CENTER FOR COMPUTER RESEARCH IN MUSIC AND ACOUSTICS
DECEMBER 1991

Department of Music
Report No. STAN-M-77

EVENT FORMATION AND SEPARATION OF MUSICAL SOUND

David K. Mellinger

CCRMA
DEPARTMENT OF MUSIC
Stanford University
Stanford, California 94305

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

© copyright 1991 by David K. Mellinger
All Rights Reserved

Abstract

This thesis reviews psychoacoustic and neurophysiological studies that show how the human auditory system is capable of hearing out one source of sound from the mixture of sounds that reaches the ear. A number of cues are used for identifying which parts of the spectrum originate with a single source: common onset, the beginning of sound energy at different frequencies at one time; harmonicity, the arrangement of the partials of a tone into a harmonic series; common frequency variation, the motion of partials in frequency at the same relative rate; common spatial location; and several others.

A multistage architecture is described for early auditory processing. After the input sound signal is transduced into a map of neural firings in the cochlea, filters extract the various cues for source separation from the cochlear image. The model uses these cues to group local features into single sound events and further groups events over time into sound sources.

The implemented model groups parts of the spectrum together over time to make separate sound events, using principles and constraints present in natural auditory systems. The model includes filters for detecting onsets and frequency variation in sound. These filters are tuned to work on musical sounds. Their output is used to find and separate notes in the signal, producing time-frequency images of the parts of the sound determined to belong to each event. This processing is applied to musical sounds made up of several notes played at a time, revealing the strengths and weaknesses of the computational model. The thesis offers directions for future work in computational auditory modelling.

Acknowledgements

My deepest thanks go to Bernard Mont-Reynaud. During the time I have been at Stanford, he has spent countless hours educating and guiding me, suggesting fruitful lines of work, listening to my sometimes misguided forays into the unknown, and generally keeping this project on track. Without him, this work would not have been possible.

I would also especially like to thank Earl Schubert, who has helped immensely in clarifying the nature of the auditory system to me and in suggesting interesting directions of study. He has also helped with the production of this thesis, giving invaluable advice at all levels of detail.

I also thank thank Roger Shepard for his insight into source separation in audition, and especially its relationship to larger perceptual questions. His lectures and demonstrations led me to think more deeply about many of the ideas discussed here, and talks with him have made this work stronger in many ways.

The model of early audition to be developed here has been informed and influenced by the participants in the weekly hearing seminar at CCRMA, the Center for Computer Research in Music and Acoustics. I would like to thank Malcolm Slaney for keeping the seminar interesting, especially during the last year when he has been running it. I wish to thank all of the participants in the seminar for interesting and lively discussions. In addition to Malcolm and my three committee members, those I would like to single out for special thanks are Dick Duda, Steve Esterly, John Pierce, Dick Lyon, Jerry Roberts, and Meg Withgott. I would also like to thank Jay M. Tenenbaum for initiating the meetings that led to the seminar.

To Malcolm Slaney I owe another special debt, for it was he who made available the program for Dick Lyon's ear model which underlies the computational model presented here.

Al Bregman has also been instrumental in furthering my understanding of source separation, and I would like to thank him as well. His book contains the following charge, to which this thesis is addressed:

> [W]e have now reached the point where we have a good appreciation of the many kinds of evidence that the human brain uses for partitioning sound, and it seems appropriate to begin to explore the formal patterns of computation by which the process could be accomplished.

Several other people at the CCRMA have helped me immensely over the years, and I thank them too: John Chowning, for making the center happen, for some interesting source-separation examples, and for arranging support when none other was available; Patte Wood and Heidi Kugler, for keeping CCRMA running; Bill, Schottstaedt, for his Common Music Notation system for producing musical scores; and Glen Diener and Jay Kadis, for keeping the computers running.

I am also grateful to the National Science Foundation and especially its grant IRI-8613574, which provided the funding for much of this work.

This work would also not be possible without personal support from several people. I thank Kristie her love and for keeping me going through times when it seemed like this project would never end, and Diana for her emotional support over the years as well. I also thank my mother and father for encouraging me to continue in school.

Finally, I would like to thank those who have been my best teachers, the ones who inspired me to learn and to keep learning: my mother, Jean Butler, and Don Buttermore. It is to you that this work is dedicated.

Contents

Abstract **iv**

Acknowledgements **v**

1 Introduction **1**

1.1 Domains for Sound Separation 1
1.2 Definition: Auditory Scene Analysis 3
1.3 Why Study Auditory Scene Analysis? 5
1.4 Perceptual Models 5
  1.4.1 Constructive Perception 6
  1.4.2 A Multi-Disciplinary Approach 6
  1.4.3 Marr's Levels of Description 8
  1.4.4 Overview of the Auditory Model 9
1.5 Previous Work 12
1.6 Scope of this Work 15
  1.6.1 What is Here 15
  1.6.2 What is Not Here 16
1.7 Structure of this Document 17

2 Architecture of the Early Auditory Model **18**

2.1 The Ear 18
  2.1.1 Outer and Middle Ear 18
  2.1.2 Cochlea 19
2.2 Feature Maps 20

2.2.1 Analog, Continuous Maps 20
2.2.2 Sampling 21
2.2.3 Locality of Processing 21
2.2.4 Spatialization of Time 22
2.2.5 Feature Maps in the Brain 23
2.2.6 Map Computation 25
2.2.7 Organizing Maps in Space 25
2.3 Autocorrelation 26
2.4 Partials 28
2.5 Features for Event Formation 28
2.6 Onset 31
2.7 Frequency Variation 35
2.8 Harmonicity 41
2.9 Amplitude Variation 43
2.10 Location 45

3 Computational Model 50
3.1 The Cochlear Model 51
3.2 Filtering Amplitude Onset 52
3.2.1 Results 66
3.3 Filtering Frequency Variation 72
3.3.1 Frequency Variation in the Cochleagram 74
3.3.2 Logarithmic Scale 74
3.3.3 Kernel Shape 77
3.3.4 Tuning the FV Kernel 83
3.3.5 Evaluation 93
3.4 Frequency Variation in the Correlogram 100
3.4.1 Cross-Correlation Output 106
3.4.2 Summing Across Lags 107
3.4.3 Lateral Inhibition 109
3.5 Comparison of FV Filters 112

3.5.1 Cochleagram 112
3.5.2 Correlogram 114

4 Event and Source Formation 120

4.1 Characteristics of Event and Source Formation 121
4.1.1 Hierarchical Perception 123
4.1.2 Competing Organizations 123
4.1.3 Allocation and Accounting 124
4.2 Natural Constraints 126
4.3 Event Formation 129
4.4 Affinity Groups 131
4.5 Source Formation 134
4.5.1 Pitch Separation 135
4.5.2 Timbre 136
4.5.3 Rate of Repetition 137
4.5.4 Number of Repetitions 137
4.5.5 Loudness 138
4.6 Summary of Event and Source Formation Mechanisms 139

5 Algorithm for Event Formation 141

5.1 Overview of Operation 142
5.1.1 Cycle of Processing 143
5.1.2 Affinity Groups 144
5.2 Tracking of Partials 145
5.2.1 Creation of Partials 145
5.2.2 Continuation of Partials 147
5.2.3 Merging of Partials 149
5.2.4 Diverging of Partials 151
5.2.5 Termination of Partials 155
5.3 Event Handling 156
5.3.1 Order Dependence 157
5.3.2 Event Continuation 157

5.3.3 Event Diverging 157
5.3.4 Event Merging 159
5.3.5 Event Termination 159
5.4 Onset and Frequency Variation Information 159
5.4.1 Onsets 160
5.4.2 Frequency Variation 162
5.5 Evaluation 163
5.6 Limitations of the Model 183
5.7 Neural Speculation 184
5.7.1 Labelling and Filtering 184
5.7.2 Cortical Oscillation 185

6 Summary and Conclusions **187**
6.1 Summary of the Model 187
6.2 Questions Revisited 187
6.3 Contributions 189
6.4 Comparison with Other Models 190
6.4.1 Weintraub 190
6.4.2 Cooke 191
6.5 Future Work 193
6.6 Conclusion 195

A Growable Triangular Matrices **197**

B Parameters **199**

C SoundExplorer **202**

Bibliography **209**

List of Tables

3.1 Summary of onset kernel characteristics. 72
3.2 Summary of FV kernel characteristics. 93
3.3 Analytical Q values for the leftmost few spots in a correlogram. . . . 117
3.4 Measured Q values for the leftmost few spots in a correlogram. 119

List of Figures

1.1 Auditory model block diagram. 10

2.1 Two-dimensional space map in the MLD of the owl. By permission from Knudsen. 24
2.2 Typical correlogram for a pitched sound. 27
2.3 Cochleagram display of saxophone and drum tones, showing partials. 29
2.4 Score for the previous cochleagram display (without drum notes). . . 30
2.5 Tone pattern in Pierce's demonstration. 31
2.6 Tones for the onset part of Bregman/Pinker experiment. 33
2.7 A tone rising continuously in frequency. 36

3.1 Cochleagram showing onsets at about 0, 120, 220, 340, 360, 470, and 580 ms. 53
3.2 Score for the previous figure. 53
3.3 A cochlear channel with two pronounced onsets. 54
3.4 Onset kernel. Horizontal axis is time, vertical axis is the function value. 54
3.5 An onset map for the Frescobaldi piano excerpt. 55
3.6 Onset kernels with rectangular (left) and exponential (right) functions. 56
3.7 Onset map computed with rectangular kernel. 57
3.8 Onset map computed with exponential kernel. 57
3.9 Onset map computed with 2.3 ms kernel. 58
3.10 Onset map computed with 4.5 ms kernel. 59
3.11 Onset map computed with 9.0 ms kernel. 59
3.12 Onset map computed with 18.1 ms kernel. 60

3.13 Onset kernels of different decay times. 61
3.14 Two tones with different vibrato. 62
3.15 Onset map computed with 18.1 ms kernel. 63
3.16 Kernel to eliminate frequency-variation artifacts. Horizontal axis is time, vertical is height. 64
3.17 Offset kernel. 66
3.18 Offset map for the Frescobaldi sound. 67
3.19 Cello (top) and bowed violin (bottom) onset responses. 69
3.20 French horn (top) and trumpet (bottom) onset responses. 70
3.21 Snare drum (top) and plucked violin (bottom) onset responses. 71
3.22 Two tones with different vibrato. 73
3.23 A correlation kernel for FV filtering. Time is the horizontal axis, height ($\log f$) the vertical. 75
3.24 A FV feature map, computed with the kernel of the previous figure, showing partials rising in frequency. 76
3.25 A typical center-surround function. 78
3.26 A partial aligned with the kernel axis. 78
3.27 A kernel for filtering descending-frequency partials. 79
3.28 A partial crossing a kernel. 80
3.29 A partial aligned with a kernel but off-axis. 80
3.30 A partial aligned with a kernel far off the axis. 81
3.31 Two partials crossing a kernels at different angles. 81
3.32 A Gabor function. 82
3.33 Positive-sum center-surround function. 83
3.34 FV filtering with the kernel above. 84
3.35 FV kernels of relatively long and short durations. 85
3.36 FV map computed from 23 ms kernel. 87
3.37 FV map computed from 57 ms kernel. 88
3.38 FV filtered at a rate of 2.5 octaves/s. 90
3.39 FV filtered at a rate of 1.25 octaves/s. 91
3.40 Two tones with different vibrato. 94

