{2}\frac{(f_0 - m_{n,j})^2}{\sigma_{n,j}^2}\right\}, \qquad \textbf{(18)}$$

where $m_{n,j}$ and $\sigma_{n,j}$ are the mean and standard deviation for the feature f_j, given membership in category *n*. We use Bayes' rule to invert this expression, yielding the likelihood of membership in each category, given the feature observation:

$$P(C_n|f_j = f_0) = \frac{P(f_j = f_0|C_n)P(C_n)}{\sum_{k=1}^{N} P(f_j = f_0|C_k)P(C_k)}. \qquad (19)$$

The denominator in Equation 19 is a normalizing factor that does not change from category to category, and is thus often ignored in practice. When more than one feature value is observed, their values are assumed to be independent of each other (the naive Bayes assumption), and the likelihood values simply multiply. Defining

$$\lambda_{n,j} = P(f_j|C_n), \qquad (20)$$

the likelihood of class *n* (ignoring the normalizing factor) is given by

$$\lambda_n = P(C_n)\prod_j \lambda_{n,j}. \qquad (21)$$

Because the product of likelihoods often results in very small values of λ_n, it is more numerically stable to compute these values using logarithms:

$$l_{n,j} = \log P(C_n|f_j) \qquad (22)$$

and

$$l_n = \sum_j l_{n,j} + \log P(C_n). \qquad (23)$$

Once the likelihood has been calculated for each category, the *maximum a posteriori* estimate of the unlabeled source's category is simply the category with the largest probability value. The current implementation of the system described in this chapter assumes that all categories are equally likely *a priori*, so the $P(C_n)$ terms are ignored, and the result is the *maximum likelihood* estimate of category membership.

5.3 Taxonomic classification

Now consider a taxonomic hierarchy, as illustrated in Figure 36. We define the structure to be a tree with a single root node, labeled *A* in the figure. A node may have any number of immediate *descendents*, or *children*. In the figure, node *A* has three children, labeled B_1, B_2, and B_3. If a node has more than one child, it is called a *decision node*. If it has none, it is called a *leaf node*. In this formulation, each node represents a category. Node *A* represents the category that contains all sound sources. Nodes B_1, B_2, and B_3, represent a partitioning of the sound sources represented by node *A* into three categories. Let the area labeled "Level 1" be a *categorization* in the sense defined in Section 5.2. Each category may be further subdivided into additional subcategories. In the figure, each of the categories B_1, B_2, and B_3 is divided into two subcategories, labeled C_1 - C_6, which also make up a categorization, labeled "Level 2." The division into subcategories may continue indefinitely, or until each category contains only a single sound source. Within this framework, each *level* of the taxonomy represents a different level of abstraction. In the figure, "Level 1" is more abstract than "Level 2."

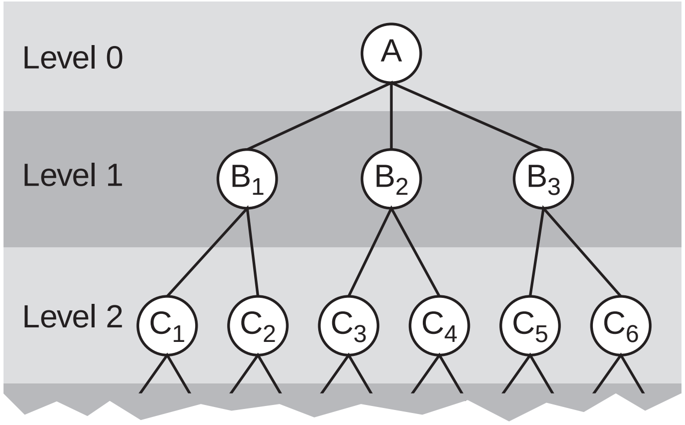

FIGURE 36. An example of a taxonomic hierarchy. Each node represents a class or category; each level is a *categorization*. Other properties are discussed in the text.

The recognition process starts at the root of the taxonomy and makes a maximum likelihood decision at each node (as described in Section 5.2), recursively stepping through decreasingly abstract levels of the hierarchy. Several variations of this algorithm have been implemented, each with different strengths and weaknesses. In Chapter 6, several of these possibilities are tested and their performance levels evaluated. The best approach in a particular scenario depends on many factors, which will be discussed in Chapters 6 and 7.

5.3.1 Extension #1: Context-dependent feature selection

One of the biggest hurdles in constructing a successful pattern-recognition system is dealing with insufficient training data. As the number of parameters in a probabilistic model increase, so does the amount of training data required to estimate the parameter values; with a fixed amount of training data, additional features can cause the performance of a classifier to decrease. This is commonly known as the *curse of dimensionality* or the *bias-variance tradeoff* (Therrien, 1989). One approach to alleviating the difficulties associated with using a large number of features is to select a small subset of the features that is best suited to the task at hand. In a taxonomic classifier, usually there are a small number of child categories to decide among at any particular node, and intuitively, we expect that the best feature subset for classification will be different for each set of categories. For example, although pitch range may be an excellent feature for distinguishing between violin and double bass, it is not as good for distinguishing between violin and viola.

In the system constructed for this dissertation, several approaches for feature selection were tested. First, we observe that the category prototypes of the children of a node can be used directly to estimate the *salience* of each feature. The *discriminability* of two normal distributions with the same variance is given by

$$d' = \frac{m_1 - m_2}{\sigma}, \tag{24}$$

where m_1 and m_2 are the means for the two distributions, and σ is their common standard deviation. The probability of error of a maximum-likelihood estimator based on a single normal feature is monotonically related to the *inverse* of d', so a the d' value of a feature increases, so does its usefulness as a feature for classification.

There are several different ways of calculating analogous measures when the variances are not equal, such as the *mutual information* or the *Bhattacharya distance* (Therrien, 1989), but a simpler approach was taken here, using the average variance

$$d'' = \frac{m_1 - m_2}{\sqrt{(\sigma_1^2 + \sigma_2^2)/2}}. \tag{25}$$

d'' is taken to be the *discriminating power* of a feature in the context of two categories. At each node in the hierarchy, the discriminating power of each feature for each pair of child categories is computed and stored at the node.

A second observation is that as sounds are heard from a sound source whose category is not known, some features may not be available at all, and some may be measured more reliably than others. Intuitively, the system should favor the most discriminating features that have been most reliably measured. It makes no sense at all to think of making a decision based on only default feature values for which there is no supporting evidence.

From these two observations, several algorithmic variations are possible. As described above, the system computes the discriminating power of each feature for each pair of nodes under consideration. These numbers are averaged, and the result is taken to be the *salience* of the feature in the current context. Further, a *reliability estimate*—a number between 0 and 1—is computed for each feature based only on the model created for the sound source being recognized. The features can then be ordered by the salience estimate, the reliability estimate, or the product of the two. The features with the highest scores are the most likely to be good discriminators, given the current *context* (defined to be the set of categories currently under consideration). The system can then either choose some subset of the features or can use all of the features, weighted by their discriminating power. In the current implementation, this second option is accomplished by multiplying the log likelihoods returned from the comparison methods by the salience, the estimated reliability, or both. These *ad hoc* computations have the effect of exponentiating the likelihood estimates and are not based on theoretical motivation. Their practical usefulness will be evaluated in Chapter 6.

5.3.2 Extension #2: Rule-one-out

If some sort of feature selection is used in the system then the calculations of feature salience may depend strongly on the particular set of categories under consideration. When more than two categories are being considered, it may be possible to do better than just choosing a set of features and computing a maximum-likelihood estimate of category membership. The algorithm adopted here is a *rule-one-out* strategy. Given a set of *k* categories, the system identifies the most salient features, computes the likelihood scores, and removes from consideration the category with the least likelihood. After a category is ruled out in this manner, the feature salience scores are recomputed in light of the new context and the algorithm repeats. With this strategy, the classifier shifts the features during the process, always choosing a suitable subset for the current context.

5.3.3 Extension #3: Beam search

One of the most significant drawbacks of hierarchical classifiers is that an error made at one level cannot be fixed by choices made later at more specific levels. If the classifier always chooses the maximum-likelihood category at each node, the probability of a correct classification is equal to the product of correct-classification probabilities at each node from the root of the tree to the leaf node corresponding to the "correct" classification. This can be a serious problem because the prototypes for the most abstract classes are necessarily the most vague, if only because they comprise many different subcategories.

To deal with this problem, the system has been equipped with a *beam search* algorithm (Winston, 1992), which expands the best *b* nodes at each level of the hierarchy until the leaf nodes are reached, at which time a final maximum-likelihood decision is made. This alleviates the error-compounding problem of the greedy branch-selection case. The *beam width*, *b*, can be varied from one (greedy branch-selection) to infinity (full search of the entire tree), trading classification performance for computational expense. If the maximum-likelihood decisions made at the most abstract nodes of the hierarchy are generally reliable, a beam width value of two or three is a reasonable compromise.

5.4 Strengths of the approach

It is worth reflecting upon how well the algorithm described above is likely to satisfy the desiderata listed at the beginning of this chapter. Unlike many pattern-recognition techniques, the algorithm does not depend on a fixed set of features for classification. Rather, it uses whatever information is available to make the best decision it can. For this reason, the algorithm's performance will degrade gracefully—rather than failing altogether—when particular features are not available. The results of the experiments in Chapter 6 will show that, given a suitable set of features and comparison functions, the system generalizes from a small number of training examples and can robustly classify previously unheard examples from learned sound-source classes. This satisfies the *robustness* criterion.

Systems based on the algorithm described above can easily be augmented with new features and new sound-source classes. Adding a new feature requires only the addition of a new slot (with its corresponding comparison function) to the description of each affected sound-source class. This can take place after the system has already been trained, and default slot values could—with only minor extensions to the algorithm—be gradually replaced by sensory data as sounds are recognized using other features. Adding a new sound-source class requires only the creation of a class prototype and the introduction of appropriate links to parent and child nodes. Again, this can take place after the system has already been trained; the only overhead is that pair-wise feature-salience ratings will have to calculated at the parent node, and some may have to be re-calculated at its ancestors. This satisfies the *extensibility* criterion.

Because the classification process operates in stages, traversing decision nodes from the abstract to the specific, the algorithm scales well to very large numbers of sound-source classes, as long as reliable classifications can be made at each node. Consider, for example, a system with N “leaf” nodes, representing the partitioning of all sound sources into the most specific set of classes that could be needed in the recognition process. If N is very large, it would be prohibitively expensive to directly compare an unlabeled sound-source model to every prototype in order to make a classification decision. If, however, each node of the hierarchy has k children (on average), the greedy branch-selection algorithm requires that only $k\log_k N$ comparisons be made—a huge savings for large values of N. The main drawback is that classification errors in the greedy algorithm compound; these are addressed by the beam search algorithm, which trades classification performance for computational expense. The multi-stage classification process is particularly advantageous, however, if fine-grained categorization is not always necessary. Often, categorization at a more abstract level suffices for the task at hand, and in such cases, even fewer than $k\log_k N$ comparisons need be made. The degree of effort required to make a decision is directly related to the logarithm of the number of categories that must be considered. This and the ability to choose appropriate features for a given context satisfies the *flexibility* criterion.

The final desideratum, *consistency*, is dependent on the implementation of the feature comparison methods and their ability to estimate their expected utility, given a particular set of sensory data. If the comparison functions are able to accurately gauge their ability to discriminate among the children of a node, the system will automatically choose the best features to make each particular decision, given the information at hand.

5.5 An example of the recognition process

To illustrate the basic recognition algorithm and the effects of some of the extensions described in the previous section, consider the simplified musical-instrument taxonomy shown in Figure 37. In the example, the classifier is configured to use context-dependent feature selection based on the average *discriminating power* given the *current context*, with the rule-one-out and beam search extensions (the beam width is set to two) described in Section 5.3.

When a new, unclassified, recording of a sound source is presented to the system, feature measurements are assembled into a frame representation. For example, an isolated tone produced by playing a violin with vibrato might give rise to the set of feature slots shown in Table 4.

5.5.1 Step one

The recognition process begins at the "All instruments" node. The current context consists of the node's children: the bowed string, brass, and double-reed groups, as shown in Figure 38. The model computes discriminating-power measurements based on the stored prototypes for these categories. The features with d'' values greater than one are shown in Table 5.

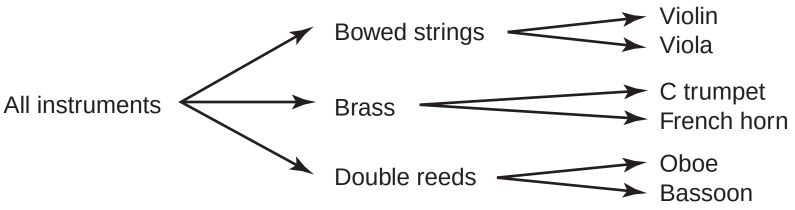

FIGURE 37. The simplified taxonomy used to illustrate the recognition process.

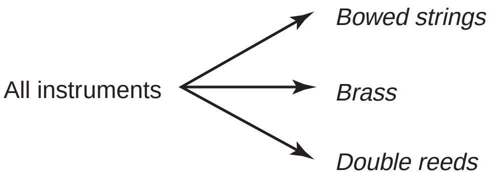

FIGURE 38. The current context at the beginning of Step one. The categories under consideration are Bowed strings, Brass, and Double reeds (shown in italics).

Feature	Details
Pitch range	1 measurement (the note falls in one 1/6-octave band)
Spectral centroid	1 measurement (at the pitch frequency)
Average relative spectrum by harmonic number	6 measurements
High-frequency rolloff rate	1 measurement (at the pitch frequency)
High-frequency cutoff frequency	1 measurement (at the pitch frequency)
Spectral irregularity	5 measurements (in various frequency bands)
Number of zeros	1 measurement (at the pitch frequency)
Tremolo: absolute and relative strength and phase	1 measurement each (at the pitch frequency)
Centroid modulation: absolute and relative strength and phase	1 measurement each (at the pitch frequency)
Individual harmonic AM: absolute and relative strength and phase	6 measurements each (at frequencies of first six partials)
Relative onset time by partial frequency	6 measurements (at frequencies of first six partials)
“Rise likelihood” by frequency and post-onset time	30 measurements (in 10 frequency bands and 3 post-onset time windows)

TABLE 4. Features measured from an example violin tone.

Feature	Average d''	Number of measurements chosen
Relative onset time by partial frequency	3.514	5
Centroid modulation (relative phase)	2.395	1
Spectral irregularity	1.378	4
Individual harmonic modulation (relative phase)	1.208	4
Tremolo (relative strength)	1.177	1
Tremolo (relative phase)	1.175	1
Individual harmonic modulation (relative strength)	1.166	3

TABLE 5. Features with d'' values greater than one, given the current context at Step one.

Considering that the stimulus is an isolated tone, it is not surprising that the most salient features are related to the tone’s attack (relative onset time by partial frequency), and to vibrato (centroid and tremolo features). As discussed in Chapter 3, bowed string attacks are much slower than brass or double-reed attacks. Also, bowed-string instruments have much more complicated resonance structures than the brass and double-reed instruments, and vibrato highlights this difference.

Log likelihood values for the three categories are computed based on the features (weighted by the d'' values). The double-reed category has the smallest likelihood value and is ruled out.

5.5.2 Step two

At the beginning of Step two, there are two categories under consideration, the bowed strings and brass groups. Because the beam width is set to two, these categories are expanded, and their children become the new context. The current context therefore consists of the violin, viola, C trumpet, and French horn groups, as shown in Figure 39. Features with d'' values greater than one are shown in Table 6.