3.41 FV maps at -5 and -2.5 octaves/s. 95
3.42 FV maps at 0 and 2.5 octaves/s. 96
3.43 FV-map at 5 octaves/s. 96
3.44 The McAdams/Reynolds oboe sound. 97
3.45 The McAdams/Reynolds oboe, FV rate 0 octaves/s. 98
3.46 The McAdams/Reynolds oboe, FV rate 0.5 octaves/s. 99
3.47 Correlogram frames for lower (left) and higher pitched tones. 101
3.48 Two-D slice through the 3-D correlogram. 102
3.49 Spot motion in a correlogram. 103
3.50 0.5 octave/s filter output at maximum (left) and minimum response. . 107
3.51 Summed filter output for -0.5 octave/s FV rate. 108
3.52 Summed filter output for +0.5 octave/s FV rate. 109
3.53 Laterally-inhibited, summed filter output for -0.5 octave/s FV rate. . 111
3.54 Laterally-inhibited, summed filter output for +0.5 octave/s FV rate. . 111
3.55 A peak in a function, with f^+ and f^- marked. 113
3.56 Halfwave-rectified sine wave. 114
3.57 Computation of autocorrelation value for fixed l. 115
3.58 Autocorrelation function $C(l)$ (solid curve) with cosine wave (dotted curve) for comparison. 116

4.1 Tones in Bregman-Rudnicky experiment. 122
4.2 Exclusive allocation in Bregman's experiment. 125
4.3 What is the underlying pattern? (Used by permission from Bregman.) 128
4.4 The underlying pattern is clear. (Used by permission from Bregman.) 129
4.5 Masking experiment (after [Warren82]). 130
4.6 Continuity of partials. Used by permission from Bregman. 132
4.7 Pierce's delayed-onset example. 134
4.8 A different part of the Bregman-Rudnicky experiment. 136
4.9 Wessel's timbre illusion. 137

5.1 Destructive interference of partials. 148
5.2 Partial divergence phenomena. 152

5.3 Frescobaldi toccata played on a piano, and score. 164
5.4 Events 1 and 2 from the Frescobaldi toccata. 165
5.5 Events 3 and 4 from the Frescobaldi toccata. 165
5.6 Events 5 and 6 from the Frescobaldi toccata. 166
5.7 Event 7 from the Frescobaldi toccata. 166
5.8 Mixture of two notes with separate vibrato. 167
5.9 Event 1 of the two-note mixture. 168
5.10 Event 2 of the two-note mixture. 169
5.11 McAdams's oboe sound. 170
5.12 Event 1 of the McAdams oboe sound. 171
5.13 Event 2 of the McAdams oboe sound. 172
5.14 An excerpt (the first) from the Beethoven octet. 174
5.15 Events 1 and 2 from the Beethoven octet. 175
5.16 Events 3 and 4 from the Beethoven octet. 175
5.17 Excerpt from Beethoven's D-major violin concerto, with score. 177
5.18 Four events from the Beethoven violin concerto. 178
5.19 The remaining events from the Beethoven violin concerto. 179
5.20 A second excerpt from the Beethoven octet, with score. 181
5.21 Four events from the Beethoven octet, excerpt 2. 182
5.22 The remaining events from the Beethoven octet, excerpt 2. 183

C.1 A feature map for a segment of piano music. 204
C.2 One-dimensional filled-in map display. 206
C.3 One-dimensional line map display. 206
C.4 FM map with convolution kernel. 207
C.5 Mouse position information. 208

Chapter 1

Introduction

Accompanied by his band, Miles Davis plays a solo. Someone talks in an office full of clattering machines and ringing telephones. A car screeches to a halt in the street outside. In each of these situations, the auditory system is faced with the problem of separating several different sources of sound from the muddled signal that reaches the ears.

In hearing out separate instruments of that jazz band, one perceives the distinct notes of each instrument. This is done effortlessly, despite the fact that the different events that sum to produce the incoming sound signal may overlap in time, in frequency, or in other characteristics important to hearing. These confusing factors make it difficult not only to find distinct events in a sound signal, but also to associate these events with the sources that produced them. Experimental science has revealed some of the processes involved in the auditory system, but most of its complex workings remain to be modelled. The aim of this thesis is to describe and test mechanisms used to hear events in musical sound and to separate them when they occur concurrently.

1.1 Domains for Sound Separation

Sounds encountered by the auditory system may be broadly classified into music, speech, and environmental sounds. The auditory mechanism for source separation operates similarly on all of these types of sounds, though different separation cues may have more importance for some types than others. What follows is an overview of the characteristics of these sounds important for source separation and an introduction to some of the processes that operate in doing it.

Musical sounds usually have a pitch, which is both beneficial and detrimental to source separation processes: beneficial in that the fundamental frequency of a pitched note provides a basis for identifying which frequencies of the sound are associated with the note, and detrimental in that harmonically related notes — differing by, say, and octave, or a perfect fifth — have coinciding harmonics that can make the separation process much more difficult. Music usually has some sort of rhythmic structure, which is again both a boon and a bane. On one hand, rhythm provides a kind of order that can be useful in grouping notes into separate sources. On the other, it can lead to false associations, as when several instruments play notes at the same instant, inducing the auditory system to believe that they all come from the same source.

Source separation in speech research is usually driven by the need to separate out a particular voice from background noise [Cooke91, Weintraub85] or from other voices. The speech problem is characterized by the mixture of voiced, harmonic vowels and unvoiced, wide-band consonants, and also by the rapid rate of change of speech sounds [Borden84, pp. 176-196] [Moore89]. Again, each of these characteristics can aid or impede the progress of event separation. Harmonic (voiced) speech offers the same positive and negative factors to a scene analysis process as does music, with the added boon that the pitches of vowels from two simultaneous talkers are less likely to be harmonically related than those from from two simultaneous instruments. Consonantal sounds have sharp attacks and decays, providing strong cues for grouping mechanisms to operate on, but have little spectral information to help identify which of the sources present they may belong to. And the rapid rate of change gives a grouping mechanism a lot of material to work with, making the task easier, but implies that any cue present does not persist for long and thus has characteristics that are less easily detected.

All other sounds can be lumped together into the loose category of environmental sounds. Such sounds are an important part of everyday life, but have not been extensively studied, perhaps because of their variety. Though there have been efforts to distinguish some types of environmental sounds [Warren84], I know of no systematic attempts at categorization. Lacking useful characterization of such sound signals leaves the separation problem difficult indeed.

The focus in this thesis is on musical sounds. Part of the reason for this is that generating musical sounds, especially electronic ones, is easier than speech sounds, allowing for more precisely controlled experiments. Also, the "right answer" is easier to see for most musical sounds than for speech, in that the correct separation can be seen more clearly for mixtures of musical tones than for mixtures of simultaneous speakers. (This condition would not necessarily apply to a re-synthesis system.) The musical problem has been called source separation above; this also goes by the name of auditory scene analysis.

1.2 Definition: Auditory Scene Analysis

This section introduces a number of terms, culminating in a definition of the central problem, auditory scene analysis.

The terms feature, event, and source refer to different levels of organization of sounds in the auditory system. Since the terms overlap somewhat, definitions may help straighten out what general levels of organization they refer to.

A *feature* is a part of the sound signal occurring at a specific time and frequency. It includes such things as an onset of sound energy of a particular frequency at a particular time, or a change in frequency in the harmonic of a pitched sound. Feature filtering is strongly data-driven, meaning that the context of higher-level objects plays very little rôle in it.

While features are instantaneous, *events* extend over time and perhaps frequency. An *event* is an auditory phenomenon that exhibits constancy or at least continuity for its relatively short duration. It has an onset and an offset and represents the lowest time-extensive perceptual entity. In music, a single note is normally an event.

Features and events are auditory phenomena, points or regions in time/frequency space. In contrast, an auditory *source* (or stream) is a perceptual object, more permanent than an event, to which an explanation is attached. Bregman defines it as "our perceptual grouping of the parts of the [...] spectrogram that go together [acting] as a center for our description." [Bregman90, pp. 9-10]. There are two important points here. The first is that a source is a grouping of lower-level phenomena — all the tones from the violin part of a sonata, for example. The second is that it unifies our mental description of the auditory field, providing a perceptual handle for explanatory processes. We attach an origin, or source, to each sound we hear, using new sources for sounds that have not been assigned one yet. We usually have some explanation for the generation of the sound, be it vocalization, a vibrating string, frictional noise, or whatever. Electronically synthesized sounds do not necessarily fit any of these common physical models. We have many ways of explaining such sounds, sometimes by reference to other sounds whose physical models are known, sometimes simply by learning that sounds that are produced in certain contexts or that have certain characteristics are electronic.

Formation is the process of explaining or accounting for objects at one level by finding a higher-level object that corresponds to them. Thus event formation accounts for a number of lower-level features by grouping them into events, while source formation provides an explanation for events by assigning them to sources. *Separation*, closely associated with formation, is the distinguishing of objects at one level from each other, a process that necessarily occurs in tandem with formation. So *event formation and separation* is the process of identifying which features of the time/frequency image belong together, and grouping these features into events.

A few other terms bear defining. *Sequential* formation is the part of the integration process that associates entities over time. The complementary process, *simultaneous* formation, applies to entities that happen concurrently. These entities can be at any of the levels mentioned above, feature, event, or source.

Finally, *auditory scene analysis* refers to the entire process from the reception of a sound signal through source formation, including event formation along the way.

1.3 Why Study Auditory Scene Analysis?

Auditory scene analysis is a hard problem, so before embarking on its study the question of why to model it merits some attention.

One answer is that it is useful for a variety of applications. A scene analysis mechanism is a useful component of any device operating with real-world sounds. A speech recognizer, if it is to work outside a sound studio, must be able to cope with background noise and separate out the voice to be recognized. A car telephone or airplane cockpit radio would do well to separate out the desired voice from competing noise. An automated music transcription system must distinguish the separate instruments before it can assign pitches and rhythmic values to the sounds it receives. The techniques uncovered in sound scene analysis may also be useful in domains beyond sound, as for example when the source signal is a recording from a patient monitoring system and the signal to be separated from noise is the electrical activity of heart-muscle action.

Another reason for studying scene analysis is that it is an important part of audition. Any complete model of the human auditory system must give an account of how it does so well at picking out the attended-to source. Also, the processing models uncovered in scene analysis work may shed light on other auditory processes, leading to a deeper understanding of the entire system.

In addition, the perceptual approach used in the study of scene analysis may prove to be an insightful method for considering higher-level thinking processes. The framework of computation — the types of processes and representations — uncovered in the study of perception may extend to other parts of the brain and provide some clues for artificial intelligence researchers who are modelling higher kinds of thinking.

1.4 Perceptual Models

A perceptual model is a description of the process that transforms a physical stimulus, such as a sound pressure wave, into perceptual objects understood by higher-level brain processes, such as notes and melodies. A model provides answers to questions about how perceptual objects are formed, about the perceived relatedness of different types of phenomena, and about illusions. In addition, a *computational* perceptual model is testable because it includes, in addition to a theory of perceptual operation, a computational process — a simulation — to which physical stimuli may be applied to produce results that should match our perceptions.

1.4.1 Constructive Perception

An important principle to bear in mind is that perception is a constructive process, not a descriptive one. That is, the percepts received by higher levels of the perceptual pathways and the brain are constructed from the raw sound signal arriving at the ear. Kanizsa made a similar point for vision [Kanizsa79]. There is quite a difference between the sound wave carrying a short sine tone and the pattern of neural firings representing that tone at higher levels of the auditory system. It is the organization of the raw signal and deduction of its important structures that provides the percepts with which higher levels can work. As Witkin and Tenenbaum [Witkin83] put it, this constructive structuring is

> ...an area that lies at the very heart of human perception: the ability to impose organization on sensory data — to discover regularity, coherence, continuity, etc., on many levels.

The goal of the auditory modeller, then, becomes that of finding out what kinds of regularity and coherence are detected in the auditory system, how these regularities are computed, and what type of representation is constructed to convey them.

1.4.2 A Multi-Disciplinary Approach

The approach to auditory scene analysis used in this work involves the study and modelling of natural processes. This approach is founded on the realization that the human auditory system does a remarkably good job of separating sound sources, and that we can learn much from an understanding of its mechanisms. The approach must integrate components from several disciplines.