As is evident from a comparison of Tables 5 and 6, the relative salience of the various features has shifted considerably. The average discriminating power of the relative onset time by partial frequency has been cut in half (but is still salient), and the spectral centroid has become very salient, as has the spectral irregularity (also evidenced by the number of zeros). As suggested by the discussion in Chapter 3, the violin and viola have much more irregular spectra and longer attacks than the C trumpet and French

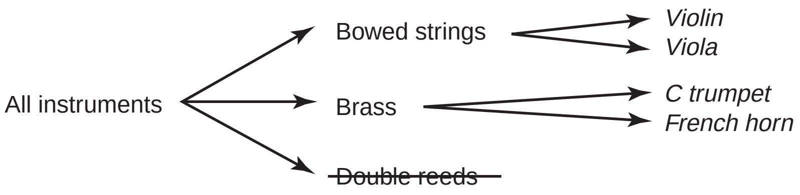

FIGURE 39. The current context at the beginning of Step two. The categories under consideration are Violin, Viola, C trumpet, and French horn (shown in italics).

Feature	Average d''	Number of measurements chosen
Centroid modulation (relative phase)	4.156	1
Centroid modulation (absolute strength)	3.992	1
Spectral centroid	3.352	3
Spectral irregularity	3.070	4
Number of zeros	2.505	1
Relative onset time by partial frequency	1.615	2
Centroid modulation (relative strength)	1.541	1
Individual harmonic modulation (relative strength)	1.385	4
Tremolo (relative strength)	1.042	1
"Rise likelihood" by frequency and post-onset time	1.021	2
Average relative spectrum by harmonic number	1.008	1

TABLE 6. Features with d'' values greater than one, given the current context at Step two.

horn. The brass instruments tend to sound "brighter" than the string instruments and thus have higher spectral centroid measurements.

Log likelihood values for the four categories are computed based on the features (weighted by the d'' values). C trumpet has the smallest likelihood value, and it is ruled out.

5.5.3 Step three

At the beginning of Step three, the current context consists of the violin, viola, and French horn groups, as shown in Figure 40. Features with d'' values greater than one are shown in Table 7.

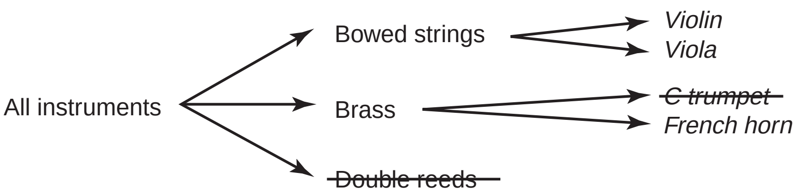

FIGURE 40. The current context at the beginning of Step three. The categories under consideration are Violin, Viola, and French horn (shown in italics).

Feature	Average d''	Number of measurements chosen
Spectral irregularity	1.993	4
Spectral centroid	1.320	1

TABLE 7. Features with d'' values greater than one, given the current context at Step three.

At Step three, only two features remain salient (and their discriminating power is somewhat reduced). Likelihood values for the three categories are computed based on the features (weighted by the d'' values). French horn has the smallest likelihood value, and it is ruled out, thereby ruling out the brass category as well.

5.5.4 Step four

At the beginning of Step four, the current context consists of the violin and viola groups, as shown in Figure 41. None of the features have d'' values greater than one (the largest is 0.267, see Table 8), highlighting the difficulty of discriminating a violin from a viola based on only one isolated tone. The system computes the likelihood values for the two categories (weighted by the d'' values). Viola has the smaller likelihood and is ruled out. The sample is correctly classified as violin.

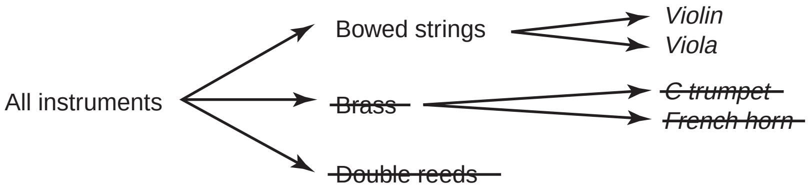

FIGURE 41. The current context at the beginning of Step four. The categories under consideration are Violin and Viola (shown in italics).

Feature	Average d''	Number of measurements chosen
Tremolo (relative strength)	0.267	1
Individual harmonic modulation (relative strength)	0.200	4
Centroid modulation (relative strength)	0.177	1
Spectral irregularity	0.169	4
Individual harmonic modulation (absolute strength)	0.161	4
Spectral centroid	0.155	1
Number of zeros	0.134	1

TABLE 8. Features with the largest values of d'', given the current context at Step four.

CHAPTER 6 Evaluation

Chapters 4 and 5 presented the components of a sound-source recognition system tailored to the recognition of solo monophonic orchestral musical instruments. In this chapter, the system is tested on a variety of classification tasks, and its performance is juxtaposed with that of human listeners and of other artificial systems.

The chapter has six sections. First, the sets of recordings used to train the recognition system and to test both the system and human experimental subjects are described. Second a human listening experiment designed to evaluate human abilities at recognizing musical instruments is described. The next three sections describe three experiments that test the recognition system under various conditions. Finally, the results are related to previous research in musical instrument recognition by both humans and machines.

6.1 A database of solo orchestral instrument recordings

Recordings for use during the evaluation process were obtained from three sources: a commercial sample library, a number of commercial compact discs, and a small set of recordings made especially for this project. An effort was made to collect solo recordings of the 27 orchestral instruments in Table 9. Whenever possible, multiple, independent recordings of performances by different artists were gathered. In all, more than 1500 isolated tones and more than 2 ½ hours of musical performance were assembled. All recordings were re-sampled to 32 kHz using professional-quality software before presentation to either human or machine.

Violin	Bassoon	Trumpet
Viola	Contrabassoon	Cornet
Cello	B-flat clarinet	Fluegel horn
Double bass	E-flat clarinet	French horn
Flute	Bass clarinet	Alto trombone
Alto flute	Soprano saxophone	Tenor Trombone
Piccolo	Alto saxophone	Bass trombone
Oboe	Tenor saxophone	Euphonium
English horn	Baritone saxophone	Tuba

TABLE 9. The 27 orchestral instruments considered in this study.

The first source of recordings was the McGill University Master Samples (MUMS) collection (Opolko & Wapnick, 1987). The collection consists of a series of chromatic scales performed on a variety of musical instruments (over most of their playing ranges) by professional musicians in a recording studio. According to the producers, careful attention was paid to making recordings that were “maximally representative” of the various instruments. For the studies presented in this chapter, a subset of the collection was used, consisting of chromatic scales by the instruments shown in Table 10.

The second source of recordings was the MIT Music Library’s compact disc collection (and a test CD produced by the European Broadcast Union). As is evident from Table 11, which details the number of independent recordings and total duration of the samples acquired for each instrument, it was much easier to find solo recordings of some instruments than others. The recording quality varies greatly from sample to sample, ranging from recordings made in modern studios to decades-old recordings made in highly reverberant concert halls with high levels of ambient noise.

To augment the collection of recordings described above, several student performers were hired from within the MIT community. Samples were recorded directly to DAT (at 48 kHz) in a converted studio control room, using a high-quality cardioid microphone placed approximately 1 meter in front of the performer. These recordings are also catalogued in Table 11.

Instrument	Notes
Violin	4 scales: bowed w/vibrato, muted, martele, pizzicato
Viola	(see violin)
Cello	(see violin)
Bass	(see violin)
Flute	2 scales: normal and flutter-tongued
Alto flute	1 scale
Piccolo	2 scales: normal and flutter-tongued
Oboe	1 scale
English horn	1 scale
Bassoon	1 scale
Contrabassoon	1 scale
B-flat clarinet	1 scale
E-flat clarinet	1 scale
Bass clarinet	1 scale
Soprano saxophone	1 scale (partial range only)
Alto saxophone	1 scale (partial range only)
Tenor saxophone	1 scale (partial range only)
Baritone saxophone	1 scale (partial range only)
C trumpet	2 scales: normal, and with harmon mute (stem out)
Bach trumpet	1 scale
French horn	2 scales: normal and (hand) muted
Alto trombone	1 scale
Tenor trombone	2 scales: normal and (straight) muted
Bass trombone	1 scale
Tuba	1 scale

TABLE 10. Description of the MUMS samples (isolated tones) used in this study. Each sample consists of a chromatic scale performed by a professional musician in a recording studio.

Instrument	Total duration	Number of performers (professional/ student)	Notes
Alto trombone	300 s	1/0	
Bassoon	406 s	2/1	39 s (authentic/period instrument); 14 s; 353 s
Bass clarinet	9 s	1/0	
B-flat/A clarinet	1242 s	5/1	323 s; 139s; 300 s; 300 s; 15 s; 165 s
Cello	627 s	2/1	128 s; 33 s; 466 s
Double bass	31 s	1/0	
English horn	190 s	2/0	181 s; 9 s
Euphonium	688 s	0/1	
Flute	2147 s	7/1	669 s; 439 s; 35 s; 31 s; 300 s; 300 s; 19 s; 354 s
French horn	382 s	2/1	250 s; 115 s; 17 s
Oboe	460 s	2/1	53 s (authentic/period instrument); 21 s; 386 s
Piccolo	7 s	1/0	
Saxophone (type not known)	14 s	1/0	
Soprano saxophone	183 s	1/0	
C Trumpet	454 s	2/2	64 s; 13 s; 224 s; 153 s
Tenor trombone	299 s	2/0	289 s; 10 s
Tuba	19 s	1/0	
Viola	452 s	3/1	55 s; 200 s; 24 s; 173 s
Violin	1451 s	5/1	572 s; 9 s; 300 s; 129 s; 30 s; 501 s

TABLE 11. Description of the recordings assembled from compact discs and from student performers. The student recordings were made in the control room of a recording studio (a space with very little reverberation); the professional recordings vary greatly in the levels of ambient reverberation and noise. Source material ranged from classical repertoire to 20th century art music and jazz.

6.2 Testing human abilities

Although the experiments described in Section 3.1 reveal some of the quirks and qualities of human instrument-recognition abilities, none of them employed a wide range of natural stimuli. Only Kendall (1986) used melodic phrases, and his stimuli were played on only three different instruments, each from a different family. The results cited from the isolated-tone studies are difficult to interpret, in part because of variations in experimental procedure (e.g., free-response versus forced-choice) and range of stimuli. In order to fairly compare the performance of an artificial system with that of human listeners, it is necessary to test human subjects with experimental protocols equivalent to those used to test the artificial system.

6.2.1 Experimental method

This section describes the method used in an experiment designed to test the ability of expert human listeners to recognize musical instruments. The experiment was divided into two components. Like nearly all of the previous musical instrument recognition experiments, the first component employed single isolated musical tones as stimuli. The second component employed more ecologically relevant stimuli consisting of ten second fragments of solo musical performances.

Fourteen human subjects participated in the experiment. Each had substantial previous exposure to the instruments of the orchestra. At the time of the experiment, subjects 1-9 were currently practicing an orchestral instrument or performing with orchestral ensembles (subjects 8 and 9 were vocalists). Subjects 10-11 had previously played in orchestras, but not in the last five years. Subjects 12-13 had never played in an orchestra but had substantial experience listening to orchestral music. Subject 14 had never performed in an orchestra but had extensive experience as a recording engineer for professional orchestras.

The experimental sessions were automated using a computer program written especially for this task. The program presented the trials comprising each particular session in random order and recorded the subject’s responses in a data file. Stimuli were played back from compact discs (over headphones) under the control of the program. Each experimental session took place in a quiet room, free from interruption.

Every subject participated in two sessions, lasting approximately 30 minutes each. The first session tested the subjects’ classification abilities with isolated tones, the second with ten second segments of solo performance taken from commercial recordings or specially recorded for this experiment. Each session was divided into separate trials, with one recording (a tone or a solo segment) tested on each trial (137 isolated tones and 102 solo segments were tested). On each trial, the subject had the opportunity to listen to the test stimulus as many times as desired. The subject was subsequently required to choose a response from a list of 27 instrument names (reproduced in Table 9 on page 118). Each subject was informed that stimuli might not be evenly distributed among the 27 categories, and that he or she should use their best judgment on each trial individually rather than attempt to distribute responses uniformly. Prior to participation, each subject confirmed having prior exposure to each of the 27 instruments in the response list.

The recordings used in the first experiment were taken from the McGill University Master Samples collection (Opolko & Wapnick, 1987). Tones at ten different pitches were used, and the set of instruments varied from pitch to pitch (in large part because playing range varies from instrument to instrument, but also due to quirks of the available set of recordings). The collection of pitches and instruments is summarized in Table 12.

Pitch (Hz)	Number of tones	Instruments
64.7 Hz	8	cello (vibrato; muted with vibrato; pizzicato), double bass (v; m; p), bassoon, tuba
91.4 Hz	12	cello (v; m; p), double bass (v; m; p), bassoon, French horn (normal; muted), tenor trombone (normal; with mute), tuba
182.4 Hz	19	viola (v; m; p), cello (v; m; p), double bass (v; m; p), English horn, B-flat clarinet, bassoon, trumpet (normal; with Harmon mute), French horn (normal; muted), tenor trombone (normal; with mute), tuba
257.7 Hz	23	violin (v; m), viola (v; m; p), cello (v; m; p), double bass (v; m; p), flute, oboe, English horn, B-flat clarinet, bassoon, trumpet (normal; with Harmon mute), “Bach” trumpet, French horn (normal; muted), tenor trombone (normal; with mute), tuba
347.6 Hz	21	violin (v; m), viola (m; p), cello (v; m; p), flute, oboe, English horn, B-flat clarinet, trumpet (normal; with Harmon mute), “Bach” trumpet, French horn (normal; muted), tenor trombone (normal; with mute), tuba
440.0 Hz	21	violin (open string; muted open string; fingered with vibrato; muted with vibrato), viola (v), cello (v), flute, oboe, English horn, B-flat clarinet, trumpet (normal; with Harmon mute), “Bach” trumpet, French horn (normal; muted), tenor trombone (normal; with mute).
647.0 Hz	13	violin (os; mos; fv; mv), viola (v), cello (v), flute, piccolo, oboe, B-flat clarinet, trumpet (normal; with Harmon mute), “Bach” trumpet
979.0 Hz	11	violin (v; m), viola (v; m), flute, piccolo, oboe, B-flat clarinet, trumpet (normal; with Harmon mute), “Bach” trumpet
1383.0 Hz	5	violin (v; m), flute, oboe, “Bach” trumpet
2094.0 Hz	4	violin (v; m), flute, piccolo

TABLE 12. List of isolated tones used in the first experiment, arranged by pitch.

The recordings used in the second experiment were of eclectic origin, as described in Section 6.1. An attempt was made to present multiple recordings of each instrument, as played by different performers. It was difficult, however, to find examples of some instruments (or to find local performers willing to be recorded), so the number of recordings (and the number of independent performers) varies by instrument. In almost all cases, two recordings were used per performer. If a particular recording was longer than ten seconds, only the first ten-second segment was played for the subjects. Typically, the segment contained a melodic phrase, often taken from the cadenza of a concerto; a few segments consisted of major-triad arpeggios. All segments were intended to be typical of a musical style commonly performed on the particular instrument. Table 10 summarizes the number and sources of recordings used in the listening experiment.

Instrument	Total number of samples	Number of professional performers	Number of student performers
Alto trombone	2	1	0
Bassoon	6	2	1
Bass clarinet	1	1	0
B-flat clarinet	13	5	1
Cello	6	2	1
Double bass	2	1	0
English horn	4	2	0
Euphonium	2	0	1
Flute	15	7	1
French horn	6	2	1
Oboe	6	2	1
Piccolo	1	1	0
Saxophone (?)	2	1	0
Soprano Saxophone	2	1	0
Trumpet	8	2	2
Tenor trombone	4	2	0
Tuba	2	1	0
Viola	8	3	1
Violin	12	5	1

TABLE 13. Summary of the stimuli used in the second experiment.

Of the 27 instruments in the response list, eight instruments were absent altogether from the stimulus sets of both experiments (alto flute, E-flat clarinet, contrabassoon, cornet, fluegel horn, bass trombone, tenor saxophone[1], and baritone saxophone). Bass clarinet, alto trombone, euphonium, soprano saxophone, and

[1] Two of the samples, which came from a collection of short solo passages, were labeled only “saxophone.” I judged them to most likely have been played on an *alto* saxophone, and that was arbitrarily deemed to be the correct response. It turns out that 50% of the subjects judged it to be a *tenor* saxophone, and only 28.6% responded *alto*. Neither interpretation changes the overall results significantly.

alto saxophone were also absent from the isolated tone stimulus set used in the first experiment.

6.2.2 Results

A confusion matrix for the isolated tone stimuli, pooled across all subjects, is shown in Table 8. Table 16 summarizes the results by family. Pooling across all subjects in the isolated-tone condition, the exact instrument was selected on 45.9% of trials, and an instrument from the correct family on 91.7% of trials (a subject would score 3.7% and 20.2% on these statistics by guessing randomly). In this condition, a within-family error is 5.5 times more likely to occur than a between-family error. All of these results are strongly significant. For the full confusion matrix, pooled across all subjects, $\chi^2(13, 26) = 8837, p \ll 0.001$ (χ^2 values for individual subjects were all strongly significant using this test). Collapsed across instrument families (still pooled across all subjects), $\chi^2(4, 5) = 5334, p \ll 0.001$ (again, each individual subject result was strongly significant).

Six of the subjects were not able to reliably distinguish double-reed instruments from clarinets in the isolated tone condition. χ^2 tests using only trials on which a double-reed or clarinet instrument was presented or responded were insignificant for subjects 4, 7, 9, 11, 13, 14. Results for the other subjects ranged from significance levels of $p < 0.05$ to $p < 0.001$.

A confusion matrix for the ten-second excerpt stimuli, pooled across all subjects, is shown in Table 15. Table 17 summarizes the results by family. Pooling across all subjects in the ten-second excerpt condition, the exact instrument was selected on 66.9% of trials, and an instrument from the correct family on 96.9% of trials (a subject would score 3.7% and 18.1% on these statistics by guessing randomly). In this condition, a within-family error is 9.7 times more likely to occur than a between-family error. All of these results are strongly significant. For the full confusion matrix, pooled across all subjects, $\chi^2(18, 26) = 13236, p \ll 0.001$ (χ^2 values for individual subjects were all strongly significant using this test). Collapsed across instrument families (still pooled across all subjects), $\chi^2(5, 5) = 6477, p \ll 0.001$ (again, each individual subject result was strongly significant).

In the ten-second excerpt condition, only one subject (#13) could not reliably distinguish double-reed instruments from clarinets. χ^2 tests for all other subjects were significant at the ($p < 0.001$) level, except for subject #14 ($\chi^2(1, 1) = 8.6, p < 0.005$).

A summary of the results from both conditions is shown in Figure 42, along with results for a hypothetical “random guesser.” Performance pooled across all subjects is summarized in Table 18. Table 19 illustrates the overall performance for each individual instrument in the two conditions.

Responded Presented	Violin	Viola	Cello	Double bass	Flute	Alto flute	Piccolo	Oboe	English horn	Bassoon	Contrabassoon	B-flat clarinet	E-flat clarinet	Bass clarinet	Trumpet	Cornet	Fluegel horn	French horn	Alto trombone	Tenor trombone	Bass trombone	Euphonium	Tuba	Soprano saxophone	Alto saxophone	Tenor saxophone	Baritone saxophone
Violin	**47.6**	33.3	13.5		0.4	0.4	0.4		0.8				0.8	0.4	0.4		0.4							0.8	0.8		
Viola	40.8	**36.7**	20.4	1.5												0.5											
Cello	9.8	13.9	**50.4**	24.8							0.4															0.4	0.4
D. bass	6.6	11.5	40.7	**39.6**		0.5					0.5												0.5				
Flute	1.0				**63.3**	22.4	12.2					1.0															
Piccolo					57.1	4.8	**38.1**																				
Oboe	3.6	1.2			2.4	1.2	2.4	**56.0**	8.3			11.9	3.6		1.2		1.2							2.4	3.6		1.2
E. horn								35.7	**46.4**	5.4		5.4	1.8			1.8									1.8	1.8	
Bassoon								3.6	7.1	**42.9**	42.9			9.5				1.8	1.8	1.8	1.8	1.8	5.4		3.6	3.6	
B-flat Clar.								11.9		1.2		**42.9**	9.5	2.4		1.2	2.4							3.6	7.1		
Trumpet	0.8							1.2	0.4			3.2	1.2	0.4	**69.8**	13.9	2.0	2.4		1.2		0.4		2.0	0.4	0.4	0.4
Fr. horn	1.4								0.7	0.7					15.7	2.1	7.1	**35.7**	7.9	22.1	3.6	0.7	1.4			0.7	
Ten. tromb.		0.7						2.1	7.9	10.0	0.7	1.4		0.7	3.6	5.7	8.6	29.3	4.3	**17.9**	4.3					2.9	
Tuba				4.3					1.4	4.3	8.6					1.4	8.6	38.6		11.4	2.9	10.0	**7.1**				1.4
Totals	12.8	11.3	14.7	7.5	4.6	1.4	1.6	4.4	2.8	2.4	1.0	3.9	0.9	0.5	10.7	2.6	1.9	6.5	0.9	3.5	0.7	0.5	0.6	0.6	0.7	0.5	0.2

TABLE 14. Confusion matrix for the isolated tone component of the experiment. Entries are expressed as percentages. The dashed boxes indicate within-family confusions.

Confusion matrix for the ten-second excerptdata, compiled across all subjects

Responded Presented	Violin	Viola	Cello	Double bass	Flute	Alto flute	Piccolo	Oboe	English horn	Bassoon	Contrabassoon	B-flat clarinet	E-flat clarinet	Bass clarinet	Trumpet	Cornet	Fluegel horn	French horn	Alto trombone	Tenor trombone	Bass trombone	Euphonium	Tuba	Soprano saxophone	Alto saxophone	Tenor saxophone	Baritone saxophone
Violin	**69.6**	28.0	2.4																								
Viola	27.7	**48.2**	23.2	0.9																							
Cello	2.4	6.0	**78.6**	13.1																							
D. bass			3.6	**92.9**																			3.6				
Flute				0.5	**81.4**	14.3	3.3					0.5															
Piccolo					35.7		**64.3**																				
Oboe								**73.8**	14.3			4.8	3.6			2.4								1.2			
E. horn	1.8							46.4	**33.9**	7.1		5.4	1.8	1.8												1.8	
Bassoon								1.2	7.1	**78.6**	4.8		1.2	2.4							1.2					2.4	1.2
B-flat Clar.	0.5							1.1		2.7		**73.6**	16.5	4.9										0.5			
Bass Clar.										14.3		7.1	14.3	**64.3**													
Trumpet												0.9			**76.8**	18.8	2.7	0.9									
Fr. horn									1.2						3.6	1.2	4.8	**73.8**	4.8	3.6	4.8	2.4					
Alt. Tromb.									3.6						7.1	7.1	17.9	39.3	**14.3**	10.7							
Ten. tromb.															3.6	1.8	10.7	30.4	3.6	**50.0**							
Euphon.									3.6								17.9	14.3		14.3	3.6	**42.9**	3.6				
Tuba																		3.6			7.1	14.3	**75.0**				
Sop. Sax.																								**7.1**	42.9	46.4	3.6
Alto Sax.																	3.6	3.6						3.6	**28.6**	50.0	10.7
Totals	10.6	7.4	6.8	2.7	12.3	2.1	1.1	6.4	2.8	5.4	0.3	10.1	2.6	1.5	6.5	1.9	1.7	6.8	0.7	2.7	0.6	1.3	1.6	0.4	1.4	2.1	0.4

TABLE 15. Confusion matrix for the ten-second excerpt component of the experiment. Entries are expressed as percentages. The dashed boxes indicate within-family confusions.

Responded Presented	Strings	Brass	Double reeds	Clarinets	Flutes	Saxophones
Strings	**97.8**	0.4	0.4	0.3	0.4	0.7
Brass	1.3	**86.4**	7.5	2.5		2.3
Double reeds	2.0	5.6	**73.5**	10.7	2.6	5.6
Clarinets		3.6	13.1	**72.6**		10.7
Flutes	0.7			0.7	**98.6**	
Totals	46.3	28.1	10.6	5.3	7.7	2.1

TABLE 16. Family confusion matrix for the isolated tone component of the experiment. Entries are expressed as percentages.

Responded Presented	Strings	Brass	Double reeds	Clarinets	Flutes	Saxophones
Strings	**99.7**	0.3				
Brass		**98.8**	0.9	0.3		
Double reeds	0.4	1.3	**89.3**	6.7		2.2
Clarinets	0.5		4.6	**94.4**		0.5
Flutes	0.4			0.4	**99.1**	
Saxophones		3.6				**96.4**
Totals	27.6	23.7	14.8	14.1	15.5	4.2

TABLE 17. Family confusion matrix for the 10-second phrase component of the experiment. Entries are expressed as percentages.

	Isolated tone Condition	Ten second segment Condition
% Exact responses	45.9	66.9
% correct family	91.7	96.9
$\frac{P(\text{within-family error})}{P(\text{between-family error})}$	5.5	9.7

TABLE 18. Summary of human performance in the two conditions, pooled across all subjects.

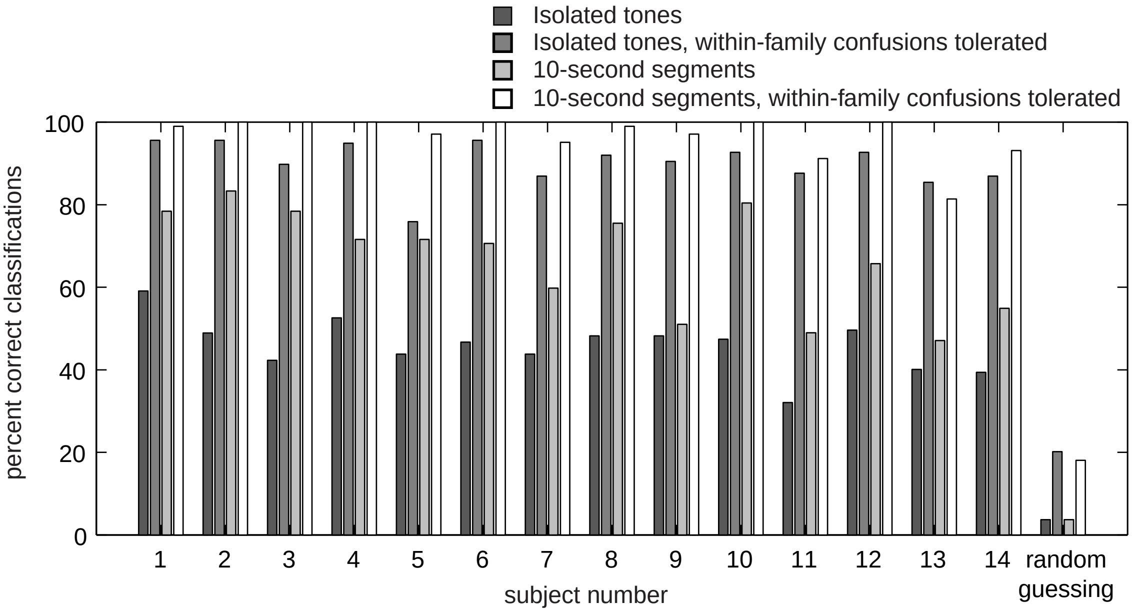

FIGURE 42. Performance by subject on the two components of the experiment. Separate results are shown for identification of the correct instrument, and of the correct family group.

Instrument	$\frac{\|S_i \cap R_i\|}{\|S_i \cup R_i\|} \times 100$		Ranking in isolated tone condition
	10-second excerpts	Isolated tones	
Flute	79.5	49.6	[2]
Trumpet	72.3	62.6	[1]
B-flat clarinet	70.0	47.2	[3]
Tuba	70.0	6.6	[14]
Bassoon	69.5	30.8	[8]
Double bass	63.4	28.4	[9]
Violin	57.6	31.8	[6]
Cello	57.4	32.4	[5]
Oboe	54.9	38.5	[4]
French horn	52.1	23.3	[11]
Piccolo	42.9	28.1	[10]
Tenor trombone	42.4	13.7	[13]
Euphonium	35.3	-	-
Bass clarinet	34.6	-	-
Viola	32.9	21.2	[12]
English horn	24.7	31.3	[7]
Alto saxophone	20.0	-	-
Alto trombone	11.8	-	-
Soprano saxophone	6.5	-	-

TABLE 19. “Recognizability” scores by instrument, calculated as the number of trials in which the instrument was correctly identified divided by the total number of trials in which the instrument appeared as either a stimulus or a response. Instruments are sorted by their rank in the ten-second excerpt condition.

6.2.3 Discussion

There are several statistics from this experiment that can be compared to previous studies, including correct-response rate and within- and between-family confusion rates. Pooling across all subjects in the isolated-tone condition, the exact instrument was selected on 45.9% of trials, and the correct family on 91.7% of trials (a random guesser would score 3.7% and 20.2% on these statistics). In this condition, a within-family error is 5.5 times more likely to occur than a between-family error. Success rates from previous studies include 35-57% exact answers on a free-response task (Eagleson & Eagleson, 1947), 85% on an 8-way forced-choice task (Strong & Clark, 1967), and 59% on a 10-way forced-choice task (Berger, 1964). Strong's subjects identified the correct family on 94% of trials (thus, within-family confusions were 1.5 times more likely than between-family confusions). Berger's subjects identified the correct family 88% of the time (within-family confusions were 2.4 times more likely than between-family confusions).

Pooling across all subjects in the ten-second-exceprt condition, subjects responded with the exact instrument on 66.9% of trials, and with the correct family on 96.9% of trials (a random guesser would score 3.7% and 18.1% on these statistics). In this condition, a within-family error is 9.7 times more likely to occur than a between-family error, rather strongly highlighting the perceptual salience of the instrument families.

Previous studies suggest that certain small groups within families are particularly difficult to distinguish. Within the string family, for example, Robertson (1961) reported common confusions between violin and viola, and between cello and double bass. The confusion matrices from both components of the current experiment (Tables 8 and 15) exhibit a strong diagonal band, with each string instrument commonly confused with its neighbors in size. Confusions occurred particularly often between violin and viola. Viola samples were also very often classified as cello, although the converse does not hold.

Robertson (1961) also reported frequent confusions between instruments of the brass family, particularly between instruments of similar size. Saldanha and Corso (1964) reported common confusions of trumpet with cornet, saxophone (not a member of the brass family!), and French horn; and trombone with French horn, saxophone, and trumpet. Berger (1964) reported common confusions between trumpet and cornet; and French horn, baritone, and trombone. Schlossberg (1960) reported confusions between trombone and trumpet; and French horn and trombone. In the ten-second-excerpt data, brass instruments were commonly confused. Trumpet samples were classified correctly on 76.8% of all trials, but were confused with cornet (18.8%), fluegel horn (2.7%), French horn (0.9%), and clarinet (0.9%). Tuba samples were classified correctly on 75.0% of trials, but were confused with Euphonium (14.3%), bass trombone (7.1%), and French horn (3.1%). Euphonium samples were classified correctly on 42.9% of trials, and were confused with fluegel horn (17.9%), French horn (14.3%), tenor trombone (14.3%), bass trombone (3.6%), tuba (3.6%), and English horn (3.6%). These statistics do not suggest particularly salient subgroups, but it is interesting to note that across all trials, most mistaken classifications were as French horn, but French horn was misclassified relatively infrequently.

Within the double-reed family, frequent confusions between oboe and English horn were reported by Robertson (1961) and Saldanha and Corso (1964). The data from this experiment support the oboe-English horn confusion pair, though oboe was selected much more often than English horn. Subjects 13 and 14 accounted for nearly all (21) of the confusions of the double-reed instruments with the clarinets (subjects 11 and 7 contributed made three such confusions; no other subjects made any).