Study of the Natural System

This integrated approach starts with study of the human auditory system or of similar animal systems, drawing on introspective, psychoacoustic, and neurophysiological experiments in the process. Introspection is a useful way to determine what is and is not reasonable in a perceptual model, especially for the overall view of what is perceived. Psychoacoustic experiments provide more thorough information, including more accurate measurements of sensory characteristics, knowledge of effects that subjects may not even be consciously aware of, and information about individual variation. Neurophysiological studies can provide information about detailed processes and representations in the brain; as will be seen from the references, this has become possible mainly within the last twenty years, thanks to advances in the technology of neural recording and in understanding of the operation of neurons. (In fact, still newer technologies such as the non-invasive positron emission tomography [Fox86] and regional cerebral blood flow measurement [Ingvar89] may provide even better information about the processes of perception.)

Functional Modelling

After finding out as much as possible about how the auditory system works, the next step is to describe a physiologically compatible model. *Physiologically compatible* means that the model is faithful to known data and could potentially exist in physiology. Researchers have uncovered the most detail about the ear and cochlea, with generally less being known about neural processing in the auditory pathway progressively inward toward the brain. This makes the requirement of being physiologically compatible more constraining at lower levels (closer to the ear) than at higher ones. Correspondingly, it may be easier to achieve a working model at lower levels, since more of the mechanism has been laid bare.

Computer Implementation and Testing

After constructing a functional model, the next step is to implement it on a computer. Then we can test it to see how well it performs, tinker with it within the bounds of physiological reasonableness to see how it is affected, and compare the results with those of other models. As this process may shed some light on what does and does not work, it can potentially contribute to physiological work in suggesting what may be present and what things to look for.

Engineering

Hopefully, engineering practices will contribute to the advancement of perceptual models and vice versa. That is, engineering approaches to such application problems as speech recognition [Paul90, Lee90, Waibel89] may help perceptual modellers by finding techniques that work well. Though the aim of such systems is to be useful — and some have been fairly successful in this regard — rather than to provide insight about perceptual processes and representations, the techniques found may become part of modellers' realm of consideration when thinking about the auditory system.

Conversely, the knowledge gained from studying the functioning of the human auditory system may prove useful in building better application systems, perhaps by having these systems integrate parts of computational perceptual models. Advances in machine vision have long stemmed from a physiological approach, where researchers have been heavily influenced by Marr's computational theory [Marr82]. Perhaps the same transfer will begin to happen more in machine hearing.

1.4.3 Marr's Levels of Description

In describing models of the visual system, Marr [Marr82, pp. 24-27] presented three different levels of computation that a model may describe. The levels, which differ in the amount of detail and specificity supplied, apply equally well to descriptions of auditory models and may help orient a reader to the type of model offered.

The most general level of description is the *computational theory*, or *functional model*. A functional model describes what the various components of a system do without reference to how they work. Think of a computational theory as a block diagram. Next is the level of *algorithm and representation*, also called the *process* level, which describes algorithms and the representation of inputs and outputs of each functional unit. The most specific level, the *implementation* or *mechanism*, includes the specific software and hardware that performs a computation; examples are programs running on some computer, or a collection of neurons that are connected to compute a particular auditory data representation.

1.4.4 Overview of the Auditory Model

The model of early audition presented here focusses on auditory scene analysis. This overview is a description at Marr's computational theory level; it does not describe all of the functional units themselves or the representations used.

Fig. 1.1 shows the overall structure of the model. A more detailed description of the system at Marr's algorithm and implementation levels will be supplied in chapters 2 through 5. This model comprises a process modelling filtering and transduction in the ear and cochlea, a series of filters responsive to various features, a process for forming features into events, and a process for grouping events into sources.

Ear Filtering and Transduction

The first stage of any auditory model must be an ear model, that is, a functional unit that models the filtering done by the outer and middle ears and the transduction of sound to neural impulses in the cochlea. The ear model includes a process representing basilar membrane action, in which sound waves propagate from high frequencies to low, triggering responses from hair cells at specific frequencies. It also contains mechanisms for gain control and intensity encoding, as the 120 dB of dynamic range perceivable by the auditory system [Moore89, p. 47] must be reduced to the much smaller dynamic range of firing rates of neurons. The ear model also confronts the filtering tradeoff in which finer time resolution leads to coarser frequency resolution, and vice versa.

Feature Filtering

Next come a series of filtering processes, each of which extracts or filters a different type of feature that may be present in the data. These filters output a number of

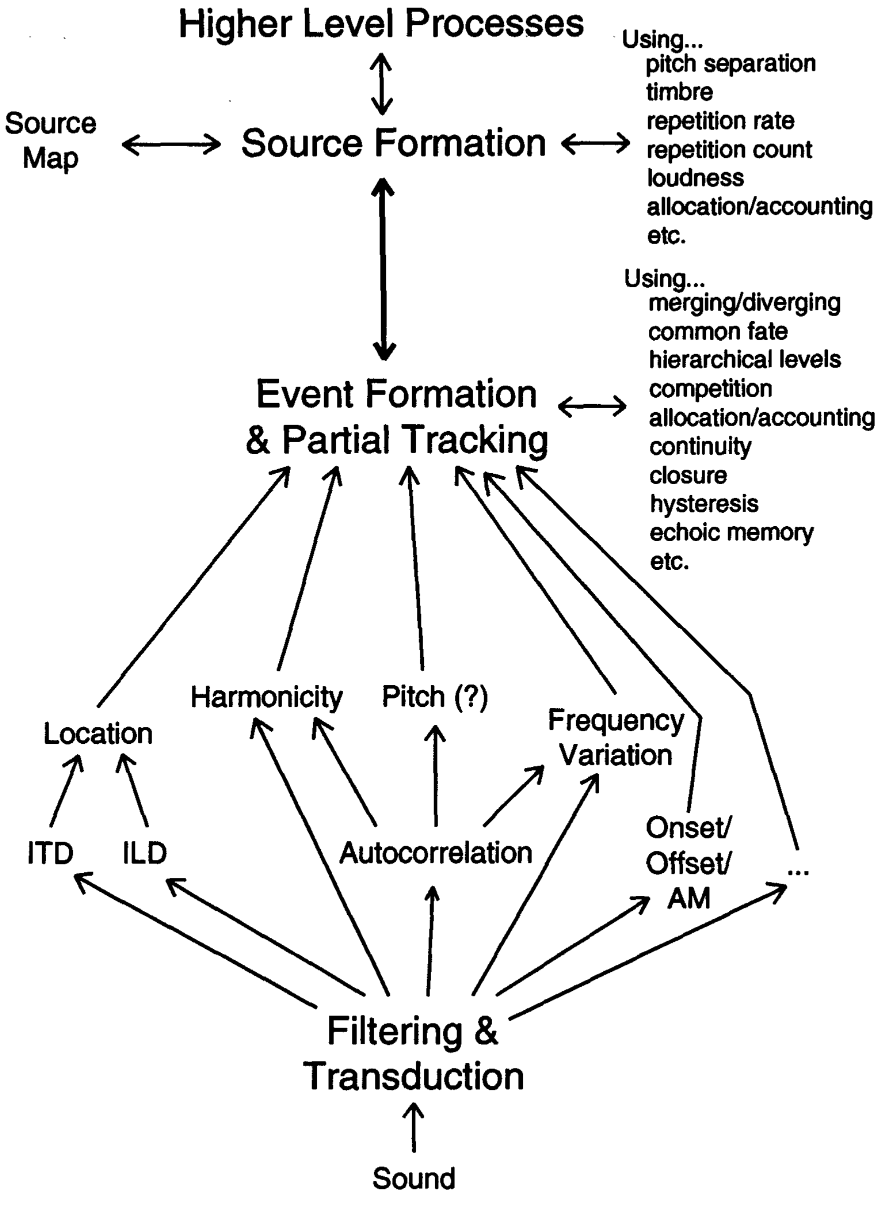

Higher Level Processes
Using...
pitch separation
timbre
repetition rate
repetition count
loudness
allocation/accounting
etc.
Source Map
Source Formation
Using...
merging/diverging
common fate
hierarchical levels
competition
allocation/accounting
continuity
closure
hysteresis
echoic memory
etc.
Event Formation & Partial Tracking
Harmonicity
Pitch (?)
Location
Frequency Variation
ITD
ILD
Autocorrelation
Onset/ Offset/ AM
...
Filtering & Transduction
Sound

Figure 1.1: Auditory model block diagram.

feature maps, representations in time and frequency and perhaps other dimensions of these features. Some of the features that may be found at this level include

- amplitude onsets and offsets, changes in the amplitude of a frequency channel over some short period of time;
- frequency variation, changes at some rate in the frequency of a sound component (partial);
- harmonicity, how well the sound energy at a particular frequency fits into a harmonic series with other energetic frequencies;
- amplitude variation, for changes in sound energy or neural firing intensity over relatively long periods of time (these are distinguished from the short-time amplitude onset because of the prevalence of the latter in neurons of the auditory system [Rhode86] [Pickles88, p. 170]); and
- interaural time and level differences, for sound localization.

Event Formation

After these features have been filtered — some directly from cochlear firings and some from previous feature maps — the features are analyzed to parcel out the energy represented by neural firings into various sound events. This event formation process uses different features in different ways and must obey a number of constraints. For example, one constraint is that if there is a sound from a particular event at one frequency at one time, then sound at the same or a nearby frequency a few milliseconds later is likely to belong to the same event. Another constraint is that onsets at different frequencies at or near the same time probably belong to the same event. Information about event formation need not be completely data-driven, that is, need not come entirely from the low-level feature filters described above: it can also include knowledge about patterns of learned or recently-heard sounds, including timbre, short-time speech patterns, and other time-frequency schemas.

Source Formation

Finally, the source formation process assigns events to separate sources. Up to this point, all of the processing has occurred on fairly short time scales. Features are represented at points in time-frequency space, and events typically range from a few tens of milliseconds to a few seconds. Sources last arbitrarily long — the fan I hear now, for example, has been running all day — and act as grouping points for all of the events that compose them. The principles obeyed by the source formation process are fairly complex, bringing in knowledge of event behavior as well as high-level knowledge about pitch, rhythm, speech, behavior of noisy objects in the environment, grammar, and even culture.

The final result of this process is a separation of the incoming sound signal into its constituent sources.

1.5 Previous Work

Previous efforts at scene analysis, or source separation, have focussed either on musical examples or on specific speech situations such as co-channel speech, that is, two simultaneous speakers in one sound signal.

Moorer [Moorer75, Moorer74] made an early effort to separate pitched sounds by a variety of filtering techniques. His effort was notable, among other reasons, for handling more than two simultaneous pitched sounds. His techniques included a waveform autocorrelator for finding periodicities, a comb filter for identifying several notes in a chord, and a bandpass filter with pitch detector. His pioneering work, aimed at automating polyphonic music transcription, did not try to model the auditory system.

Parsons [Parsons76] built a system for speech separation based on the harmonicity of vowels. His system tabulated peaks in the spectrum that were assumed to be harmonics of a vowel fundamental, dealing carefully with overlapping peaks; picked two pitches that best accounted for the peaks, using Schroeder's pitch mechanism [Schroeder68]; assigned each peak to one pitch or the other, interpolating missing harmonics; and resynthesized the separate voices from the two sets of spectral peaks.

Though consonants were ignored in his method, some of their characteristics survived in consonant-to-vowel transitions, and the output of his system was reasonably intelligible speech.

A project at Stanford's Center for Computer Research in Music and Acoustics begun in the early 1980's began a movement toward the study of human audition to improve computerized music analysis. It identified some techniques that could be used for intelligent music analysis, including scene analysis [Chowning84]. One technique [Mont-Reynaud85] aimed at identifying patterns in musical rhythms; though not polyphonic, it could potentially be used as a source-separation technique. Another [Chafe85] used a log-frequency transform [Kashima85] as a front end for an algorithm that finds the abrupt changes in amplitude envelope characteristic of percussive instruments. Further work focussed on source separation, including event detection and segmentation, periodicity estimation, and the use of metrical context for note identification [Chafe86].