The clarinet family did not exhibit any strong subgroups, except possibly between B-flat and E-flat clarinet. The E-flat clarinet is used much less frequently than the other clarinets in performances, and no recordings of it were used in this study. Similarly, the flute family did not exhibit strong subgroups, except possibly between flute and alto flute. Again, however, no recordings of the alto flute were used, and only one piccolo recording was available, so no strong conclusions can be drawn. So few recordings of saxophones were used in the study that analysis of confusions is impossible.

Several previous studies indicated that some instruments are easier to recognize than others, but such effects appear to depend rather strongly on the details of the experiment. For example, Eagleson and Eagleson (1947) found that violin and trumpet were easiest to identify in a free-response task, and that alto-horn, piccolo, and flute were the most difficult. It is likely, however, that the violin score was elevated because it is the most well-known string instrument and no other strings were used as stimuli. Trumpet is similarly well known, and alto horn (a relatively rare instrument) was the only other brass instrument used in the study. Piccolo and flute may have been confused with each other, leading to poor identification scores.

Saldanha and Corso (1964) found that B-flat clarinet, oboe, and flute tones were most easily classified, and that violin, cello, and bassoon tones were most difficult. Their study, however, did not include English horn or piccolo tones, which may have elevated the oboe and flute scores respectively. The fundamental frequencies they tested were very high in the bassoon playing range (and relatively high in the cello playing range as well), possibly contributing to its low score.

Berger (1964) found that oboe tones were easiest to classify, and that flute and trumpet tones were most difficult. His study, however, included no double-reeds other than oboe, thereby elevating its score, but several brass instruments, including the cornet, which is easily confused with trumpet.

The correct-classification scores for the instruments used in the current study are shown in Table 19 on page 129, sorted in decreasing order based on the ten-second-excerpt portion of the data. Scores for the isolated-tone data are shown alongside. Flute, trumpet, and B-flat clarinet scored well in both conditions. Viola and tenor trombone scored poorly in both conditions.

A final point of comparison with previous studies is the relative performance of individual subjects. Figure 42 on page 128 shows the performance of each subject on the two conditions, with separate scores for identifications with and without toleration of within-family confusions. Only one subject (#4) was a professional musician; the others were university graduate and undergraduate students. Interestingly, subject #1, who scored highest on the isolated-tone classification task, is the only subject with “perfect pitch”. In a post-experiment interview, he admitted to using rather exact knowledge of the pitch-ranges of the various instruments to improve his judgments, particularly in the isolated-tone condition.

6.3 Computer experiment #1: Isolated tone pilot study[1]

While the recognition architecture described in Chapters 4 and 5 was under development, a short pilot study was conducted to test some of the feature-extraction techniques described in Chapter 4 and to evaluate their usefulness for recognizing the sources of isolated musical tones. 1023 tones were selected from the MUMS collection, covering the full pitch ranges of fourteen instruments (violin, viola, cello, bass, flute, piccolo, B-flat clarinet, oboe, English horn, bassoon, trumpet, trombone, French horn, and tuba) playing several different articulation styles (e.g., pizzicato, bowed, muted).

For this study, 31 one-dimensional features were computed from the weft representation of each instrument tone. These included the pitch, spectral centroid, attack asynchrony (both the relative onset times of partials at different frequencies, and their overall variation), ratio of odd-to-even harmonic energy (based on the first six partials), and the strength of vibrato and tremolo. Many of the 31 features were subtle variations of other features included in the set, measured in a slightly different manner. The feature set was intended to be representative of the features described in Chapter 3 but certainly not exhaustive. For example, the *shape* of the spectral envelope was not considered at all in this study. Table 20 contains a list of the features that were extracted.

Several instrument-class taxonomies were constructed and various pattern-recognition techniques were used to build statistical classifiers at each node. Statistical classifiers require a set of training data whose size grows exponentially with the number of feature dimensions, and with 31 features, the necessary data set size is much larger than what was available. To reduce the training requirements, Fisher multiple discriminant analysis (McLachlan, 1992) was employed at each node of the taxonomy. The Fisher technique projects the high-dimensional feature space into a space of fewer dimensions (the number of dimensions is one fewer than the number of data classes at the node) where the classes to be discriminated are maximally separated. The analysis yields the mean feature vector and covariance

[1] The results of this study were reported in (Martin & Kim, 1998). This section is a condensed version of the paper written for that conference. The statistical classifiers were implemented and tested by Youngmoo Kim.

matrix (in the reduced space) of a single normal density for each class, which can be used to form maximum *a posteriori* (MAP) classifiers by introducing prior probabilities. The taxonomy that resulted in the best overall classification performance (of those that were tested—the search was not exhaustive) is shown in Figure 43. Figures 44 and 45 show the decision spaces found at two of the nodes of the taxonomy.

Average pitch over steady state	Tremolo frequency
Average pitch Δ ratio[a]	Tremolo strength
Pitch variance	Tremolo heuristic strength[b]
Pitch variance Δ ratio[a]	Spectral centroid modulation frequency (Hz)
Average spectral centroid (Hz)	Spectral centroid modulation strength
Spectral centroid Δ ratio[a]	Spectral centroid modulation heuristic strength[b]
Variance of spectral centroid	Normalized spectral centroid modulation frequency (Hz)
Spectral centroid variance Δ ratio[a]	Normalized spectral centroid modulation strength
Average normalized spectral centroid	Normalized spectral centroid modulation heuristic strength[b]
Normalized spectral centroid Δ ratio[a]	Slope of the onset harmonic skew[c]
Variance of normalized spectral centroid	Intercept of the onset harmonic skew[c]
Normalized spectral centroid variance Δ ratio[a]	Variance of the onset harmonic skew[c]
Maximum slope of onset (dB/msec)	Post-onset slope of amplitude decay
Onset duration (msec)	
Vibrato frequency (Hz)	Odd/even harmonic ratio
Vibrato amplitude	
Vibrato heuristic strength[b]	

TABLE 20. List of features extracted from each tone in the pilot study.

[a] The Δ ratio is the ratio of the feature value during the transition period from onset to steady state (~100 ms) to the feature value after the transition period.

[b] The heuristic strength of a feature is the peak height from the DFT divided by the average value surrounding the peak.

[c] The onset harmonic skew is a linear fit to the onset times of the harmonic partials (defined as time the partial reached an energy level 3 dB below the steady-state value) as a function of frequency.

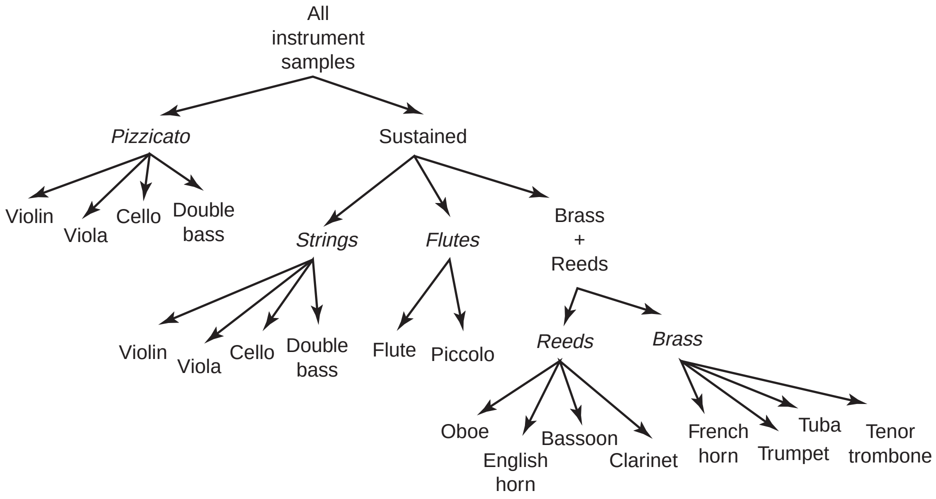

FIGURE 43. Taxonomy used in the pilot study. Instrument family groups are shown in italics. The leaf nodes are the individual instrument classes.

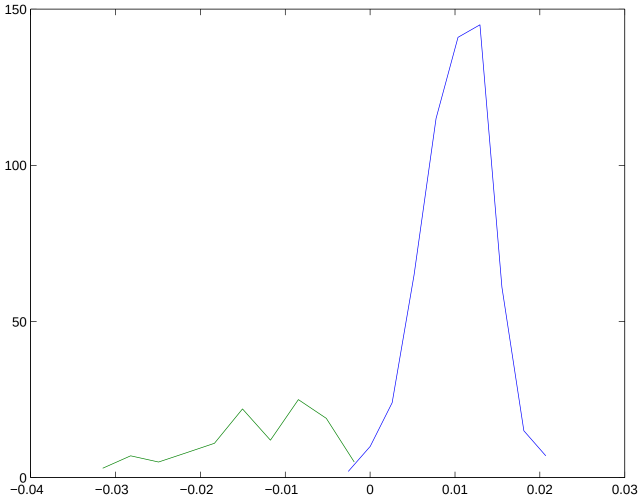

FIGURE 44. Fisher projection for the Pizzicato vs. Sustained node of the taxonomy. Since there are two classes, the projection is one-dimensional. There are “modes” in the projection: the one on the left-hand side corresponds to Pizzicato tones; the one on the right to Sustained tones. The Sustained tone distribution is favored by prior probability and therefore appears larger. The axes are not labeled; the abscissa is a linear combination of the 31 features.

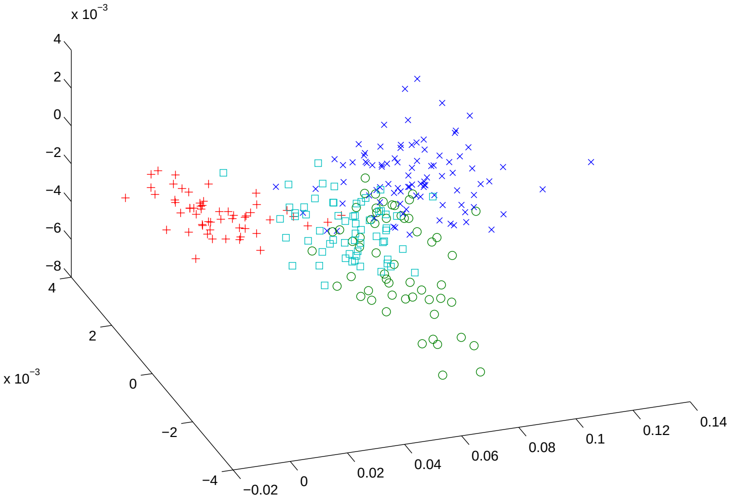

FIGURE 45. Fisher projection for classifying the individual string instruments. There are four classes and thus three dimensions in the projection. Violin data points are plotted with X's, viola with O's, cello with plus symbols and double bass with squares. The axes are not labeled. Each axis is a linear combination of the 31 features.

In addition to the Fisher projection technique, two varieties of *k*-nearest neighbor (*k*-NN) classifiers were tested. A *k*-NN classifier works by memorizing the feature vectors of all of the training samples. When a new sample is to be classified, the system finds the *k* nearest training samples in the feature space (usually using a Euclidean distance metric), and the new sample is classified by majority rule based on the labels of the *k* training samples.

To evaluate the performance of the various classifiers, each was trained with 70% of the MUMS tones, leaving 30% as independent test samples. Table 21 contains a summary of the classification performance of the hierarchical Fisher classifier, a hierarchical *k*-NN classifier, and a non-hierarchical *k*-NN classifier. The results are averaged over 200 test runs with different training/test data splits. The hierarchical Fisher classifier performs best, particularly at the individual instrument level.

Level of taxonomy	Hierarchical Methods		Non-hierarchical *k*-NN
	Fisher + MAP	*k*-NN	
Pizzicato vs. sustained	98.8%	97.9%	97.9%
Instrument family	85.3%	79.0%	86.9%
Individual instruments	71.6%	67.5%	61.3%

TABLE 21. Classification results for the three classifiers tested. Each result was cross-validated with 200 test runs using 70%/30% splits of the training/test data.

Although Fisher and *k*-NN techniques yield successful classifiers, they provide little insight into the relative importance of the various individual features. It would be valuable to know if particular features are good at characterizing particular instruments or families. To that end, a step-forward algorithm was used to find the best features for isolating each instrument family. A step-forward algorithm works by testing each feature individually and choosing the best as the *current set*. The algorithm continues by testing all combinations of the current set with each of the remaining features, adding the best of these to the current set and repeating. For computational simplicity, only *k*-NN classifiers were used in this part of the study. This procedure was followed using three different 70%/30% splits of the training/test data, iterating 10 times to find the 10-feature combination that provided the best average performance over the three different data sets.

By using only the 10 best features at each node, the system's success rate for instrument family identification increased to 93%. Some of the features were generally salient across many of the instrument families, and some were particularly useful in distinguishing single families. The most common features selected for each subgroup are listed in Table 22.

Vibrato strength and features related to the onset harmonic skew (roughly, the relative onset times of the various partials) were selected in four of the five instrument subgroups, indicating their relevance across a wide range of isolated instrument tones. One interesting omission occurs with the clarinet group. One of the 31 features was the ratio of odd to even harmonic energy. The conventional-wisdom about the clarinet is that its odd partials are much stronger than its even partials, but this is not true over the clarinet's entire range, and this study did not find it to be a very useful feature.

This pilot study has two results worth noting. First, it demonstrates the utility of a hierarchical organization of sound sources, at least for the limited range of sources it considered. Second, it demonstrates that the acoustic properties suggested by the musical acoustics and analysis-by-synthesis literature (see Chapter 3) are indeed useful features for musical instrument recognition.

Subgroup	Selected features
Strings	Vibrato strength
	Onset harmonic skew
	Average spectral centroid
Brass	Vibrato strength
	Variance of spectral centroid
	Onset harmonic skew
Clarinets	Pitch variance
	Onset duration
	Vibrato strength
	Onset harmonic skew
Flutes	Pitch
	Onset duration
	Tremolo strength
	Spectral centroid
	Vibrato frequency
Double reeds	Vibrato strength
	Average spectral centroid
	Spectral centroid modulation
	Onset harmonic skew

TABLE 22. Features that were particularly useful in distinguishing single instrument families.

Not surprisingly, the hierarchical classifier performs better than humans on this classification task. It is unfair, however, to compare its performance directly with the results from Section 6.2.[1] The classifier has learned to identify the instruments from the MUMS collection with great success, but it is not in any way a demonstration of performer-independent generalization. Because of the particular form of cross-validation used in this study, on any given trial the computer had been trained with tones produced by the same performer. The human listeners did not enjoy the same advantage. The next two sections address this limitation of the pilot study.

[1] Although the comparison is unfair, to save you the trouble of looking up the result, the human subjects averaged 45.9% exact identifications (91.7% allowing within-family confusions). The computer program scored better on exact classifications, but not quite as well on determining the family. It should also be noted that the stimulus set was not the same in the two experiments, though there was substantial overlap.

6.4 Computer experiment #2: 6- to 8-way classification

Although the isolated-tone pilot study showed that the features used by the system enabled good classification results on isolated-tone stimuli, two troublesome aspects of the study make it difficult to draw any strong conclusions from it. To address these issues, a second experiment was performed using more realistic stimuli and more principled cross-validation.

Of the 27 instruments considered in the human experiment, recordings of more than three independent performers were available for only five: violin, viola, trumpet, B-flat clarinet, and flute (bassoon, cello, French horn, and oboe each had three; each of the others had fewer). Three sub-experiments were conducted with subsets of this list, using 6, 7, and 8 instruments respectively. Violin, viola, cello, trumpet, B-flat clarinet, and flute were used in the first sub-experiment. The second sub-experiment added French horn, and the third added oboe (bassoon was omitted because the available recordings of two of the performers were very short).

In each sub-experiment, the stimuli from the human experiment corresponding to the selected instruments were used to test the system. For each trial, the computer system was trained with all of the recordings available for those instruments—except those by the particular performer being tested. This form of *leave-one-out* cross-validation makes good use of the available training data, yet still provides a fair test because on every trial the system was not trained on any recordings by the performer playing on the sample being tested.

Because the number of classes in each sub-experiment was so small, the system was configured to use a flat hierarchy (i.e., there was only one decision node, and each instrument formed a leaf node). With the flat hierarchy, beam search is meaningless, so it was not used. Four variations of context-dependent feature selection (see Section 5.3.1 on page 107) were tested: (1) no salience weights, (2) average salience score based on the classes currently under consideration, (3) salience based only on reliability estimates, and (4) the product of (2) and (3). In all cases, the "rule-one out" extension (see Section 5.3.2 on page 109) was used.

Table 23 shows the main results of the experiment, organized by the number of instrument classes tested and by the form of context-dependent feature selection. In each sub-experiment, the best configuration employed average feature-salience scores based on the current set of classes under consideration (case 2). The worst-performing configuration in each sub-experiment used salience weights based only on reliability estimates (case 3). Unsurprisingly, performance improves as the number of instrument classes decreases.

Tables 24-26 show the confusion matrices for the best-performing configuration in each sub-experiment. Like the human subjects, the computer system tends to confuse violin with viola, and viola with cello. Other mistakes are consistent across the three sub-experiments but do not bear obvious relationships to the mistakes made by the subjects in the human experiment. They may be due to quirks of the particular feature-extraction algorithms, but are probably just due to an insufficient feature set or insufficient training data.

Condition	No Salience Weights (1)	Salience weights based on current set of classes (2)	Salience weights based on confidence ratings (3)	Combined salience weights (4)
8-way	68.9% (78.4%)	73.0% (83.8%)	67.6% (79.7%)	68.9% (82.4%)
7-way	75.0% (85.3%)	77.9% (89.7%)	72.1% (85.3%)	73.5% (86.8%)
6-way	77.4% (88.7%)	82.3% (95.2%)	71.0% (85.5%)	77.4% (93.6%)

TABLE 23. Results of computer experiment #2. In all cases, performance was best in the second salience-weight configuration, which chooses feature subsets based on their ability to discriminate among the particular sound-source classes under consideration. In each box the percentage of exact responses is given (along with the percentage of correct responses if within-family confusions are tolerated).

Responded Presented	Violin	Viola	Cello	Trumpet	B-flat clarinet	Flute	French horn	Oboe
Violin	**75.0**	25.0						
Viola	12.5	**50.0**	12.5					
Cello		16.7	**83.3**					
Trumpet		12.5		**75.0**		12.5		
B-flat clarinet					**84.6**	7.7	7.7	
Flute						**93.3**		6.7
French horn		33.3			16.7		**50.0**	
Oboe	16.7			50.0				**33.3**
Totals	16.2	14.9	9.5	12.2	16.2	21.2	5.4	4.0

TABLE 24. Confusion matrix for the 8-way classification experiment. Results are reported as percentages. The classifier answered correctly on 73.0% of trials (83.8% allowing within-family confusions).

Responded Presented	Violin	Viola	Cello	Trumpet	B-flat clarinet	Flute	French horn
Violin	**75.0**	25.0					
Viola	12.5	**50.0**	12.5				
Cello		16.7	**83.3**				
Trumpet	12.5			**75.0**		12.5	
B-flat clarinet					**84.6**	7.7	7.7
Flute						**100.0**	
French horn		33.3			16.7		**50.0**
Totals	17.7	14.7	10.3	8.8	17.7	25.0	5.9

TABLE 25. Confusion matrix for the 7-way classification experiment. Results are reported as percentages. The classifier answered correctly on 77.9% of trials (89.7% allowing within-family confusions).

Responded Presented	Violin	Viola	Cello	Trumpet	B-flat clarinet	Flute
Violin	**75.0**	25.0				
Viola	12.5	**50.0**	12.5			
Cello		16.7	**83.3**			
Trumpet	12.5			**75.0**		12.5
B-flat clarinet					**92.3**	7.7
Flute						**100.0**
Totals	16.2	14.9	9.5	12.2	16.2	21.2

TABLE 26. Confusion matrix for the 6-way classification experiment. Results are reported as percentages. The classifier answered correctly on 82.3% of trials (95.2% allowing within-family confusions).

6.5 Computer experiment #3: Direct comparison to human abilities

A final experiment was performed to enable a more direct comparison between human abilities (based on the experiment described in Section 6.2) and the abilities of the recognition system. The full stimulus set used in the human experiment was employed to test the system. As with Computer experiment #2, on each trial the computer system was trained with all of the available recordings—except those by the particular performer being tested. This form of *leave-one-out* cross-validation makes good use of the available training data, yet still provides a fair test because on every trial the system is guaranteed to have never heard any performances by the musician playing on the recording being tested.

The system was configured to use the taxonomy shown in Figure 46, which includes an instrument-family layer based on the discussion in Chapter 3. The best-performing configuration from Computer experiment #2 was employed, using salience weights based on the average discriminating power of each feature for the particular categories being considered at any time. In all cases, the "rule-one-out" extension was used. Three values were tested for the beam-width parameter (1, 3, and infinite).

Table 21 shows the main results of the experiment, organized by beam width and by experiment component. With a beam width of 3 or greater, the computer system performs better than subjects 9, 11, and 13 on the ten-second excerpt component. With an infinite beam width, the system performed better than subject 11 on the isolated-tone component. All of the human subjects, however, scored much better than the computer system if within-family confusions are tolerated. Tables 28 and 29 show the computer system's confusion matrices for the two conditions.

Experimental condition	**Beam Width**		
	1	**3**	**Infinite**
Isolated tones	32.2% (69.3%)	32.9% (72.3%)	38.7% (75.9%)
Ten-second excerpts	41.2% (53.9%)	55.9% (70.6%)	56.9% (74.5%)

TABLE 27. Percentage of correct classifications for the computer recognition system configured to use the taxonomy shown in Figure 46, with beam searches of various widths. Values in parentheses indicate performance if within-family confusions are allowed.

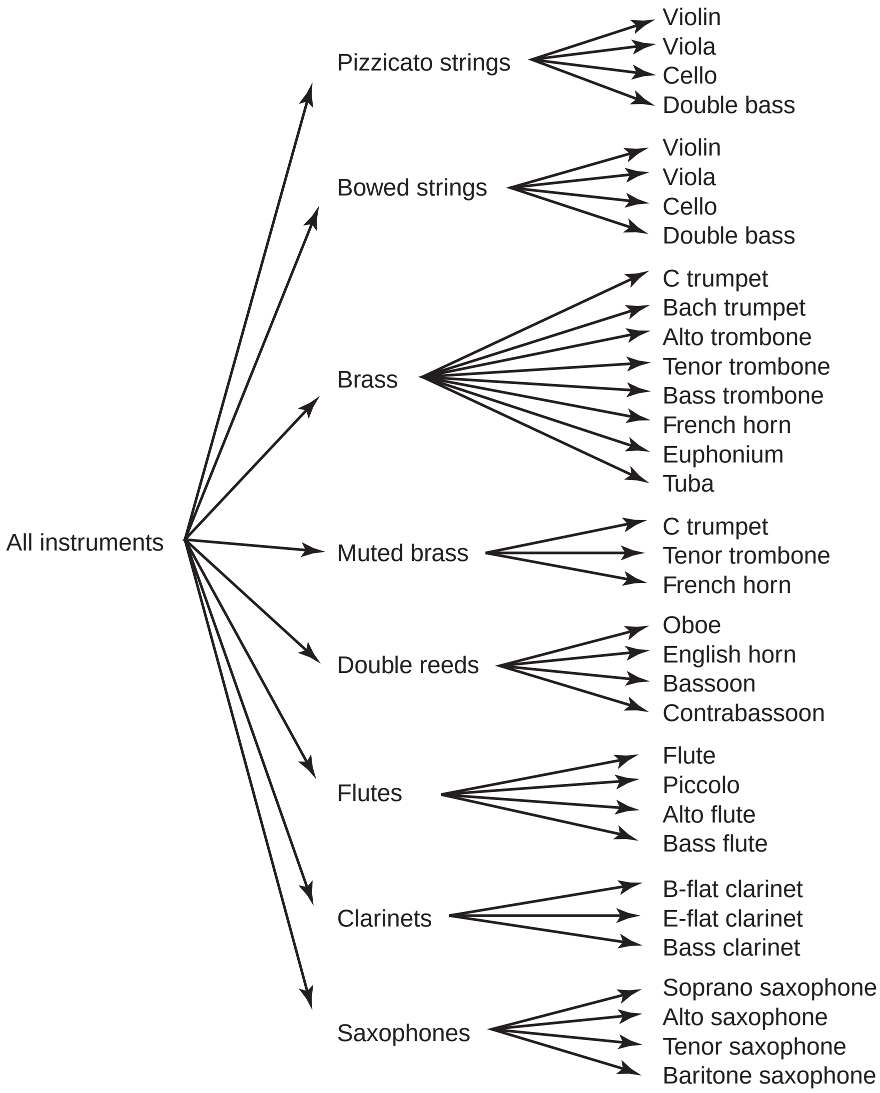

FIGURE 46. Taxonomy used in Computer experiment #3 to test the recognition system.

Responded Presented	Violin	Viola	Cello	Double bass	Flute	Alto flute	Piccolo	Oboe	English horn	Bassoon	Contrabassoon	B-flat clarinet	E-flat clarinet	Bass clarinet	Trumpet	Cornet	Fluegel horn	French horn	Alto trombone	Tenor trombone	Bass trombone	Euphonium	Tuba	Soprano saxophone	Alto saxophone	Tenor saxophone	Baritone saxophone
Violin	**44.4**	27.8	11.1												16.7												
Viola	28.6	**50.0**						7.1		7.1		7.1															
Cello	10.5	5.3	**42.1**	31.6											10.5												
D. bass		7.7	69.2	**15.4**	7.7																						
Flute			42.9		**28.6**		28.6																				
Piccolo	33.3	33.3																						33.3			
Oboe	33.3							**33.3**							16.7				16.7								
E. horn										25.0					50.0				25.0								
Bassoon			25.0							**50.0**								25.0									
B-flat Clar.	16.7							16.7				**16.7**	16.7	16.7										16.7			
Trumpet	22.2		5.6												**66.7**				5.6								
Fr. horn															10.0			**50.0**		10.0		20.0	10.0				
Ten. tromb.																			40.0	**40.0**			20.0				
Tuba				20.0														20.0				60.0					
Totals	16.1	11.0	17.5	6.6	2.2	0.0	1.5	2.9	0.0	2.9	0.0	1.5	0.7	0.7	15.3	0.0	0.0	5.1	5.1	3.7	0.0	0.0	2.2	1.5	0.0	0.0	0.0

TABLE 28. Confusion matrix for Computer experiment #3: Isolated-tone condition.

Responded Presented	Violin	Viola	Cello	Double bass	Flute	Alto flute	Piccolo	Oboe	English horn	Bassoon	Contrabassoon	B-flat clarinet	E-flat clarinet	Bass clarinet	Trumpet	Cornet	Fluegel horn	French horn	Alto trombone	Tenor trombone	Bass trombone	Euphonium	Tuba	Soprano saxophone	Alto saxophone	Tenor saxophone	Baritone saxophone
Violin	**83.3**	8.3																						8.3			
Viola	37.5	**37.5**	12.5																						12.5		
Cello		16.7	**66.7**																			16.7					
D. bass				**100.0**																							
Flute	6.7	6.7			**80.0**															6.7							
Piccolo					100.0																						
Oboe								**33.3**							66.7												
E. horn					50.0				**25.0**										25.0								
Bassoon					16.7					**83.3**																	
B-flat Clar.					15.4							**76.9**												7.7			
Bass Clar.												100.0															
Trumpet	12.5							12.5							**75.0**												
Fr. horn										16.7								**33.3**		16.7		33.3					
Alt. Tromb.																				100.0							
Ten. tromb.																			100.0								
Euphon.			50.0															50.0									
Tuba			50.0																				**50.0**				
Sop. Sax.			50.0																50.0								
Alto Sax.			100.0																								
Totals	14.1	5.9	9.8	2.0	17.7	0.0	0.0	2.9	1.0	5.9	0.0	10.8	0.0	0.0	9.8	0.0	0.0	2.9	5.9	3.9	0.0	2.9	1.0	2.0	1.0	0.0	0.0

TABLE 29. Confusion matrix for Computer experiment #3: Ten-second excerpt condition.

6.6 General discussion

Although the human experiment described in Section 6.2 and Computer experiment #3 (described in Section 6.5) afford the most direct comparison between human and computer performance on musical instrument classification tasks, the comparison is still not completely fair. As described in Chapter 2, there are several criteria that must be kept in mind when making such comparisons. Considering each in turn:

- **Do the computer system and humans exhibit the same level of generalization?** No. The computer system described here demonstrates the most general performer-independent recognition of musical instruments of any system described to date. However, the tests used to demonstrate this ability were limited, and it is not possible to make strong claims about generalization. It is very interesting to note that the human listeners who participated in the experiment made particular kinds of mistakes that suggest that they have succeeded in generalizing abstract groups of instruments—namely the instrument families. The computer system did not exhibit this particular generalization.
- **Do the computer system and humans handle real-world complexity equivalently?** No. Both the computer system described here and the human experimental subjects exhibit robust classification performance with typical commercial music recordings, which include reverberation (and, occasionally, high levels of ambient noise). With the possible exception of the systems described by Brown (1998a; 1999) and Marques (1999), the computer system described here is much more robust in this regard than any other system described to date. However, although it was not tested, complexity arising from mixtures of sounds would surely cripple the computer system. I speculate that the performance of human subjects would degrade somewhat, but would be much more robust than that of the computer system with this particular kind of complexity.
- **Are the computer system and humans “equivalently scalable”?** No. Humans are capable of recognizing examples from a vastly larger set of sound sources. The computer system described here could be extended to a much larger range of sound-source classes, but doing so would require the addition of many more feature extractors and quite a lot more training data. The taxonomic recognition structure is intended to make the system more scalable than previous systems, but this aspect has not been adequately tested. Judging by the system’s classification performance in Computer experiment #3, the representations of the instrument families would have to be improved significantly to make the classifier robust with the narrow width beam-search technique.
- **Do both systems exhibit equivalently graceful degradation?** No. The computer system was designed to make good decisions based on whatever evidence is available, and its performance does degrade smoothly as particular features are removed from consideration, but it has not been tested under conditions similar to those that would be caused by masking in normal listening situations. Again, human abilities are much more robust.

- **Do both systems exhibit a flexible learning strategy?** No. The computer system requires a supervised framework in order to learn to recognize sound sources. Humans can also learn without explicit instruction (though not for this particular forced-choice task).
- **Do both systems operate in real-time?** No. The computer system operates two to three orders of magnitude more slowly than “real” time. This is due in large part to the exploratory nature of this work, but a better criticism is that the recognition architecture does not provide any means for refining its decisions over time.

Although human listeners satisfy the foregoing criteria more thoroughly than the computer model, it is worthwhile to compare human and machine performance in light of these differences. Figure 47 summarizes the published performance data for experiments using isolated tones as stimuli. The first four entries represent human performance, and as should be expected, human performance decreases somewhat as the number of categories in a forced-choice task increases. The results from the 27-way forced-choice task described in Section 6.2 are approximately equal to performance observed by Eagleson and Eagleson (1947) in a free-response task.

The other five entries in the figure show the results from tests of computer recognition systems. Again, as the number of classes increases, the performance worsens. However, the results reported by Bourne (1972), Kaminskyj & Materka (1995), Fujinaga (1998), and Computer experiment #1 (Section 6.3) are not fair estimates of the systems’ performance with truly independent test stimuli. In these four experiments, the systems had exposure during training to performances by the same musicians (in the same acoustic environment) who performed the test samples. This critical failure of the experimental protocol probably elevates the reported performance levels significantly. Only Computer experiment #3 (Section 6.5) used a fair test of performance; the results can most fairly be compared to the human 27-way forced-choice task (again, keeping in mind differences in the satisfaction of the performance criteria).