As part of this project, Schloss [Schloss85] developed a system for identifying musical events based on the amplitude envelope. His method was to segment the sound in time according to the slope of the amplitude envelope, to pick out notes based on their attack and sustain periods, and to identify different types of notes by their attack characteristics. His method was used to separate drum beats in time and classify them according to type of drum strike. While his aim was not the separation of simultaneous sounds, some of his ideas of onset discrimination could be useful in the polyphonic case.

Throughout this period, development continued of improved ear and cochlear models, which simulated low-level auditory processes more exactly. They will be reviewed in more detail in section 3.1.

Weintraub [Weintraub85] also advanced the art of computational source separation. He identified many of the regularities present in a sound signal that enable the auditory system to separate sources, and suggested an auditory model as a prerequisite to further work in sound separation. He proposed that a per-frequency-channel autocorrelation function of the type suggested by Licklider [Licklider51], implemented as a neural delay line, could be the basis for a perceptual pitch segregator. His system was focussed on the separation of two simultaneous vowels. It used pitch and onset as cues for source separation and sought to identify periodicities that would aid in separation. He identified the need to track harmonics over time, employing a simple finite state machine to describe the action of each harmonic.

Vercoe [Vercoe88] presented a neurally-inspired technique for source separation. His method, which aimed to be physiologically compatible, used a network of connected simple computational units to filter frequency variation in sound. The goal was to use this frequency variation information as a source of event or source separation.

More currently, Assman and Summerfield [Assman89a, Assman89b] identified two concurrent vowels by summing the per-channel autocorrelation measure across channels. This "compound autocorrelogram," which highlights common periodicities, is then searched for the two highest distinct peaks and looked at for for how well the different frequency channels in the (plain) autocorrelogram matched these peaks. Their model matched the behavior of human subjects fairly well, but was again limited to segregating simultaneous vowels.

Serra developed a method for tracking harmonics and non-harmonic partials [Serra88]. His system, based on finding the sinusoidal components of a sound [McAulay86, Smith87], tracks partials over time and models the non-sinusoidal component as noise filtered by low-order filters, permitting a number of musically interesting transformations before re-synthesis. The system did not originally do source separation, though Schottstaedt [Schottstaedt91] is currently adapting it to do so.

Recently, Cooke [Cooke91] presented a model of the auditory system oriented toward source separation. Like the present study, Cooke's work proposes a layered auditory processing model, with a cochlear layer computing a place-based representation, filtering elements that respond to various kinds of features in the sound signal, and a grouping mechanism that tracks harmonics over time. He also identifies many of the types of features useful for source separation, and suggests a grouping mechanism that brings all of the information together to make decisions about what sources are present. His model is fairly similar to the one presented here; the differences are delineated in chapter 6.

1.6 Scope of this Work

1.6.1 What is Here

This thesis studies the auditory model overviewed in the preceding sections. It includes

- a review of the introspective, psychoacoustic, and neurophysiological evidence for the response in the auditory system to various features useful for scene analysis;
- an implementation of filters for some of these features;
- a description of the operating principles and constraints of the event formation mechanism;
- an implementation of enough of these principles to solve some source-separation problems in music; and
- a description of some of the characteristics that the source formation process must have.

With this model, I intend to answer the following questions:

- How might the auditory system separate sources?
- What is a computational theory suitable for auditory scene analysis?
- What are physiologically compatible processes modelling the parts of this theory that compute low-level features for frequency variation and for amplitude onset?
- What representations for various kinds of auditory features are both physiologically compatible and useful?
- What principles influence the process that integrates information from the feature extractors to make decisions about sound events?
- How can such principles be incorporated into a computer implementation?

- How well does such an implementation work?
- What are some of the principles that influence the grouping of events into sources?

1.6.2 What is Not Here

All of Audition

This is not a theory of all of audition, or even all of early audition, though the framework of computation from feature map to feature map could perhaps be extended to a wider range. The part covered here is focussed on source separation. Also, the event formation mechanism implemented here is incomplete, omitting some of the grouping principles discussed here and no doubt other as-yet-undiscovered ones. And no attempt is made to take into account the many subtle factors that affect source formation.

All of Auditory Scene Analysis

This work does not cover in depth the factors that affect the scene analysis process; this was done ably by Bregman [Bregman90]. All that is here is an in-depth review of the types of features used for event formation, an overview of the principles that apply to grouping these features into events, and a shallower overview of some of the principles that apply to source formation. The parts of the model that are implemented in software are even fewer, comprising filters for some of the feature filters and an event-formation algorithm.

Learning

Nothing in the system described here makes any modifications in its behavior on the basis of its input. Learning is undoubtedly a part of the auditory system at many levels, but none of it is modelled here.

Auditory Pattern Recognition

In this thesis, there are no auditory pattern recognizers or classifiers beyond the simple feature filters of chapters 2 and 3; nor is there an attempt to survey such methods. Such components are undoubtedly a part of the auditory system at some level, just as object recognition is a part of vision, but because we have little or no evidence about the workings of such a process in the auditory system, there is no attempt to model one here.

Re-synthesis

Though re-synthesis would be a useful way to present the output of a source separation mechanism, it is not developed in this work. The emphasis is on an auditory model, whose output is a pattern of neural firing sent to higher levels of the auditory system. An alternative approach might make a real-time signal processing system that outputs two or more sound signals from a single input signal; this is not done here.

1.7 Structure of this Document

This thesis is organized from the ear inward to the brain, or from signals to sensations to percepts, or from low level information to high if you will. Chapters 2 and 3 cover early audition, up to the level of low-level features: Chapter 2 covers the cochlear model used, feature-map data representation, and the psychoacoustic and neurophysiological evidence for various types of features. Chapter 3 presents a software implementation of some of these feature filters and their output maps. Chapters 4 and 5 model the higher perceptual processes: Chapter 4 covers grouping principles for event and source formation. Chapter 5 implements some of the event-formation mechanisms and principles discussed in chapter 4 and presents some examples of sounds run through the computational model. Chapter 6 summarizes the previous chapters and suggests future work.

Chapter 2

Architecture of the Early Auditory Model

This chapter begins the development of the model of early audition for scene analysis. It covers the features that may be useful for data-driven simultaneous separation.

2.1 The Ear

2.1.1 Outer and Middle Ear

Sound processing by the human auditory system begins with the filtering done by the outer and middle ear as the sound travels to the eardrum. The outer ear, or pinna, reflects sound waves in various frequency- and direction-dependent ways that aid sound localization at higher levels of the auditory system. The middle ear matches the impedance of the air column in the ear canal to that of the fluid in a cochlear duct [Pickles88, pp. 15-22]. The middle ear also performs some gain control, effective mainly below 1-2 kHz, by muscular contraction, stiffening the chain of bones in the presence of vocalization, loud sounds, or movement [*ibid.*, p. 23].

2.1.2 Cochlea

Next the sound waves travel along the cochlea, causing the basilar membrane within it to vibrate as the waves move down it. This causes the inner hair cells, which are attached to the basilar membrane, to vibrate in turn, allowing charged ions to flow into the cell bodies. Also on the basilar membrane are the outer hair cells, which probably perform active gain control [Pickles88]. This charge flow, if sufficiently strong, triggers neural firings, completing the transduction of information from sound pressure waves to a representation compatible with higher levels of the nervous system.

Phase Locking

Several important features of the pattern of neural firing on the basilar membrane should be mentioned. Up to a maximum of about 5 kHz, neurons are synchronized to the waveform, firing only when the sound wave is in the rarefaction portion of its cycle and not in the compression part. Thus the firings maintain a constant phase relationship to the wave, a feature which will be important later. Above a few hundred cycles per second, neurons do not fire on every peak of the waveform; it is just that when they *do* fire, they are restricted to firing during one half of the wave. Since a large population of neurons responds to each frequency present in a sound, some unit of this population responds in each period of the wave. Thus the firing of the population as a whole encodes individual peaks of the wave.

Frequency Ordering and Height

Also, frequencies on the basilar membrane are ordered from high frequency at the basal (input) end to low frequency at the apical end. Such an arrangement is called tonotopic. In fact, frequencies are spaced approximately logarithmically, so that as great an extent of membrane encodes the 1-2 kHz band as encodes the 2-4 kHz one. This mapping is not precisely logarithmic, as the encoding becomes progressively more linear at frequencies from 1 kHz on down. The mel scale describes psychophysical pitch distances which may correspond to basilar membrane frequency spacing [Stevens37, Stevens40].

This quasi-log-frequency dimension is ubiquitous in the auditory system, common enough that it deserves a name: *height.* Height h is defined in terms of frequency f by $h = \log f$.

2.2 Feature Maps

Next in this auditory model lie a number of separate feature-extraction processes. Before describing them, I must first introduce a data representation capable of holding these features.

A *feature map* is an n-dimensional array of real-valued numbers, where each number, usually positive, represents the intensity of some feature present in a signal. The number of dimensions n is sometimes two, sometimes three, and sometimes two and a half, meaning a small number of two-dimensional arrays.

2.2.1 Analog, Continuous Maps

Feature maps are analog and continuous. Analog in this sense means just that the numbers in a map are real values, rather than, say, boolean truth values. Continuous means that the axes of a feature map are smoothly varying domains, such that a small change in position in this domain represents only a small change in the underlying properties. Also, the range is continuous, in that a small change in the intensity value at a particular position in a feature map corresponds to a small change in the represented property. A typical two-dimensional feature map has time (measured relative to some designated starting time) as one axis and height as the other, with a value at a particular position representing the intensity of neural firing on the basilar membrane for that time and height.

Such an analog, continuous map may be contrasted with the symbolic representations that have been dominant in artificial intelligence. Symbolic values are neither analog nor continuous: a small change in value can change their meaning arbitrarily. Words are a typical symbolic representation; consider the difference between the meaning of *hem* and *hen.* Though the change in representation is small either alphabetically or phonetically, the change in meaning is not.

There is an ongoing debate about whether signals, including multidimensional ones like feature maps, or symbols are the best way to represent objects of thought. Symbols clearly have a place in cognition and reasoning, but signals are required for the lower-level constructs found in early perception. It could be said that *the* task of perception is to establish and maintain a meaningful correspondence between analog reality (signals) and the more symbolic representations at higher levels [Mont-Reynaud91]. Feature maps begin to bridge the gap; they are a step away from signals toward symbols, in that they represent the intensity of some symbolically chunked percept such as an onset that may be represented at higher, symbolic levels.

2.2.2 Sampling

The implementation of feature maps, using arrays, implicitly includes the idea of sampling. The auditory sysmtem may or may not contain such sampling, so a possible breakdown of physiological compatibility occurs here. For instance, the rôle or presence of time sampling in various auditory processes in unclear. However, since any signal of finite bandwidth can be accurately represented by a sampled signal, information is not necessarily lost. The sampling interval in a computational model must reflect the auditory system's sensitivity along each dimension of variation. Topographic maps in the auditory system are common, and each position in such a map represents a sample of the parameters that vary along the dimensions of the map.

The aim, in the implementational model in chapter 3, is to sample the dimensions of each feature map finely enough to capture all of the important information and represent all of the important processes. In thinking about feature maps, it is reasonably safe, because of their analog, continuous nature, to think of them as continuous representations.

2.2.3 Locality of Processing

Another property of feature maps is that the computations that produce them are primarily *local.* A local computation from one feature map to another is one that uses only points in a small neighborhood of the location to be computed. So to compute a value at a particular position in a map representing some feature, a local process looks only at values near that position in the input map(s). Local computations between feature maps have come to dominate the field of machine vision (see, for example, [Julesz81, Watson85, Adelson86, Heeger88, Malik89]). Local computations, as well as feature maps representing their results, are often called *topographic.*

Some hearing processes are apparently non-local. For instance, the part of pitch perception that is place-based must integrate information from across the spectrum, since harmonics from many widely-spaced frequencies can contribute to pitch perception. It may be that there is some neural mapping of the tonotopic order of auditory nerve fibers that brings different frequencies together in such a way that local computations can perform these non-local computations.