Figure 48 makes a similar comparison between experiments that used real music as stimuli. Again, the trend for both humans and computer systems is that performance suffers as the number of classes in the forced-choice task increases. The listening experiment described in Section 6.2 is the first to test human listeners with stimuli consisting of real music from a wide variety of musical instruments. All of the computer systems shown in the figure were tested fairly, with principled cross-validation techniques equivalent to those used in Computer experiment #3. The most direct comparison between human and computer is the human 27-way forced-choice task and Computer experiment #3 (a 25-way forced-choice task). On average, the human listeners scored somewhat higher than the computer model, as described in Section 6.5.

Of the computer systems shown in the figure, the most direct comparison can be made between Marques’s (1999) 8-way classifier and the 8-way classifier from Computer experiment #2 (Section 6.4). Although the reported performance levels

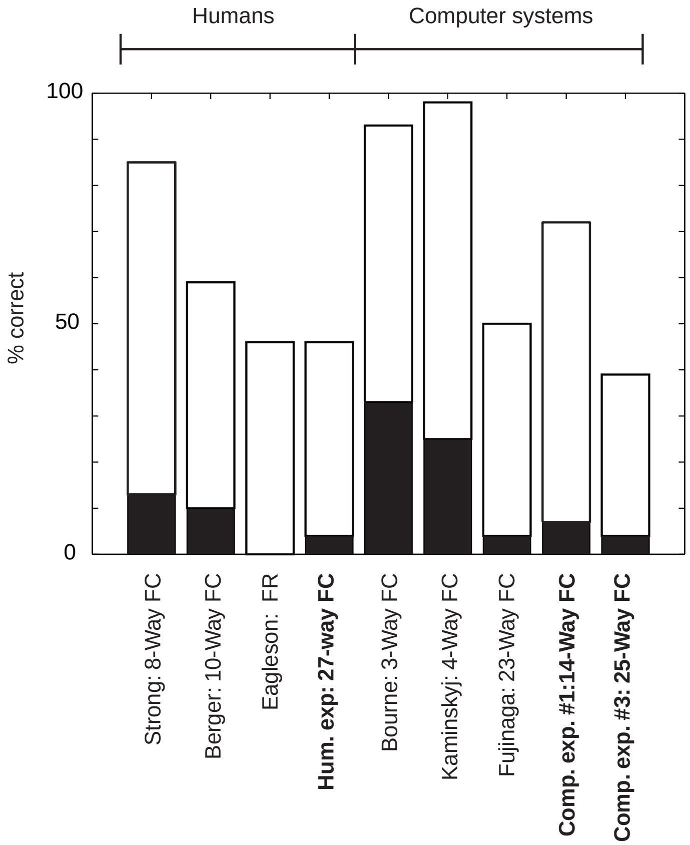

FIGURE 47. Comparison of human and computer abilities on isolated-tone recognition tasks. The open bars indicate the percentage of correct responses on the task; the filled bars indicate the level achieved by uniform random guessing. Human results are shown for Strong's (1967) 8-way forced choice experiment, Berger's (1964) 10-way forced choice experiment, Eagleson & Eagleson's (1947) free-response task (with tones from nine instruments as stimuli), and the human experiment discussed in Section 6.2. Computer results are shown for Bourne's (1972) 3-way classifier, Kaminskyj & Materka's (1995) 4-way classifier, Fujinaga's (1998) 23-way classifier, and Computer experiments #1 and #3 from Sections 6.3 and 6.5. Of the computer systems, only Computer experiment #3 employed performer-independent cross-validation.

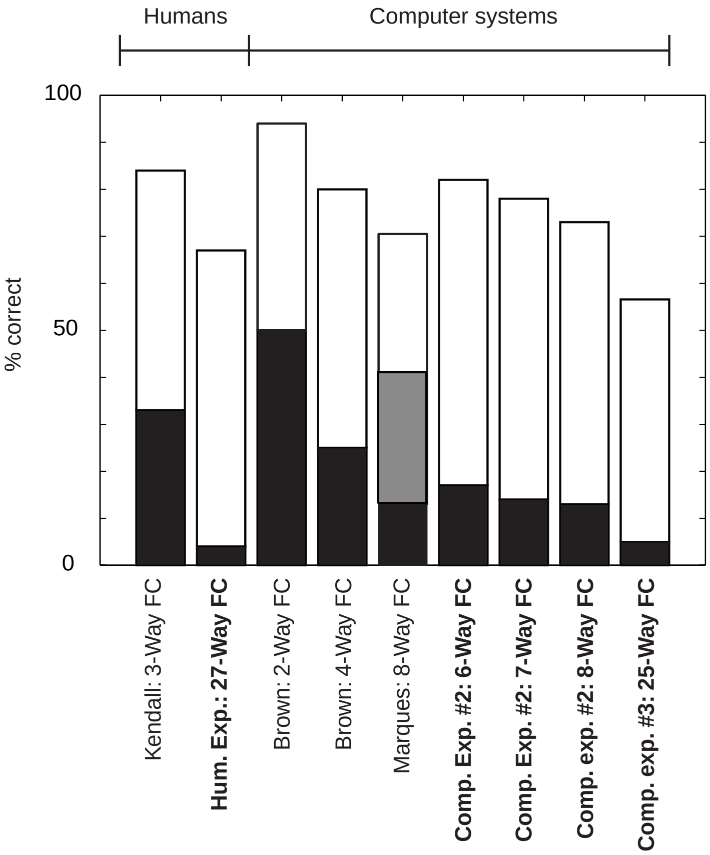

FIGURE 48. Comparison of human and computer abilities on recognition tasks using realistic musical signals. The open bars indicate the percentage of correct responses on the task; the filled bars indicate the level achieved by uniform random guessing. Human results are shown for Kendall's (1986) 3-way forced choice experiment and the human experiment discussed in Section 6.2. Computer results are shown for Brown's (1999) 2-way classifier, Brown's (1998) 4-way classifier, Marques's (1999) 8-way classifier (9-way, actually, but only 8 choices were instruments; the grey bar shows the performance level when non-commercial recordings were added to the stimulus set), and Computer experiments #2 and #3 from Sections 6.4 and 6.5.

of the two systems appear to be similar, the classifer tested in Section 6.4 appears to have generalized more thoroughly than Marques's classifier. Marques reports a performance level of 71.6% for professionally recorded music (from compact discs). After introducing "non-professional" recordings (a subset of the student recordings described in Section 6.1) to the test set, the system's performance dropped to 44.6%. This suggests that the classifier has not *generalized* as well as the classifer tested in Section 6.4, which scored 73.0% with both professional and "non-professional" recordings as test data. I speculate that this difference is due to the different feature sets used by the two classifiers. Both Marques and Brown use MFCC coefficients as features. These capture short-term properties of the spectrum, but do not represent temporal properties of the sound, such as attack transients or vibrato. The failure of the MFCC-based computer system to generalize from obervations of these features may be related to the sensitivity of human talker recognition systems—which often use the same feature set—to variations in channel conditions (Reynolds, 1995).

CHAPTER 7 Summary and conclusions

Chapters 4-6 described the implementation and evaluation of a sound-source recognition system, based on the theory presented in Chapter 1 and on extensive perceptual analysis and modeling described in Chapters 2 and 3. In this chapter, I will take a step back and consider how well the original goals of the research have been met and what implications the work has for the fields of research from which it draws.

7.1 Summary

I began this dissertation by outlining a broad theory of sound-source recognition, considered from the standpoint of the question "what is recognition *for*?" I described sound-source recognition as a process of gathering information about an object in the listener's environment so as to enable the listener to infer unobserved properties of the object. The ability to detect predators or prey has obvious evolutionary significance, but sound-source recognition can also enable the listener to subconsciously infer the properties of sounds that are partially masked by other sounds, and this kind of inference may be the key to understanding mixtures of sounds. From this perspective, sound-source recognition is essential to the hearing process, but it is absent from the current generation of computational auditory scene analysis models.

In Chapter 2, I presented a list of desiderata for sound-source recognition systems. In light of these, I compared the abilities of the state-of-the-art in artificial recognition systems to those of humans. The general conclusion was that human listeners are much better able to recognize examples from general classes of sound sources than are the current generation of artificial systems. This is hardly surprising, but it suggests that there may be aspects of human perception that could be modeled more closely in order to improve the performance of artificial systems.

In Chapter 3, I restricted attention to the class of orchestral musical instruments. Human abilities for recognizing musical instruments were reviewed, and acoustical and perceptual research was scoured for insight into the acoustic properties most likely to account for human recognition abilities in this limited domain. My conclusion was that the most significant acoustic properties are related to the excitation and resonance structures of the musical instruments. The chapter concluded with a summary list of properties both likely to be useful during the recognition process and known to be perceivable by human listeners.

In Chapters 4 and 5, I described a musical instrument recognition system based on the insights gained from the previous chapters. In Chapter 4, I described a series of representational transformations, beginning with the acoustic waveform and resulting in an abstract model (based on perceptually salient acoustic features) of the sound source's excitation and resonance structure. The representations were functionally matched to current models of information processing in the human auditory system. In Chapter 5, I presented an improvisational classification framework for sound-source recognition based on the theory outlined in Chapter 1 and using the representational scheme from Chapter 4. The framework is the result of an attempt to satisfy many of the criteria outlined in Chapter 2, and is sufficiently general to be used with many sound-source classes in addition to the musical instruments considered here.

In Chapter 6, I tested the recognition model on a battery of classification tasks and compared its performance to that of human listeners on similar tasks. Although, the human subjects in the experiments performed better overall than the computer program, the computer model performed better than at least one musically-skilled human subject in each test condition and at least as well (and with improved generalization) as other computer systems that have been tested on similar tasks. Many aspects of this performance are of interest. For example, the same model configuration performed well with both isolated-tones and real music as test stimuli. The context-dependent feature selection extension enables the model to choose appropriate features for different contexts—attack features for isolated tones, vibrato features whenever available, and spectral features for whole-phrase stimuli—without explicit instruction (indeed, the two kinds of stimuli were never distinguished during training). The model's success on this variety of stimuli is unprecedented, and these results suggest that the approach has considerable merit for musical-instrument recognition and considerable potential for sound-source recognition in general.

7.2 Future developments

It goes without saying that there are many ways in which the work presented here could be extended or improved. Not all of the goals set out at the beginning of this undertaking have been met, and many portions of the implementation were developed only far enough to see how the system as a whole might behave. Some of the possibilities for future development of the work include:

- **Integration with a CASA framework.** The system described here was purposely based on the representations used in Ellis's PDCASA architecture (Ellis, 1996), which I view as the most promising line of current research in computational auditory scene analysis. As was described in Chapter 1, sound-source recognition is only useful insofar as it allows the perceiver to infer properties of the sound source. Ellis's micro-taxonomy of noise cloud, transient, and quasi-periodic tonal elements is an example of the way recognition can be used at a very low level to improve the interpretation of mixtures of sounds. By extending the taxonomy to include more elaborate sound-source models such as those discussed here, CASA systems may someday be better equipped to understand more complicated mixtures. This integration will by no means be a trivial step.
- **Addition of multiple, overlapping taxonomies.** The system described here employs a single taxonomy as its knowledge base. In contrast, the organization of knowledge in the human brain is much more complicated. Perhaps many different taxonomies are superposed over the same set of object classes, organizing them according to different principles. It is not at all obvious how a recognition system based on multiple, overlapping taxonomies might operate. Perhaps one or another is selected according to the problem at hand. Or perhaps one taxonomy might be chosen in a given situation because of the particular feature set that is available. Perhaps taxonomic structures are too rigid altogether—other, more general models could be based on spreading activation (Maes, 1989) or something like Hofstadter's Slip Net (Hofstadter, 1995).
- **Integrating more general learning techniques.** When I began this work, my goal was to build a system that would not require explicit training. I envisioned a system that could listen to real music performances and determine for itself what features were important and what the relevant classes of sounds are. Over time, I gradually whittled this vision down to the system presented in the preceding chapters. There are, however, many interesting ways that machine learning techniques could be applied to the problem of sound-source recognition. For example, it would be interesting to have the system form its own taxonomy rather than have one specified in advance. Perhaps Bobick's techniques for evaluating the usefulness of particular categorizations (Bobick, 1987) could be spun into a method for generating and refining taxonomies, or maybe other statistical clustering techniques could be used. Perhaps the system could start with a few supervised training examples, build preliminary representations, and then refine them by trying to recognize unlabeled sounds in an unsupervised framework. Another interesting direction is multi-modal integration. In particular, there may be ways in

which visual and auditory object recognition systems could help each other learn more robustly and quickly than either could do on its own.

- **Extending the knowledge base to include other kinds of sound sources.** In Chapter 3, I concluded that musical instruments must be recognized on the basis of features arising from the excitation and resonance structures of the instruments. This may also be true of a much wider range of sound sources. For example, vowels in human speech appear to be identified on the basis of vocal-tract resonances, or *formants* (e.g., Peterson & Barney, 1952). Also, the distinction between “bouncing” and “breaking” events appears to be due in large part to the excitation structure of the events—in particular their temporal properties (Warren & Verbrugge, 1984). Many of the features used in the system presented here could be useful for recognizing the sources of pitched sounds in addition to the orchestral instruments (one promising set of possibilities is animal vocalizations). Of course, in order to extend the work to other kinds of sound sources, new features would have to be added to the system’s repertoire. Happily, the conceptual division of the sound source into excitation and resonance is a useful tool for guiding the search for new features, and the architecture described here is sufficiently flexible for new features to be added as they are discovered.
- **Using model alignment to improve early decisions.** In the visual object recognition literature, model alignment is an obvious and important aspect of the recognition process (e.g., Ullman, 1996). In order to compare local features of a model to sensory data, there must first be a stage of rough positioning or alignment to determine the correspondence of portions of the perceptual data to parts of the model. It is not as obvious that such a step is important in audition, but I believe that classification at upper (more abstract) levels of a taxonomy could be improved greatly by some form of model alignment. Consider, for example, the brass and string families, for which each instrument is—to a first approximation—a scaled version of a single prototype. The changes in scale from one instrument to another shift many feature properties—including the spectral centroid, pitch range, and cutoff frequencies—uniformly. By taking these shifts into account, abstract prototypes could become much better predictors of unobserved features, and high-level classifiers could be made much more robust, thereby alleviating the need for techniques like beam search. In addition, this could enable the system’s performance to become more like that of expert human listeners, who rarely confuse instruments from different families.
- **Taking advantage of inheritance.** One of the most important conceptual strengths of frame-based semantic networks (Minsky, 1974) is that slots in some frames can inherit default values from other frames. Within a taxonomy, the inheritance structure is obvious: a node’s slots inherit default values from the node’s ancestors unless they are overridden by evidence from training data. This style of inheritance is related to the statistical technique termed *shrinkage*, which has been used to advantage in text-document classification tasks (McCallum et al., 1998), and to *deleted interpolation*, which has been used in speech recognition systems (Jelinek & Mercer, 1980). The basic idea is that, instead of using a single probability model for each feature

based only on training data applicable to a particular node of the hierarchy, the system forms a *mixture model* based on the probability models at the node and all of its ancestors. The intuitive reason for using this kind of technique is that it improves estimates of probability-model parameters that would otherwise be uncertain due to limited amounts of training data. Empirical results show that the technique improves classifier performance, with the biggest improvement occurring when training data is sparse (McCallum et al., 1998).