The axes of a feature map other than time and height (log frequency) are typically the variation of some parameter of the filter that computes the map. In a map that represents frequency variation, for example, the parameter might be the rate of variation — positive for upward frequency variation, negative for downward. In a map for amplitude onsets, the dimension of variation might be the characteristic duration of onset.

2.2.4 Spatialization of Time

The computer implementations of the model presented in later chapters represent time as an explicit dimension of feature maps. In the brain, of course, time is implicit — things happen over the course of time, so that "time is its own representation." There is no known spatialization of long periods of time (> 0.5 s) as in the feature-map arrays of this model. A reasonable point, therefore, is that the implementation's feature maps are not physiologically compatible.

The answer to this objection is that the spatialization of time in the implementation is merely a computational convenience.

The processes of this model implemented as in chapters 3 and 5 do not make use of this spatialization to use information from widely-separated times; at each step, they use information from only a small extent in time. The spatialization provides only an aid in viewing the data, not a computational advantage. Given a sufficiently fast and/or parallel processor, they could run in real time, using information from only the last few tens of milliseconds. Such delays are physiologically compatible with the speed of neural processing; indeed there is evidence for neural response to features of such sizes [Schreiner86, Schreiner88b].

2.2.5 Feature Maps in the Brain

Part of the justification for using feature maps, aside from the reasons mentioned in the signal-*vs.*-symbol debate above, is that maps similar to feature maps are found in the brain. Many different types of maps have been seen in the auditory system to date. Since the tonotopic arrangement of frequencies on the basilar membrane is preserved in the auditory system, nearly all cortical maps have approximate frequency (or height) as one axis: "The neurones are tonotopically organized in all specific central auditory nuclei." [Pickles88, p. 165]

Suga [Suga90] has found several such maps in the auditory cortex of the bat, many specialized for echo-location. "FM-FM" maps compare pulsed pitches emitted by the bat with returning echoes, and have axes of echo delay, for target range, and amplitude, for target size. Another map has axes of frequency (tonotopy) and frequency-change rate, for Doppler-shifted target velocity measurement. Schreiner and Mendelson [Schreiner90] found a map in the cat encoding "spectral tilt," where an amplitude change from one frequency to another induced firing, and Shamma found strong evidence for a similar map in the ferret [Shamma91]. Amplitude modulation maps in the cat [Schreiner88a] encode short-time amplitude fluctuations, perhaps as a first step toward sound localization. Maps encoding the spatial location of a sound have been found in several animals; Yin and Chan [Yin88] have a good survey. The diagram in fig. 2.1, taken from Knudsen [Knudsen81], shows graphically how the directions in space around the head, azimuth and elevation, are smoothly mapped onto two spatial dimensions of a structure in the owl brain, the magnocellularis lateralis pars dorsalis.

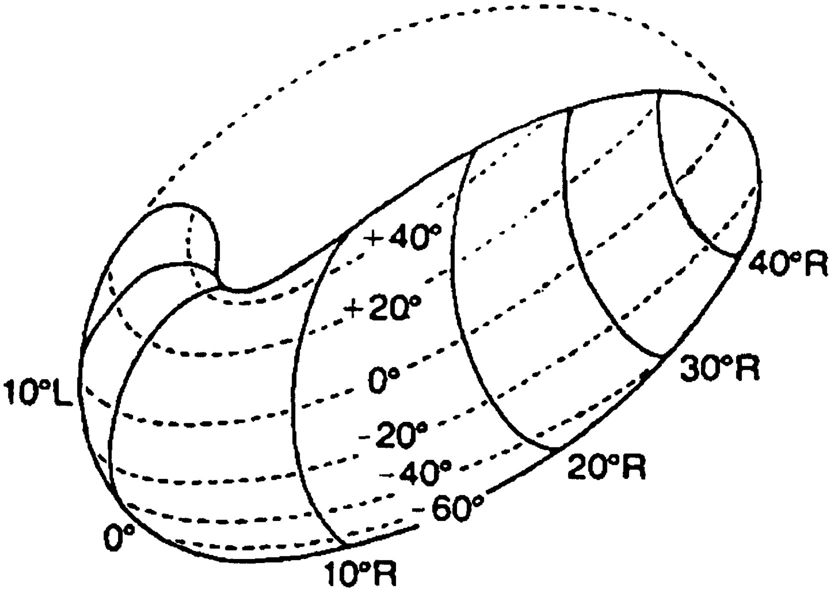

+40°
+20°
0°
-20°
-40°
-60°
10°L
0°
10°R
20°R
30°R
40°R

Figure 2.1: Two-dimensional space map in the MLD of the owl. From "The hearing of the barn owl," by E. I. Knudsen. Copyright © 1981 by Scientific American, Inc. All rights reserved.

2.2.6 Map Computation

The above maps are offered as evidence that topographically organized maps are ubiquitous in the auditory system, and are a productive way to compute and represent auditory features. Computation from one map to another is the basic structure of this auditory model. (Topographic maps may not represent every feature in the auditory system [Shepard89], but they are the only representation used here.)

The primary map from which all others are computed is the time × height pattern of firing of cochlear nerve cells, with time measured relative to some reference time t_0. All other maps are computed from this one, either directly or indirectly via intermediate maps, generally using local computational operators. These computed maps extract features of various types for use by the grouping process to be described later.

2.2.7 Organizing Maps in Space

One question that arises about the auditory system is how maps representing different features are arranged. If each map is kept in a different location in the auditory pathway, then there arises the problem of bringing together the data that belongs together. The data for a specific frequency, say, would be scattered over several areas if several maps have frequency as a dimension, and such information must be brought together for many types of computation. This recombination problem is not easily solved. If, on the other hand, maps are overlapping in space, then the different cues could be computed next to each other. The only problem with this plan is that there is simply not enough space in the three dimensions of the brain to represent all of the various maps. Pickles comments,

> "In a common analogy, the first case might be compared to a photograph, in which each point in the photograph represents one point in space, whereas the opposite end of the spectrum might be compared to a hologram, in which each point on the hologram represents many points in space, and in which individual points in space can be reconstructed only

> by the integration of information from many points on the hologram." [Pickles88, p. 164]

Fortunately, a computer implementation of the model need not suffer from this dilemma, thanks to the random-access nature of computer memories. We can simply have maps in separate locations in memory and associate values at, say, a common frequency by just accessing the maps with the right index for that frequency in each map. To what degree this is a violation of physiological compatibility is a question that may be answered as more is learned about the auditory system.

2.3 Autocorrelation

One of the most useful types of feature maps is the autocorrelation function of each channel of cochlear output. The autocorrelation of a function $f(t)$ is defined by

$$A(l) = \int_{-\infty}^{\infty} f(t)\ f(t+l)\ dt,$$

where l is the lag, or delay, in correlation. The autocorrelation is usually windowed with some window W to localize it in time:

$$A(l,t) = \int_{-\infty}^{\infty} W(t_1) f(t+t_1)\ f(t+t_1+l)\ dt_1.$$

W is often a rectangular window. If so, we may re-write this more clearly as

$$A(l,t) = \int_{W} f(t+t_1)\ f(t+t_1+l)\ dt_1.$$

The autocorrelation of a signal reveals periodicities in the signal. A nerve cell on the basilar membrane which responds to frequencies below 4-5 kHz fires in synchrony with the wave — that is, at a roughly constant phase with respect to the periodicity in the signal at the cell's center frequency. An autocorrelation applied to the output of such a neuron, or to the sum of firings of a population of neurons, will reveal the periodicity in the signal.

Applying an autocorrelation to each cochlear nerve channel independently reveals the periodicity at each frequency, if any is present. Fig. 2.2 shows such a result, called

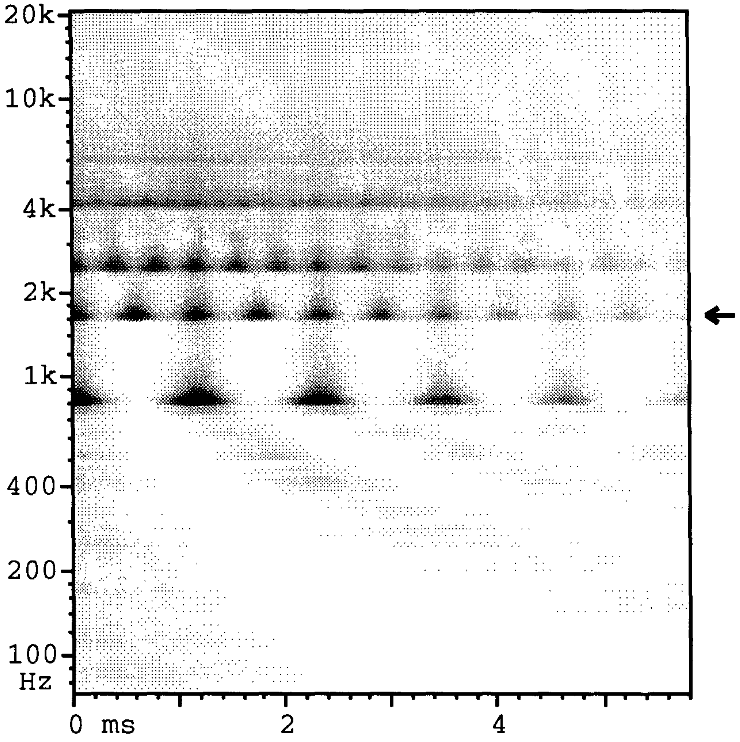


Figure 2.2: Typical correlogram for a pitched sound.

a correlogram. Height is encoded on the vertical axis, and lag on the horizontal. Each frequency that has a noticeable amount of energy, such as the one shown with an arrow, has a series of dark spots across the image in a horizontal line. Each spot going from left to right corresponds to one period of the waveform. Spots tend to have a smear extending upward due to the pattern of excitation on the basilar membrane: hair cells with center frequencies slightly higher than a given frequency f in a signal tend to fire in constant phase with those at f because of the way a sound wave travels down the basilar membrane [Shamma85, Pickles88, pp. 37-47].

2.4 Partials

In an image of a spectrogram or neural spectrogram of a pitched sound, certain features immediately stand out. These are the dark horizontal lines, usually lined up in parallel groups in the image. Fig. 2.3 shows a typical sound with a number of such horizontal lines. These lines correspond to *partials*; when their frequencies are in a harmonic series, or an approximate harmonic series, they are known as *harmonics*. An important task of any event formation and detection system is to identify partials and group together the ones that belong to the same event. The algorithm for event formation in chapter 5 presents a representation for partials and uses them as a primary means to track changes in the spectrum of a sound over time.

2.5 Features for Event Formation

The remainder of this chapter comprises a review of various low-level features that can be used for sound event formation and separation, including psychoacoustic and neurophysiological evidence for the presence in the auditory system of mechanisms responsive to the features. The features to be described at this level are what Bregman calls *cues for simultaneous integration*, because they are short-time cues that can be thought of as operating at slices in the same short interval [Bregman90, p. 213]. They are amplitude onset, frequency variation, harmonicity, amplitude variation, and spatial location.