- **Considering "cognitive" cues.** Many of the features experienced listeners use to recognize sound sources are not related directly to the acoustics of sound production. High-level contextual cues, such as the particular piece of music being played, can be used to zero in on the particular instrument being heard. Similarly, particular phrasing styles (e.g., portamento in bowed string or vocal performance) can be emblematic of particular instrument classes or performers. As another example, human speakers may have characteristic speaking rhythms or pitch contours. There are so many possibilities that small systems like the one described in this dissertation may never be able to compete with humans on recognition tasks using real-world sounds. Systems may require vast degrees of experience (equivalent to years of listening)—and orders of magnitude more feature detectors and memory—to compete directly with human listeners.
- **Using multiple prototypes for each sound-source category.** The classification system described in this dissertation employed a single prototype for each sound-source category, and an obvious extension is to use multiple prototypes, particularly for the categories that vary the most from instance to instance. Systems that take this approach will need to carefully balance the additional processing requirements against the degree of improved performance that the approach may provide.
- **Constructing better feature-salience estimates**. Because the set of sound sources explored in this dissertation was relatively small, the extensions to the basic classification architecture proposed in Section 5.3 were not adequately explored. The results of Computer experiments #2 and #3 (Sections 6.4 and 6.5) suggest that feature selection based on local estimates of discriminating power do improve performance, but the *ad hoc* estimates of measurement reliability did not help. This is not to say, however, that reliability estimates are not a promising avenue for future research, but only that the issues involved are subtle and worthy of more extensive investigation.
- **Improving the feature detectors (and adding additional features).** It should be obvious from the presentation in Chapter 4 that many of the signal processing techniques used to extract features from the correlogram representation were invented in an *ad hoc* manner. Many could be made more robust by more thorough analysis of the properties of the signals being analyzed, or by more principled statistical approaches. In addition, the repertoire of feature detectors could of course be expanded to include more of the range known to be important to human perception. In particular, I believe that note-to-note transitions are the single most promising enexplored feature for musical-instrument recognition. However, so little work has been

done to explore these features (see Strawn, 1985, 1986, 1987, for some initial analyses), that it is difficult to know where to begin.

- **Providing more training data**. I had hoped that a good set of features would enable the recognition system to generalize from very little training data, and it is conceivable that the right features could make this possible. Although it would be interesting to see how the system's performance would improve if more labeled examples were provided to it, I do not view this as one of the more interesting paths to explore. At this stage, I believe that exploring a wider range of sound sources, another set of features, or alternate recognition algorithms could yield more insight than such a brute-force approach.
- **Improving the system's efficiency.** The system's current implementation is painfully slow. The front end, which is implemented in C++, runs at about ten times real time on a desktop PC (mine is a 150 MHz Pentium). The recognition algorithm is implemented in MATLAB and is even slower. Although it would probably be possible to implement a real-time front end with technology available today, I do not believe that it would be a useful exercise. Much more work has to be done to develop the recognition framework—particularly in regard to how the recognition process evolves over time—before it would be worth attempting to build a real-time system.

7.3 Insights gained

In this dissertation, I have described a computer model based on portions of a new theory of sound-source recognition. Although many parts of the implementation were exploratory (and certainly sub-optimal), several key insights can be gained from this work. For example:

- **Serious consideration of psychoacoustics can lead to successful computer perception systems.** The recognition system described here was engineered rather than "hill-climbed." Instead of blindly applying a black-box pattern-recognition technique to a fully general—but not interpretable—feature set (as is done by many purveyors of artificial neural network techniques), I purposely discarded much of the information in the acoustic signal based on evidence that it is not used by human listeners. The human sense of hearing is more robust than current machine systems, and we have much to learn as engineers and as scientists by carefully studying human perception.
- **"Timbre" is useless as a scientific concept.** There is no fixed set of parameters that determine what something "sounds like," any more than there is for what something "looks like." There are infinitely many ways to describe objects in the world, and worrying about holistic concepts such as timbre or appearance is a waste of time.
- **Introspection is misleading.** Previous research in auditory perception—particularly in computational auditory scene analysis—has in general vastly underestimated the ubiquitous nature of the perceptual illusions our brains create. Our perceptual systems are incredibly robust, and we are constantly deluded into believing that we perceive more than is actually there to be dis-

cerned. When we "hear out" the guitar solo in a pop song, we do not do so by "separating out" the waveform generated by the guitar. We do it by subconsciously making huge inferential leaps to fill in the perceptual gaps created by competing sounds. I rather strongly believe that the only reason our brains can fool us so well is that we are unknowingly making extensive use of contextual information and background knowledge.

- **Resynthesis is not a necessary component of a successful computer listening system.** It disturbs me greatly to see how much emphasis is placed on using computational auditory scene analysis systems to "extract" sounds from mixtures and resynthesize them as isolated components. The human auditory system surely does not do this, so why should computer models of the human system? Even if the human auditory system *could* perform this task, what would be the point?—who would listen to the result? This is a version of the *homunculus paradox*. The solution to the paradox in this case is to realize that the system can only progress toward its goal—which is to make sense of objects in the world and their interactions—by abstracting away from the acoustic signal to a point where aspects of the sound can be related to prior experience. To be sure, we do not know exactly how this happens in the human brain, but what would be the point of re-representing the world at the same level of abstraction? My best guess is that abstract auditory representations refer to the low-level sensory data for support of hypotheses about mixtures of sounds; there is no need to separate their contributions explicitly, and there certainly is no need for resynthesis.

7.4 Conclusions

The theory of sound-source recognition outlined in Chapter 1 is necessarily vague and should probably be viewed mainly as a collection of constraints that will need to be part of a more developed theory. There are many possible recognition systems that would be consistent with the general theory I have proposed; the particular implementation described here is but one.

To my knowledge, this theory is the first of its kind. Many of its components can be found in the computer vision and cognitive science literature, and parts of it are hinted at by Bregman's *Auditory Scene Analysis*, but this particular assemblage of ideas is new to hearing science, and it is my hope that I have provided a viable jumping-off point for future research in this area. Our current scientific understanding of perception is so limited that we do not even know all of the right questions to ask of a perceptual theory. It is encouraging, however, that the approach I have described has yielded such promising initial results. Sound-source recognition remains a promising avenue for future research—one that will eventually lead to a deeper understanding of audition in general.

References

American National Standards Institute (1973). *American national psychoacoustical terminology*: New York: American Standards Association. (As cited by Houtsma, 1997)

American Standards Association (1960). American Standard Acoustical Terminology. Definition 12.9, Timbre, p. 45. New York.

Beauchamp, J. W. (1974). Time-variant spectra of violin tones. *J. Acoust. Soc. Am.* **56**(3), 995-1004.

Beauchamp, J. W. (1982). Synthesis by spectral amplitude and “brightness” matching of analyzed musical instrument tones. *J. Audio Eng. Soc.* **30**(6), 396-406.

Beauchamp, J. W. (1993). Unix workstation software for analysis, graphics, modification, and synthesis of musical sounds. *Audio Engineering Society Preprint* 3479, L1-7.

Benade, A. H. (1990). *Fundamentals of Musical Acoustics*. New York: Dover.

Berger, K. W. (1964). Some factors in the recognition of timbre. *J. Acoust. Soc. Am.* **36**, 1888-1891.

Bobick, A. & Richards, W. (1986). *Classifying Objects from Visual Information*. Massachusetts Institute of Technology A.I. Memo No. 879.

Bobick, A. F. (1987). *Natural Object Categorization*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Bourne, J. B. (1972). *Musical Timbre Recognition Based on a Model of the Auditory System*. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA.

Boyk, J. (1997). There's life above 20 kilohertz! A survey of musical instrument spectra to 102.4 kHz. http://www.cco.caltech.edu/~boyk/spectra/spectra.htm

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

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth Intl. Group.

Brown, G. J. (1992). *Computational Auditory Scene Analysis: A Representational Approach*. Ph.D. thesis, Univeristy of Sheffield.

Brown, G. J. & Cooke, M. (1994). Perceptual grouping of musical sounds: A computational model. *Journal of New Music Research* **23**, 107-132.

Brown, J. C. (1996). Frequency ratios of spectral components of musical sounds. *J. Acoust. Soc. Am.* **99**(2), 1210-1218.

Brown, J. C. (1997a). Computer identification of musical instruments using pattern recognition. *Presented at the 1997 Conference of the Society for Music Perception and Cognition,* Cambridge, MA.

Brown, J. C. (1997b). Cluster-based probability model for musical instrument identification. *J. Acoust. Soc. Am.* **101**, 3167 (abstract only).

Brown, J. C. (1998a). Personal communication.

Brown, J. C. (1998b). Computer identification of wind instruments using cepstral coefficients. *J. Acoust. Soc. Am.* **103**, 2967 (abstract only).

Brown, J. C. (1998c). Computer identification of wind instruments using cepstral coefficients. In *Proceedings of the 16th International Concress on Acoustics and 135th Meeting of the Acoustical Society of America* (pp. 1889-1890). Seattle.

Brown, J. C. (1999). Musical instrument identification using pattern recognition with cepstral coefficients as features. *J. Acoust. Soc. Am.* **105**(3) 1933-1941.

Brown, R. (1981). An experimental study of the relative importance of acoustic parameters for auditory speaker recognition. *Language and Speech* **24**, 295-310.

Campbell, W. & Heller, J. (1979). Convergence procedurces for investigating music listening tasks. *Bulletin of the Council for Research in Music Education* **59**, 18-23 (As cited by Kendall, 1986).

Campbell, W. C. & Heller, J. J. (1978). The contribution of the legato transient to instrument identification. In E. P. A. Jr. (ed.) *Proceedings of the research symposium on the psychology and acoustics of music* (pp. 30-44). University of Kansas, Lawrence. (As cited by Kendall, 1986)

Cariani, P. A. & Delgutte, B. (1996a). Neural correlates of the pitch of complex tones. I. Pitch and pitch salience. *Journal of Neurophysiology* **76**, 1698-1716.

Cariani, P. A. & Delgutte, B. (1996b). Neural correlates of the pitch of complex tones. II. Pitch shift, pitch ambiguity, phase invariance, pitch circularity, rate pitch, and the dominance region for pitch. *Journal of Neurophysiology* **76**, 1717-1734.

Casey, M. A. (1996). Multi-Model Estimation and Classification as a Basis for Computational Timbre Understanding. Unpublished manuscript.

Charbonneau, G. R. (1981). Timbre and the perceptual effects of three types of data reduction. *Computer Music Journal* **5**(2), 10-19.

Clark, M., Luce, D., Abrams, R., Schlossberg, H. & Rome, J. (1963). Preliminary experiments on the aural significance of parts of tones of orchestral instruments and on choral tones. *J. Audio Eng. Soc.* **11**(1), 45-54.

Cooke, M. (1993). *Modelling Auditory Processing and Organisation*. Cambridge: Cambridge University Press.

Cosi, P., De Poli, G. & Lauzzana, G. (1994a). Auditory modelling and self-organizing neural networks for timbre classification. *Journal of New Music Research* **23**, 71-98.

Cosi, P., De Poli, G. & Lauzzana, G. (1994b). Timbre classification by NN and auditory modeling. In *Proceedings of the International Conference on Artificial Neural Networks*.

Cosi, P., De Poli, G. & Prandoni, P. (1994c). Timbre characterization with Mel-Cepstrum and neural nets. In *Proceedings of the 1994 International Computer Music Conference* (pp. 42-45).

Crummer, G. C., Walton, J. P., Wayman, J. W., Hantz, E. C. & Frisina, R. D. (1994). Neural processing of musical timbre by musicians, nonmusicians, and musicians possessing absolute pitch. *J. Acoust. Soc. Am.* **95**(5), 2720-2727.

Dawant, B. & Jansen, B. (1991). Coupling numerical and symbolic methods for signal interpretation. *IEEE transactions on Systems, Man Cybernetics* **21(1)**, 115-124.

De Poli, G. & Prandoni, P. (1997). Sonological models for timbre characterization. *Journal of New Music Research* **26**, 170-197.

De Poli, G. & Tonella, P. (1993). Self-organizing neural networks and Grey's timbre space. In *Proceedings of the 1993 International Computer Music Conference* (pp. 441-444).

Dennett, D. (1991). *Consciousness Explained*. Boston: Little, Brown.

Dillon, H. (1981). *The Perception of Musical Instruments*. (Doctoral dissertation, University of New South Wales Australia, 1979). *Dissertation Abstracts International* **41**, 2703B-2704B (As cited by Kendall, 1986).

Dubnov, S. & Rodet, X. (1998). Timbre recognition with combined stationary and temporal features. In *Proceedings of the 1998 International Computer Music Conference* (pp. 102-108).

Duda, R. O., Hart, P. E. & Stork, D. G. (1997). *Pattern Classification*. Wiley (draft manuscript only).

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

Eagleson, H. V. & Eagleson, O. W. (1947). Identification of musical instruments when heard directly and over a public-address system. *J. Acoust. Soc. Am.* **19**(2), 338-342.

Elliott, C. (1975). Attacks and releases as factors in instrument identification. *Journal of Research in Music Education* **23**, 35-40 (As cited by Kendall, 1986).

Ellis, D. & Rosenthal, D. (1995). Mid-level representations for computational auditory scene analysis. In *Proceedings of the International Joint Conference on Artificial Intelligence workshop on Computational Auditory Scene Analysis.*

Ellis, D. P. W. (1994). A computer implementation of psychoacoustic grouping rules. In *Proceedings of the 12th Intl. Conf. on Pattern Recognition*. Jerusalem.

Ellis, D. P. W. (1996). *Prediction-driven computational auditory scene analysis*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Erickson, R. (1975). *Sound Structure in Music*. Berkeley: University of California Press.

Feiten, B. & Gunzel, S. (1994). Automatic indexing of a sound database using self-organizing neural nets. *Computer Music Journal* **18**(3), 53-65.

Fletcher, H. (1964). Normal vibration frequencies of a stiff piano string. *J. Acoust. Soc. Am.* **36**, 203-209.

Fletcher, H., Blackham, E. D. & Stratton, R. (1962). Quality of piano tones. *J. Acoust. Soc. Am.* **34**(6), 749-761.

Fletcher, H. & Sanders, L. C. (1967). Quality of violin vibrato tones. *J. Acoust. Soc. Am.* **41**(6), 1534-1544.

Fletcher, N. H. & Rossing, T. D. (1998). *The Physics of Musical Instruments*. New York: Springer.

Foote, J. (1997). A similarity measure for automatic audio classification. In *Proceedings of the 1997 AAAI 97 Spring Symp. Intelligent Integration and Use of Text, Image, Video and Audio* (SS-97-03). AAAI Press.

Fourier, J. B. J. (1822). *[The Analytical Theory of Heat]*.

Freed, D. J. (1990). Auditory correlates of perceived mallet hardness for a set of recorded percussive sound events. *J. Acoust. Soc. Am.* **87**(1), 311-322.

Freedman, M. D. (1967). Analysis of musical instrument tones. *J. Acoust. Soc. Am.* **41**, 793-806.

Fujinaga, I. (1998). Machine recognition of timbre using steady-state tone of acoustic musical instruments. In *Proceedings of the 1998 International Computer Music Conference* (pp. 207-210).

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. New York: Houghton Mifflin.

Gordon, J. W. (1984). *Perception of Attack Transients in Musical Tones*. Ph.D. thesis, Stanford University.

Gordon, J. W. & Grey, J. M. (1978). Perception of spectral modifications on orchestral instrument tones. *Computer Music Journal* **2**(1), 24-31.

Grey, J. M. (1975). *An Exploration of Musical Timbre*. Ph.D. thesis, Stanford University.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *J. Acoust. Soc. Am.* **61**(5), 1270-1277.

Grey, J. M. (1978). Timbre discrimination in musical patterns. *J. Acoust. Soc. Am.* **64**(2), 467-472.

Grey, J. M. & Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. *J. Acoust. Soc. Am.* **63**(5), 1493-1500.

Grey, J. M. & Moorer, J. A. (1977). Perceptual evaluations of synthesized musical instrument tones. *J. Acoust. Soc. Am.* **62**, 454-462.

Grimson, W. E. L. (1990). *Object Recognition by Computer: The Role of Geometric Constraints*. Cambridge: MIT Press.

Hajda, J. M., Kendall, R. A. & Carterette, E. C. (1994). Perceptual and acoustical analyses of impulse tones. In I. Deliege (ed.) *Proceedings of the 4th International Conference on Music Perception and Cognition* (pp. 315-316).