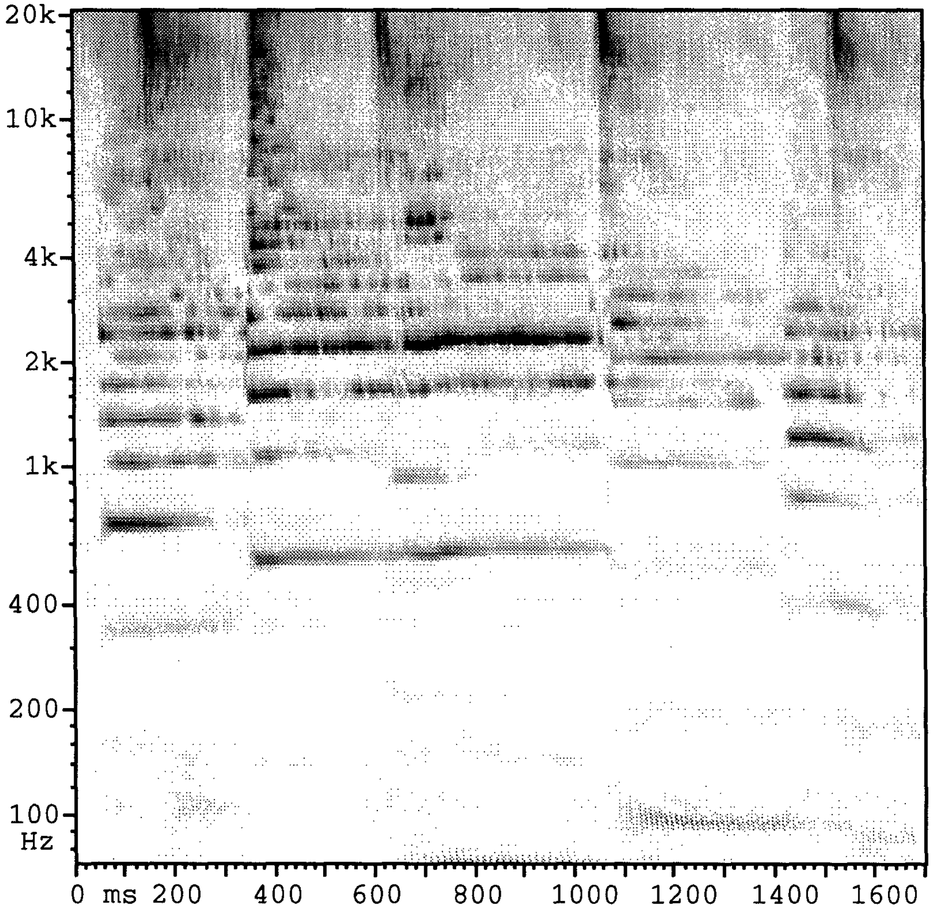


Figure 2.3: Cochleagram display of saxophone and drum tones, showing partials.

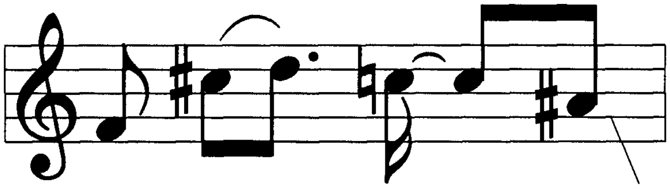


Figure 2.4: Score for the previous cochleagram display (without drum notes).

The features to be covered in this section generally do not include commonalities such as common amplitude onset, common frequency variation, etc. All that is of interest here is the raw feature such as amplitude onset, frequency variation, or location. Most of these features are the basis of event formation when several partials in the spectrum have them in common. For instance, common amplitude onset of several partials can be seen at about 340 ms in fig. 2.3, where a number of partials appear at different frequencies. Common frequency variation can be seen around 1500 ms, where several partials have a parallel decline in frequency. Grouping of partials by common characteristics of features is left until later, in the event formation section.

This chapter is a review of evidence for several of the features used in event formation and, later, source formation. This evidence is mainly of two forms. Psychoacoustic evidence demonstrates that the feature in question is an important part of the scene analysis process. Neurophysiological evidence shows that there exist neurons in the auditory system that respond to a certain feature. Such evidence suggests that these features are filtered fairly early in the auditory pathway, enabling them to be used as information for an event and/or source formation process. In this review, I also mention a few introspective demonstrations that are widely known or are especially telling, even if they have not been put through the rigors of a psychoacoustic test.

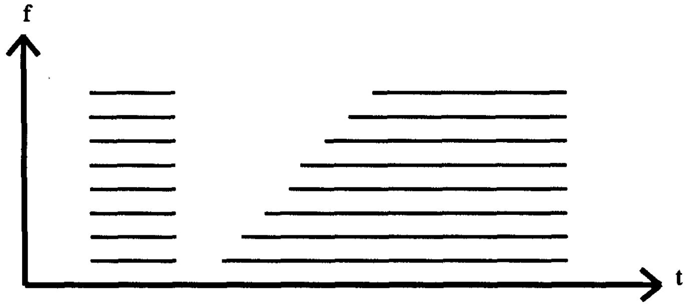


Figure 2.5: Tone pattern in Pierce's demonstration.

2.6 Onset

One of the features acting most strongly for grouping of related partials is common amplitude onset, or simply common onset. Hartmann notes this:

> It is appropriate that onset time be give first place in this list, because it is probably the primary cue by which sounds from different sources are segregated in everyday listening to speech and music. [Hartmann88, p. 625]

In nearly all musical instrument tones, the start of a note is marked by the rapid rise of all strong partials within a period of about 40 ms [Grey75]. This common onset gives the auditory system a strong cue that the partials belong to a single source. In many types of ensemble playing, performers take great pains to have a common onset. This does not greatly inhibit the ear's ability to hear the separate instruments, but if one instrument leads or lags it stands out noticeably.

Pierce [Pierce83, p. 226] presents an illustration of the power of onset asynchrony to pull apart a single source to make several. He plays a short tone made up of a number of harmonics of the fundamental frequency at 220 Hz, shown schematically as the note at the left in in fig. 2.5. He then plays the same set of harmonics, but introduces each of them one second after the previous one. As each harmonic enters, it stands out briefly as a separate tone before merging with the existing mass of partials. Since each new partial is part of a harmonic series with the existing ones, the only evidence present for the grouping process in the auditory system is that the partial has a different onset from the others. This makes it stand out initially as a separate event. Why it later merges with the other partials will be discussed in chapter 4.

Psychoacoustic Evidence

Gordon [Gordon84] studied the differences between various different sound characteristics that fall under the general name of "onset." The ***physical onset time*** of a note is the instant sound is physically present. With sufficient intensity and a suitable amplitude envelope, *perceptual onset*, the instant when a note is first audible, occurs at the same time. *Perceptual attack time* (PAT) is a tone's moment of attack relative to the physical onset; the attack time can be later than the onset in, for example, reed and bowed string tones. Because of these differences in perceived attack of different instruments, it is important in measuring perceptual onset-time differences to define a "standard instrument" for comparison.

Gordon measured PAT by placing a given tone A and a standard instrument tone B from an E-flat clarinet in a repeating A-B-A-B... pattern. Subjects could adjust the onset of the A tones relative to the B tones until the two were perceived to be evenly spaced in time — so the A-B-A-B rhythm was as uniform as possible. Gordon also had subject adjust tones that were meant to have simultaneous attacks. He found that in the former case, subjects could be consistent in their judgment to 1-2% of the beat interval, and in the latter case, to about 2 ms provided perceptual fusion of the two instruments did not occur. Gordon tested a variety of physical measures to try to match the perceptual data. Several of these worked well. The best one marked the instant when the amplitude envelope's slope reached a certain threshold value; then measured the amount of time that the slope stayed above threshold (the rise time); and finally added a fraction of this latter time to the marked instant.

An important characteristic of onset synchrony is that it forms an important part of the timbre of a musical instrument. Grey's multidimensional scaling study of instrument timbres [Grey75] identified three dimensions of important timbral variation,

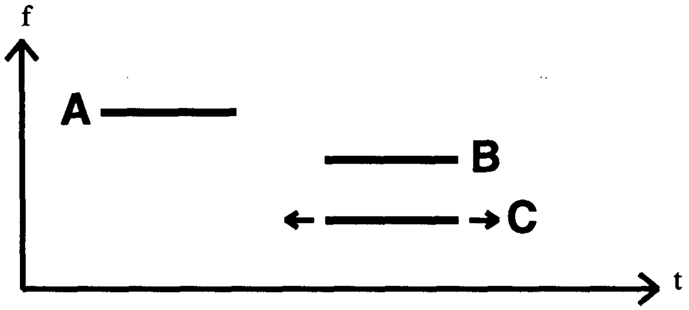


Figure 2.6: Tones for the onset part of Bregman/Pinker experiment.

reporting of one of them that

> [it] would appear to relate to the form of the onset-offset patterns of tones, especially with respect to the synchronicity in the collective attacks and decays of higher harmonics. [*ibid.*, p. 61]

In other words, the auditory system is quite capable of detecting small differences in the onset (and offset) times of the different partials.

Bregman performed a number of experiments that confirm the effect of onset synchrony in grouping partials together. A classic experiment, done with Pinker [Bregman78b], involves varying the characteristics of three sine tones to change the way they are grouped into streams by the auditory system. Fig. 2.6 illustrates the portion of the experiment which relates to common onset. Tones A, B, and C are sine tones presented repeatedly to the listener for about 150 ms each. Bregman and Pinker found that the degree of onset synchrony between partials B and C strongly affected the perception of whether partial B stood by itself or combined with C to make a richer timbral tone complex BC. A 30 ms difference in onset time of the two partials caused a significant reduction in the degree to which they were heard as a single tone. The experimenters conclude, "We did demonstrate ... a clear overall effect of B's asynchrony upon the stream segregation judgements."

Rasch [Rasch79] measured onset asynchrony among different instruments in an ensemble and found that the asynchrony (RMS) varied from 27 to 49 ms. This large a difference, in light of Bregman and Pinker's 30 ms difference, suggests that the auditory system is able to hear out separate instruments in ensemble music partly by virtue of onset asynchrony. This conclusion fits with Gordon's work with attack transients, which showed that perceptual attack times of a single instrument may themselves be as long at 30 ms. [Gordon84]

Interestingly, Rasch also found in another experiment that listeners were able to detect large perceptual differences in two test sounds that differed by 30 ms in the onset synchrony of their partials, even though they could not otherwise say what the difference was.

Darwin [Darwin84] did a number of experiments with vowel perception that show the influence of onset asynchrony. He added extra energy near the first formant of a vowel sound, an action that changes the vowel quality. However, he found that the extra energy has no effect if it has a sufficiently different onset time — as small a difference as 32 ms from that of the vowel. "A tone that starts or stops at a different time from a vowel is less likely to be heard as part of that vowel than if it is simultaneous with it." [Darwin84]

Neurophysiological Evidence

Evidence for neurons that fire upon onset of a particular frequency is fairly extensive. Pickles describes a set of cells in the cochlear nucleus:

> Cells showing onset response produce a sharp peak in [firing] at the beginning of a tone burst, and then either no activity, or a low level of sustained activity. Such cells are found throughout the cochlear nucleus.... We might suppose that there is an excitatory input, and then a delayed inhibitory input." [Pickles88, pp. 170-172].

This last sentence gives a clue to the implementation which will be used in the algorithmic model in chapter 3.

Young *et al.* [Young88] point out that precisely timed onset responses are necessary for sound lateralization, and suggest that cochlear nucleus cells may be specialized for this purpose. This could well be the case, though it does not preclude projections of cochlear nerve cells to higher areas which could produce the perception of onsets found in the psychoacoustic experiments.

Schreiner and Urbas [Schreiner86] report neurons in the auditory cortex — well past the central nucleus of the inferior colliculus, at which lateralization is believed to take place — that respond best to signals that rise or fall with a 10 ms characteristic time. There is no proof that such units are encoding onsets as we experience them perceptually, but they certainly are encoding onsets for some reason.

2.7 Frequency Variation

Frequency variation (FV) refers here to the change in frequency of a partial. It is reflected in a cochleagram by a partial with a non-zero slope. Fig. 2.7 shows a tone steadily rising in frequency.

Filtering of frequency variation may be based on the output of some pitch extraction process. *Pitch* is a subjective measure (or combination of several subjective measures) of the height of a tone, as on a musical scale. It is a perceptual phenomenon, and is related to the periodicity or fundamental frequency of a harmonic or quasi-harmonic sound. It may be contrasted with frequency, which is a physical measure of the repetition rate of a pure tone.

As will be explained page 51, many pitch detectors have been proposed, some of which are physiologically compatible. These put together information from across the spectrum, finding the fundamental frequency or frequencies that are present in the sound. As Bregman points out, however, pitch perception is a complex phenomenon that can be influenced by the source-segregation process [Bregman90, p. 246]. For this reason, the interest here is in mechanisms that act independently of pitch perception — that could be part of the input, directly or indirectly via a source-separation process, for a pitch mechanism.

A note on terminology: I use the term frequency *variation*, as opposed to the more common frequency *modulation*. While the words modulation and variation both mean a change from one state to another, the term frequency modulation has come to signify a periodic or nearly-periodic variation in frequency. Typically such

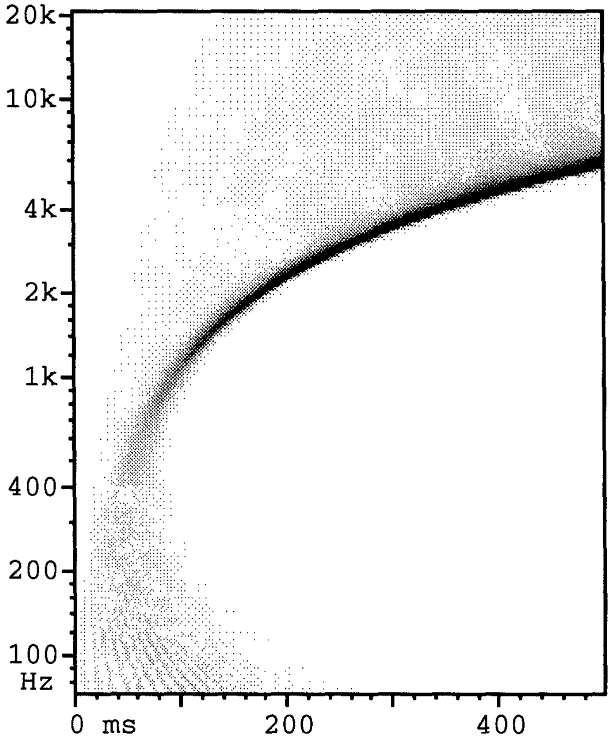


Figure 2.7: A tone rising continuously in frequency.

periodic modulation is used to introduce rich timbre in electronically synthesized music, to enrich a vocal or instrumental tone with vibrato, or to encode information in a transmitted signal. To emphasize that the frequency motion in question here need not be periodic, the term frequency variation is preferred.

Psychoacoustic Evidence

Chowning [Chowning80] noticed that a synthesized singing voice — a vowel sound — became more coherent when identical slight vibrato was added to the synthetic partials. This suggests that there is a mechanism which uses the vibrato as a feature for grouping. Lacking this feature, the synthesized partials sound like a collection of buzzy separate sounds. With it, the partials crystallize into a single voice in a sudden and striking way.

McAdams studied this micromodulation extensively [McAdams84]. He tested whether constant frequency *ratios* or constant frequency *differences* are important for source formation. For instance, if a partial at 220 Hz is frequency-modulated so it attains a maximum frequency of 224 Hz, then under constant-ratio modulation a partial at 440 Hz would go up to 448 Hz, while a constant-difference one would only go to 444 Hz. McAdams discovered that it is the constant ratios that have the strongest effect for grouping. Indeed, if a sound has sufficiently deep constant-difference vibrato, then the high partials split off from the low ones near the maximum excursion of frequency change to become a separate note. One way to restate McAdams's result is that constant partial spacing on a pitch-height scale is essential to spectral fusion.

McAdams's result is not too surprising in light of the behavior of natural sound sources. Suppose some process — a voice, an instrument, or any pitched sound — is producing a periodic signal by means of some vibrating medium. If the rate of vibration is increased while the waveform being produced stays constant in shape, then the frequencies of all partials will increase by the same ratio. So nearly all natural vibrato processes produce the constant-ratio variety, and it is not surprising that our auditory system is attuned to this type.

McAdams also tested the importance of FV for hearing one source in a mixture of sounds [McAdams89]. He synthesized sounds with three simultaneously sung vowels; each vowel could have FM imposed or not, and if so, this FM could optionally be coherent with that of another of the vowels. Subjects were asked to judge the prominence of a particular vowel in the mixture. McAdams found that such prominence judgements increased significantly when the target vowel was modulated, and suggested that an auditory grouping mechanism was coming into play in subjects' judgements. Surprisingly, prominence of the target vowel did not decline significantly in the presence of another vowel with identical modulation. One might think that partials of the other vowel might get grouped with the target vowel on the basis of FM coherence, and thus interfere with the prominence of the target. Lack of this result may be due to the fact that the main effect of FM is to make partials become a coherent source in the first place rather than remaining distinct objects. In other words, it may be that the difficulty of the listening task is not in separating the vowels from one another, but rather in having the partials form a source at all.

Rasch [Rasch78] played notes with and without 4% vibrato against a background of a masking tone. He found that tones could be decreased in level by 17.5 dB and still be equally audible against the background when vibrato was added. This amount of vibrato, typical of musical instruments, makes notes stand out even when they are much less intense than a masking tone.

Carlyon and Stubbs [Carlyon89] did a number of experiments to determine whether changes in the frequency of one partial of a harmonic or inharmonic complex were heard because of direct frequency change, or because the partial became mistuned with respect to the rest of the complex. They found that harmonic complexes were more easily heard than inharmonic ones against a noise background, and that FM of a partial is also more easily heard in the harmonic case. Their conclusion could be seen as suggesting that response to FV is solely dependent on harmonicity, but this would be a mistake: They claim only a rôle for harmonicity, but don't deny a rôle for FV. Both are probably operating as grouping factors.

Kay and Matthews did an interesting adaptation study of frequency-modulated tones. They found that playing a modulated tone of one FM depth caused subjects to become less sensitive to FM at a shallower depth, provided that the shallower modulation was at approximately the same rate as the deeper. When the two modulation rates were the same, the threshold of modulation of the test tone had to be about three times as deep after adaptation as before. This suggests that there are units in the auditory system which are susceptible to adaptation and which are sensitive to a certain rate of FM. The effect was not due to pure-frequency adaptation, for Kay and Matthews tried adapting with plain and amplitude-modulated sine tones, and did not get as strong an adapting effect as to frequency-modulated ones.

One hypothesis about the salience of frequency-varying sound is that it is due entirely to the passage of a partial into new frequency channels as it moves. The idea is that there are units in the auditory system especially responsive when a sound first enters a frequency channel — onset units. Perhaps there are not really FV units in the auditory system, just onset units. Gardner and Wilson tested this hypothesis with a clever adaptation study [Gardner79]. They played a series of tones which swept downward in frequency until the subject had become adapted to such tones and his or her threshold for hearing them increased. They then played an upward-sweeping tone through the same frequency band. If the subject were hearing frequency sweeps only because of onsets in channels, then after adaptation to down-sweeps, an up-sweep would have the same threshold as a down-sweep. This was not the case. The adaptation-inducing down-sweeps affected only the threshold for down-sweeps, not for up-sweeps. This result, and the analogous one for down-sweeps after adaptation to up-sweeps, suggests that there are mechanisms sensitive to the direction of frequency movement. It would be an interesting experiment to see if the same holds true of one rate of sweeping versus another.

Neurophysiological Evidence

Evidence for auditory neurons sensitive to FV comes from several sources. The earliest is probably Whitfield and Evans [Whitfield65], who recorded units in the auditory cortex. The found many units more responsive to FM tones than to any steady tone, and some that responded only to FM. They also found that FM-responsive units tended to fire at a particular point in the modulation sinusoid — for example, when the tone was rising. This led them to conclude,

> In nearly all cases it was found that the tone occupied part or all of one particular half of the modulation cycle, *i.e.*, a half-cycle whose limits were the maximum and minimum deviation of frequency. This firing occurred at a certain frequency or range of frequencies within the modulation cycle, but only when one frequency was changing in one particular direction.

Whitfield and Evans also found units responsive to wide frequency sweeps but not to steady-state tones or to narrow FM, and units more responsive to sweeps in one direction than the other. The latter echoes Gardner and Wilson's psychoacoustic study.

Møller extended the above study in a series of experiments with sweep tones [Møller77], recording responses in the cochlear nucleus. He swept tones at different rates and discovered that for some units, an "optimal rate of frequency change exists" in eliciting a neural response from a unit. These units show a steady increase in firing rate as the sweep rate is increased up to a point, followed by a steady decrease as it is further increased. In other words, units are tuned to filter an optimal FV rate. One may object that these rate-specific units were sensitive only to a tone whose frequency stayed within a given range for only a given period of time. Møller overcame the objection with tests of a number of tones of short, varying duration, showing that these tones did not excite the units as much as FV sweeps.

Mendelson and Cynader [Mendelson85] also found units in the cat auditory cortex responsive to the direction and rate of a frequency sweep. Based on responses of successive units in neural penetrations, they suggest that there may be a columnar organization to the auditory cortex — a map of FV sweep rate. They also report the somewhat surprising result that, at least in the cat cortex, there are about twice as many units responsive to downward sweeps as to upward ones. It is an interesting question whether the latter predominance is present in humans, and if so, whether it has had any consequence for spoken language sounds, perhaps being ties to an increased incidence of downward glides.

2.8 Harmonicity

Harmonicity refers to the degree to with which a partial falls into a harmonic series with other partials. It is important in the source separation process because many sound sources in nature have harmonic or nearly harmonic partial structures resulting from their origin in a vibrating medium. Since different vibrational sources usually vibrate at different frequencies, their partials form separate harmonic series, enabling a scene analysis mechanism to distinguish them.

The relationship between harmonicity and perception of harmony, or pleasantness, is a complex one that I will touch on only briefly here. The existence of beautiful pieces of music that rely on partials with non-harmonic spacing (*i.e.*, spacing other than integer multiples of the fundamental) demonstrates that perception of pleasantness is not completely tied to harmonicity. Such pieces as *Stria*, by John Chowning, and *Inharmonique*, by Jean-Claude Risset, exemplify the beauty possible in inharmonicity. Bregman puts forth the interesting idea that we have two separate mechanisms for hearing harmony, one that gives rise to a perception of consonance or dissonance, and a second that produces fusion of partials into a sound source [Bregman90, p. 508]. This is closely related to the observation of Moore *et al.* that low-frequency mistuned partials "appear to stand out from the complex tone as a whole, [while] for high harmonics, the mistuning is detected as a kind of 'beat' or roughness" [Moore85b]. Shepard has investigated the rôle of harmonic relations in musical structure [Shepard82], and Balzano suggests a possible underlying group-theoretic basis for 12-tone harmonic relations.

Or perhaps, as Mathews and Pierce suggest [Mathews80], harmony is merely "a matter of brainwashing."

Psychoacoustic Evidence

The strength of harmonicity as a grouping cue is well enough established psychoacoustically that it is usually used as a basis against which experimenters measure competing forces for segregation. For instance, McAdams' thesis work (summarized on page 37) starts out with harmonic tones, on which he imposes (in one experiment) different types of frequency modulation. One type of modulation, constant-difference modulation, introduces inharmonicity into a tone. The tone breaks into two sources when a sufficient modulation depth is reached [McAdams84].

Moore *et al.* performed a similar type of experiment. Their aim was to see how much inharmonicity in a partial is tolerated by the auditory system before the partial becomes a separate source. Again, harmonicity is presumed to have a grouping effect, and it is deviations from harmonicity that are the independent variable in the experiment.

Cohen [Cohen84] also used harmonically-spaced partials as a standard against which to measure inharmonically-spaced ones. Her work showed that pitch perception is not completely tied to harmonicity, in that subjects were able to assign consistent pitch values to tones with small departures from harmonicity.

Also, people studying perception of multiple simultaneous vowels have typically used the harmonicities present in the vowels as the main, or only, grouping cue [Parsons76, Weintraub85, Assman89b].

Duifhuis *et al.* [Duifhuis82] suggested, as part of a pitch-detection mechanism, the idea of a harmonic seive. In processing a sound with many partials, this sieve filters out those partials that best fit a harmonic series. Such a sifting process could underlie a harmonicity filtering mechanism, one that responds according to the fitness of its input sound to a harmonic series.