Hajda, J. M., Kendall, R. A., Carterette, E. C. & Harshberger, M. L. (1997). Methodological issues in timbre research. In I. Deliege & J. Sloboda (eds.), *Perception and Cognition of Music*. Psychology Press, East Essex, UK.

Han, K.-P., Park, Y.-S., Jeon, S.-G., Lee, G.-C. & Ha, Y.-H. (1998). Genre classification system of TV sound signals based on a spectrogram analysis. *IEEE Trans. on Cons. Elect.* **44**(1), 33-42.

Handel, S. (1989). *Listening*. Cambridge: MIT Press.

Handel, S. (1995). Timbre perception and auditory object identification. In B. C. J. Moore (ed.) *Hearing*. New York: Academic Press.

Hawley, M. J. (1993). *Structure out of Sound*. Ph.D. thesis, Massachusetts Institute of Technology, Program in Media Arts and Sciences, Cambridge MA.

Helmholtz, H. (1954). *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. (A.J. Ellis, translator, original work published in 1885, Second English ed.) New York: Dover.

Hewitt, M. J. & Meddis, R. (1991). An evaluation of eight computer models of mammalian inner hair-cell function. *J. Acoust. Soc. Am.* **90**(2), 904-917.

Hofstadter, D. (1995). *Fluid Concepts and Creative Analogies*. New York: Basic Books.

Houtsma, A. J. M. (1997). Pitch and timbre: Definition, meaning and use. *Journal of New Music Research* **26**, 104-115.

Hutchins, C. M. (1998). The air and wood modes of the violin. *J. Audio Eng. Soc.* **46**(9), 751-765.

Iverson, P. & Krumhansl, C. L. (1993). Isolating the dynamic attributes of musical timbre. *J. Acoust. Soc. Am.* **94**(5), 2595-2603.

Jelinek, F. & Mercer, R. (1980). Interpolated estimation of Markov source parameters from sparse data. In S. Gelsema & L. N. Kanal (eds.), *Pattern Recognition in Practice* (pp. 381-402).

Kaminskyj, I. & Materka, A. (1995). Automatic source identification of monophonic musical instrument sounds. In *Proceedings of the 1995 IEEE International Conference on Neural Networks* (pp. 189-194).

Kashino, K. & Murase, H. (1997). Sound source identification for ensemble music based on the music stream extraction. In *Proceedings of the 1997 International Joint Conference on Artificial Intelligence.*

Kashino, K. & Murase, H. (1998). Music recognition using note transition context. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Seattle.

Kashino, K., Nakadai, K., Kinoshita, T. & Tanaka, H. (1995). Application of Bayesian probability network to music scene analysis. In *Proceedings of the 1995 International Joint Conference on Artificial Intelligence.*

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, Faculty of Engineering, University of Tokyo.

Kashino, K. & Tanaka, H. (1993). A sound source separation system with the ability of automatic tone modeling. In *Proceedings of the 1993 International Computer Music Conference.*

Kendall, R. A. (1986). The role of acoustic signal partitions in listener categorization of muscial phrase. *Music Perception* **4**(2), 185-214.

Kendall, R. A. & Carterette, E. C. (1991). Perceptual scaling of simultaneous wind instrument timbres. *Music Perception* **8**, 369-404.

Kendall, R. A., Carterette, E. C. & Hajda, J. M. (1994). Comparative perceptual and acoustical analyses of natural and synthesized continuant timbres. In I. Deliege (ed.) *Proceedings of the 3rd International Conference for Music Perception and Cognition* (pp. 317-318).

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

Krumhansl, C. L. (1989). Why is musical timbre so hard to understand? In S. Nielzen & O. Olsson (eds.), *Structure and perception of electroacoustic sound and music: Proceedings of the Marcus Wallenberg symposium held in Lund, Sweden, on 21-28 August 1988* (pp. 43-53). Netherlands: Excerpta Medica.

Krumhansl, C. L. & Iverson, P. (1992). Perceptual interactions between musical pitch and timbre. *J. Exp. Psych.* **18**(3), 739-751.

Laakso, T. I., Valimaki, V., Karjalainen, M. & Laine, U. K. (1996). Splitting the unit delay. *IEEE Signal Processing Magazine* **13**(1), 30-60.

Langmead, C. J. (1995a). Sound analysis, comparison and modification based on a perceptual model of timbre. In *Proceedings of the 1995 International Computer Music Conference.*

Langmead, C. J. (1995b). *A Theoretical Model of Timbre Perception Based on Morphological Representations of Time-Varying Spectra*. Master's thesis, Dartmouth College.

Li, X., Logan, R. J. & Pastore, R. E. (1991). Perception of acoustic source characteristics: Walking sounds. *J. Acoust. Soc. Am.* **90**, 3036-3049.

Lichte, W. H. (1941). Attributes of complex tones. *J. Experim. Psychol.* **28**, 455-481.

Licklider, J. C. R. (1951). A duplex theory of pitch perception. *Experientia* **7**, 128-133.

Luce, D. (1963). *Physical Correlates of Nonpercussive Musical Instrument Tones*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Maes, P. (1989). How to do the right thing. *Connection Science* **1**(3), 291-323.

Mammone, R., Zhang, X. & Ramachandran, R. P. (1996). Robust speaker recognition: A feature-based approach. *IEEE Signal Processing Magazine* **13**(5), 58-71.

Marques, J. (1999). *An Automatic Annotation System for Audio Data Containing Music*. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: Freeman.

Martin, K. D. & Kim, Y. E. (1998). Musical instrument identification: a pattern-recognition approach. *Presented at the 136th meeting of the Acoustical Society of America.* Available at: http://sound.media.mit.edu/papers.html.

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.

Mathews, M. V., Miller, J. E., Pierce, J. R. & Tenney, J. (1966). Computer study of violin tones. Bell Telephone Laboratories Technical Report, Murray Hill, NJ.

McAdams, S. (1993). Recognition of sound sources and events. In *Thinking in Sound: the Cognitive Psychology of Human Audition* (pp. 146-198). Oxford University Press.

McAdams, S. & Cunible, J.-C. (1992). Perception of timbral analogies. *Phil. Trans. R. Soc. Lond. B* **336**, 383-389.

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G. & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychol. Res.* **58**(3), 177-192.

McCallum, A., Rosenfld, R., Mitchell, T. & Ng, A. Y. (1998). Improving text classification by shrinkage in a hierarchy of classes. In *Proceedings of the 1998 ICML*.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York, NY: Wiley Interscience.

Meddis, R. & Hewitt, M. J. (1991a). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *J. Acoust. Soc. Am*. **89**, 2866-2882.

Meddis, R. & Hewitt, M. J. (1991b). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. II: Phase sensitivity. *J. Acoust. Soc. Am.* **89**, 2883-2894.

Mellinger, D. K. (1991). *Event Formation and Separation in Musical Sound*. Ph.D. thesis, Stanford University.

Milios, E. E. & Nawab, S. H. (1992). Signal abstraction concept for signal interpretation. In *Symbolic and Knowledge-Based Signal Processing*. Prentice Hall.

Milner, P. (1963). *Interrelationships Between the Timbre and the Intensity of Musical Instruments*. Bachelor's thesis, Massachusetts Institute of Technology Cambridge, MA.

Minami, K., Akutsu, S., Hamada, H. & Tonomura, Y. (1998). Video handling with music and speech detection. *IEEE Multimedia* **5**(3), 17-25.

Minsky, M. (1974). *A Framework for Representing Knowledge*. Massachusetts Institute of Technology AI Lab Memo #306.

Minsky, M. (1986). *The Society of Mind*. New York: Simon & Schuster.

Moore, B. C. J. (1989). *Introduction to the Psychology of Hearing*. London: Academic Press.

Moore, B. C. J. & Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *J. Acoust. Soc. Am.* **74**(3), 750-753.

Moorer, J. A. & Grey, J. M. (1977). Lexicon of analyzed tones - Part I: A violin tone. *Computer Music Journal* **1**(2), 39-45.

MPEG Requirements Group (1999). MPEG-7: Context, objectives, and technical roadmap. Doc. ISO/IEC JTC1/SC29/WG11/N2729, MPEG Seoul Meeting.

Nakatani, T., Kashino, K. & Okuno, H. G. (1997). Integration of speech stream and music stream segregations based on a sound ontology. In *Proceedings of the 1997 International Joint Conference on Artificial Intelligence.*

Nooralahiyan, A. Y., Kirby, H. R. & McKeown, D. (1998). Vehicle classification by acoustic signature. *Mathl. Comput. Modelling* **27**(9-11), 205-214.

Opolko, F. & Wapnick, J. (1987). McGill University Master Samples [Compact disc], Montreal, Quebec: McGill Univeristy.

Patterson, R. D. & Holdsworth, J. (1990). A functional model of neural activity patterns and auditory images. In W. A. Ainsworth (ed.) *Advances in speech, hearing and language processing.* London: JAI Press.

Patterson, R. D. & Moore, B. C. J. (1986). Auditory filters and excitation patterns as representations of frequency resolution. In B. C. J. Moore (ed.) *Frequency Selectivity in Hearing.* London: Academic.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kauffman.

Peterson, G. E. & Barney, H. L. (1952). Control methods used in a study of the vowels. *J. Acoust. Soc. Am.* **24**(2), 175-184.

Pfeiffer, S., Fischer, S. & Effelsberg, W. (1996). Automatic audio content analysis. Universität Mannheim Technical Report, Mannheim, Germany.

Pickles, J. O. (1988). *Introduction to the Physiology of Hearing*. Academic Press.

Plomp, R. (1970). Timbre as a multidimensional attribute of complex tones. In R. Plomp & G. G. Smoorenburn (eds.), *Frequency Analysis and Periodicity Detection in Hearing.* A.W. Sijthoff, Leiden.

Plomp, R. (1976). *Aspects of Tone Sensation.* London: Academic Press.

Plomp, R., Pols, L. C. W. & Geer, J. P. v. d. (1967). Dimensional analysis of vowel spectra. *J. Acoust. Soc. Am.* **41**(3), 707-712.

Popper, A. N. & Fay, R. R. (1997). Evolution of the ear and hearing: Issues and questions. *Brain, Behaviour and Evolution* **50**, 213-221.

Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication* **17**, 91-108.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Risset, J. C. (1966). Computer study of trumpet tones. Bell Laboratories Technical Report, Murray Hill, New Jersey.

Risset, J.-C. & Wessel, D. L. (1982). Exploration of timbre by analysis and synthesis. In D. Deutsch (ed.) *The Psychology of Music* (pp. 26-58). New York: Academic.

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

Robertson, P. T. (1961). *The Aurally Perceptual Basis for the Classification of Musical Instruments by Families*. Bachelor's thesis, Massachusetts Institute of Technology, Cambridge, MA.

Robinson, K. & Patterson, R. D. (1995). The duration required to identify the instrument, the octave, or the pitch chroma of a musical note. *Music Perception* **13**, 1-15.

Roederer, J. G. (1973). *Introduction to the Physics and Psychophysics of Music*. New York: Springer-Verlag.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (eds.), *Cognition and Categorization*. Hillsdale, NJ: Lawrence Erlbaum.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M. & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology* **8**, 382-439.

Rossing, T. D. (1990). *The Science of Sound.* Reading: Addison-Wesley.

Saint-Arnaud, N. (1995). *Classification of Sound Textures.* Master’s thesis, Massachusetts Institute of Technology, Cambridge, MA.

Saldanha, E. L. & Corso, J. F. (1964). Timbre cues and the identification of musical instruments. *J. Acoust. Soc. Am.* **36**, 2021-2026.

Sandell, G. J. & Chronopoulos, M. (1996). Identifying musical instruments from multiple versus single notes. *J. Acoust. Soc. Am.* **100**, 2752 (abstract only).

Sandell, G. J. & Chronopoulos, M. (1997). Perceptual constancy of musical instrument timbres; generalizing timbre knowledge across registers. In A. Gabrielsson (ed.) *Proceedings of the Third Triennial ESCOM Conference* (pp. 222-227).

Sasaki (1980). Sound restoration and temporal localization of noise in speech and music sounds. *Tohoku Psychologica Folia* **39**, 79-88 (As cited by Warren, 1999).

Sayre, K. M. (1965). *Recognition: A Study in the Philosophy of Artificial Intelligence*. Notre Dame: University of Notre Dame Press.

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

Schloss, W. A. (1985). *On the Automatic Transcription of Percussive Music - From Acoustic Signal to High-Level Analysis*. Ph.D. thesis, Stanford University

Schlossberg, H. R. (1960). *The Relative Importance of Transients and Steady States in the Recognition of Musical Instruments from their Tones*. Bachelor's thesis, Massachusetts Institute of Technology, Cambridge, MA.

Serra, X. (1989). *A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic Plus Stochastic Decomposition*. Ph.D. thesis, Stanford University.

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

Slaney, M. (1993). An efficient implementation of the Patterson-Holdsworth auditory filter bank. Apple Computer Technical Report #35.

Slaney, M. & Lyon, R. F. (1990). A perceptual pitch detector. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 357-360).

Slaney, M. & Lyon, R. F. (1993). On the importance of time - a temporal representation of sound. In M. Cooke, S. Beet & M. Crawford (eds.), *Visual Representations of Speech Signals.* John Wiley & Sons.

Spina, M. & Zue, V. (1996). Automatic transcription of general audio data: Preliminary analyses. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 594-597).

Strawn, J. (1985). *Modeling Musical Transitions*. Ph.D. thesis, Stanford University.

Strawn, J. (1986). Orchestral instruments: Analysis of performed transitions. *J. Audio Eng. Soc.* **34**(11), 867-880.

Strawn, J. (1987). Analysis and synthesis of musical transitions using the discrete short-time Fourier transform. *J. Audio Eng. Soc.* **35**(1/2), 3-13.

Strong, W. & Clark, M. (1967). Synthesis of wind-instrument tones. *J. Acoust. Soc. Am.* **41**(1), 39-52.

Strong, W. J. (1963). *Synthesis and Recognition Characteristics of Wind Instrument Tones*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Stumpf, C. (1926). *Die Sprachlaute*. Berlin: Springer-Verlag (As cited by Kendall, 1986).

Thayer, R. (1972). The effect of the attack transient on aural recognition of instrumental timbre. In J. Heller & W. Campbell (eds.), *Computer analysis of the auditory characteristics of musical performance* (pp. 80-101). Final Report (Project No. 9-0564A), U.S. Department of Health, Education, and Welfare, Bureau of Research. (As cited by Kendall, 1986)

Therrien, C. W. (1989). *Decision, Estimation, and Classification.* New York: Wiley.

Ullman, S. (1996). *High-level Vision*. Cambridge: MIT Press.

van Dommelen, W. A. (1990). Acoustic parameters in human speaker recognition. *Language and Speech* **33**(3), 259-272.

Vercoe, B. L. (1984). The synthetic performer in the context of live performance. In *Proceedings of the 1984 International Computer Music Conference*. Paris.

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 International Computer Music Conference*. Burnaby BC, Canada.

Volodin, A. (1972). [The perception of transient processes in musical sounds]. *Voprosy Psikholgii* **18**(4), 51-60 (As cited by Kendall, 1986).

von Békésy, G. (1960). *Experiments in Hearing*. New York: McGraw Hill.

von Hornbostel, E. M. & Sachs, C. (1961). Classification of musical instruments. *Journal of the Galpin Society* **14**, 3-29.

Warren, H. & Verbrugge, R. R. (1984). Auditory perception of breaking and bouncing events: A case study in ecological acoustics. *J. Exp. Psychol: Hum. Percept. Perform.* **10**, 704-712.

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: Cambridge University Press.

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

Wedin, L. & Goude, G. (1972). Dimension analysis of the perception of instrumental timbre. *Scand. J. Psychol.* **13**, 228-240.

Wessel, D. L. (1983). Timbre space as a musical control structure. *Computer Music Journal* **3**(2), 45-52.

Winston, P. H. (1992). *Artificial Intelligence*. Reading, Massachusetts: Addison-Wesley.

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