Neurophysiological Evidence

I know of no work that reports finding neurons in the auditory system which respond to harmonicity — *i.e.*, which integrate information from widely-spaced parts of the spectrum and respond only if the frequencies present are in a harmonic series. It is not even clear how harmonicity would be represented, since it must tie together different parts of the spectrum that are normally represented at different places at each stage of processing. Perhaps the cortical oscillation theory to be discussed in chapter 5 plays a part.

One element of physiology does seem important for its potential use in harmonicity and pitch detection: phase locking. The synchrony of neural firings on the basilar membrane to the peaks in a periodic sound signal could be used at higher levels to detect harmonicity at frequencies up to 4-5 kHz, as units responding to frequencies that are multiples of a fundamental will stay in a constant phase relationship with one another. Perhaps this relationship — or a slowly-varying relationship for nearly-harmonic partials — is the basis of a harmonicity or pitch mechanism.

2.9 Amplitude Variation

Common amplitude variation is the parallel variation in intensity of a number of partials. It derives its importance as a grouping cue from the fact that most physical processes that change the intensity of one partial will change the intensity of all of the others at the same time.

Psychoacoustic Evidence

Békésy found that sine waves at 750 and 800 Hz, when presented to opposite ears, could be made to form a single "sound image" by the imposition of coherent AM of 5 to 10 Hz [Békésy63].

The strongest line of evidence for the salience of common amplitude variation comes from the series of co-modulation masking release (CMR) experiments begun by Hall *et al.* [Hall84].

The fundamental discovery was that noise bands that vary in amplitude *coherently*, with the same modulating function, are not as effective at masking other within-band signals as those that are modulated *incoherently*. In other words, the signal which was not modulated coherently with the noise bands was more easily separated when the other parts were coherent than when they were incoherent. The suggestion is that a scene analysis mechanism is better able to hear out a target tone from within its masking noise band because it can use AM evidence from flanking noise bands to group the noise bands together, leaving the signal more perceptible because it is a separate source.

Many experiments have followed this initial discovery, only a few of which will be mentioned here. Moore [Moore90] has a recent review. Cohen and Schubert [Cohen87] measured the frequency range over which a flanking noise band contributes to CMR, finding it to be about 2/3 octave below the signal frequency and 1/3 octave above. They also showed that when the signal is coherently modulated with the masker it is less easily detected than an equally loud uncorrelated component; in this case, a scene analysis mechanism would not have be able to place the signal in a separate source, and so could not contribute to hearing it out.

Schooneveldt and Moore [Schooneveldt87] tested CMR under a wide variety of stimulus conditions, finding a weak dichotic effect across a wide spectrum of flanking bands and a stronger effect, present only for monotic stimuli, restricted to frequencies near the signal frequency. The latter sharply-tuned component of CMR worked best at high frequencies, in the range 2-8 kHz. They also found that CMR happened more when both noise bands were near the same signal frequency, and that it worked better for narrow (25 Hz) noise bands than wide (100 Hz) ones, especially for the sharply-tuned component. They suggest that beating between carrier frequencies of the two noise bands produces sharply-tuned CMR — that the auditory system listens to the target signal in the silent periods of beats. This is supported by the fact that CMR fails for frequency-modulated, instead of amplitude-modulated, sounds [Schooneveldt88]. However, Schooneveldt and Moore's finding of dichotic CMR that cannot depend on within-channel cancellation suggests that a mechanism similar to one useful for scene analysis may also be present. This and other data led Moore to conclude that CMR probably results from

> the operation of flexible mechanisms which can exploit a variety of cues or combination of cues depending on the specific stimuli used [Moore90, p. 135].

Bregman *et al.* [Bregman85] performed another A-B-C study like that in fig. 2.6 above. This time, the cue used to integrate B and C was common amplitude modulation, while the competing "force" that tended to group A and B was again frequency proximity. The conclusion they reached from the results of their experiment was that the AM rate of B and C did significantly affect their tendency to fuse.

Neurophysiological Evidence

Schreiner and various co-workers have performed an extensive series of recordings of neurons responsive to amplitude change at various places in the auditory system. These neurons could be part of a mechanism that uses amplitude modulation as a cue for source separation, though there is no evidence as yet for this function. Schreiner and Urbas found that units in auditory cortex could be characterized by a *best modulation frequency* (BMF) as determined by phase-locking of neural firings to the signal, and that BMF varied in a consistent way with frequency [Schreiner86]. They also found [Schreiner88b] that units varied in sharpness of tuning, some being tuned to relatively exact AM rates ($Q_{10dB} = 5$) and some to relatively broad ones ($Q_{10dB} = 0.5$).

Langner and Schreiner, in studies of the central nucleus of the inferior colliculus, also found a BMF function for rate of neural firing which closely followed that for phase-locking. They found [Langner88] that units, in addition to varying by BMF and sharpness, varied in response time from 4 to 120 ms, with 95% of them responding in less than 21 ms and all but two in less than 30 ms. Also, they found that units with higher BMF's tended to have faster response time, and conversely. In a second study, they found an arrangement of units ordered by BMF spatially across the inferior colliculus, with highest BMF's at one point and progressively lower ones concentrically around it. Sharpness of tuning also varied in an orderly concentric map, but with a different center than the BMF map. Thus each unit embodies a different BMF-sharpness pair.

2.10 Location

The ear performs a significantly difficult task in deciding what spatial location a sound originates from. Spatial location is a significant cue for sound separation, though not an indispensable one. To convince yourself of the latter, simply listen to two people talking at once with one ear covered, or better yet, listen over headphones or a loudspeaker. A few personal demonstrations have convinced me that a voice is usually readily distinguished from another voice under these conditions unless the competing voice is quite similar, and even then, only in rare cases do they fail to be separable.

Reverberation and the Precedence Effect

The problem of locating a sound in space is complicated by the presence of reverberation and room resonances. The reflections of reverberant sound make measuring differences in time of arrival between the two ears much more difficult, as it becomes uncertain whether a particular feature — a rise in amplitude, for example — is a direct or reverberant sound. The auditory system likely has some mechanism for gaining information from the effects of reverberation in order to characterize the physical environment and possibly for better localizing a sound.

The auditory system has a mechanism to remove the effect of reverberation, which is evinced by the *precedence effect* [Henry51] [Hall36] [Fay36] [Zurek80][Wickesberg90]. This mechanism apparently makes the auditory system insensitive to a given micropattern in time and frequency for roughly 30 ms after that the onset of that pattern [Gardner68]. This duration, corresponding to a 10 m round-trip distance, is roughly the reflection time of walls in a medium-sized room, and provides increased clarity to the listener. A good demonstration of the power of the precedence effect can be done by recording a voice or music in any room. Listen to the recording, then listen to it played backwards. Reverberant effects not noticeable in the normal situation stand out clearly in the time-reversed sound [Houtsma87, no. 35].

Localization Cues

There are a number of different ways that the auditory system may detect features for sound location. For lateralization (left-right position), it can use an interaural level difference (ILD). A sound on one side of the head is usually more intense in the near ear than the far ear. (An exception can occur in reverberant rooms, where harmonic nodes may create local loud spots or holes of certain frequencies at various places in the room.) ILDs are normally frequency-specific, as more high-frequency energy normally is lost due to blocking by the head. Also, the outer ear or pinna has a complex frequency- and location-dependent filtering effect that makes available a great deal of information to the auditory system, the only information available for vertical (elevation) localization.

Another localization cue is interaural time differences (ITD). Waves from a sound on one side of the head arrive at the nearer ear sooner than at the far ear, providing a localization cue. At low frequencies, ITD is expressed in differences in phase of the waveform arriving at the two ears. Phase information is communicated to higher regions of the auditory system by the phase-locking of neural firings to peaks in the waveform. At higher frequencies, roughly above the 1 kHz half-wave size of the head, phase information is ambiguous — the auditory system has no way of knowing which waveform peak of a given frequency at one ear matches a given peak at the other ear. It can compensate for this ambiguity by using information from many different frequencies: The ITD for each frequency constrains the location to a set of possible places, and the intersection of these sets for all frequencies present may disambiguate the location. Unfortunately this will not work for periodic sounds with fundamental frequencies above 1 kHz, for then all of the constraint sets of different frequencies are supersets of the fundamental's set. No additional information is gained. In this case, ITD information must come from changes in the amplitude envelope of the sound. An amplitude change can be readily detected and its ITD measured for localization. Tobias and Zerlin [Tobias59] measured human jnd's for sounds arriving at the two ears at slightly different times, finding sensitivity to an ITD of just 6 μs. Yin and Chan [Yin88] point out a number of specialized physiological features between the ear and the superior olivary complex, where localization is thought to be done, that preserve such amazing time resolution.

In addition to ILD and ITD, the cue of ratio of direct to reverberant sound may be used to determine the distance of a sound. Chowning [Chowning70, Chowning74] notes that this ratio is an important perceptual feature and may also be used to determine characteristics of the environment such as room size. He has an interesting demonstration of the effect. Sheeline made a more thorough study of the effect of reverberant sound in distance perception, concluding that "reverberation has an absolute effect on perceived distance, independent of individual subjective differences or response biases." [Sheeline82, p. 62]

Psychoacoustic Evidence

A series of experiments with a common theme support the idea that spatial location is an important grouping cue. An experimenter plays a signal and a masking noise diotically and measures the signal level threshold necessary to hear it in the noise. He or she then introduces a delay into either the signal or the masker in one ear and measures the threshold again. The threshold is usually lower in the second case; the difference in thresholds is called the masking level difference (MLD).

MLD was discovered by Jeffress *et al.* [Jeffress56]. Several explanations have been advanced for it [Durlach64, Jeffress72, Hafter71]. The one that seems to have survived is that the delay in one ear makes the auditory system localize the signal and masker at different places. Thus, the signal is better separated as a source and can be heard more clearly. Jeffress [Jeffress72] proposed a cross-correlation technique, implementable as neural delay lines, that could provide a means for source separation based on spatial location. This model has recently been extended and implemented on a computer by Lindemann [Lindemann86], who added lateral inhibition and a contribution from a monaural filter to produce a method that appears to work well for localizing monophonic signals and looks like it should work for polyphonic ones as well. This could potentially be the basis for a source separator.

Durlach's Equalization and Cancellation theory [Durlach63] [Durlach64] [Metz68] suggests that MLD results from a process that shifts the mental representation of the stimulus in one ear in time and/or amplitude until it matches the stimulus in the other ear as well as possible. The matching part of the sound — which is the noise in the MLD effect — can then be subtracted (cancelled) from each ear, leaving behind the target signal. Such a mechanism, if present, could be considered a part of the scene-analysis process that separates sources by spatial location.

A number of other experiments have been conducted in which different signals are played to the two ears. The presumption is that the two signals are localized to opposite sides of the head, providing a scene-analysis mechanism with a strong cue for distinguishing the two signals as being separate. Van Noorden [van Noorden75] played identical tones to the two ears and found it difficult to tell if the tones were evenly spaced in time, that is, if the right-ear tones were exactly half way between the left-ear tones.

Recently, Lakatos [Lakatos91] performed a similar experiment, playing alternate tones at different locations 30° apart in space vertically and horizontally. At sufficiently high rates, it becomes impossible to order the tones that come from different locations. Each sequence of tones at one location breaks off from the other locations and becomes its own stream.

Neurophysiological Evidence

Evidence for extraction of information for spatial localization is well established. Knudsen's recordings in the owl have established topographic maps in the magnocellularis lateralis pars dorsalis, which corresponds to the inferior colliculus in humans. These structures map out the spatial field in the space around the owl's head [Knudsen78, Knudsen81]. In the two-dimensional map of the magnocellularis seen in fig. 2.1, each point corresponds to a specific azimuth and elevation in the sphere around the owl's head at which a sound may be located. A unit in this map increases its firing rate when a sound occurs at the unit's specific location.

Yin and Chan [Yin88] present evidence for neural units in the cat that are probably precursors for the type of space-map found by Knudsen. They found units responsive to ITDs in the lateral superior olive and review evidence for other units responsive to ILD. One example of the latter is the experiment of Takahashi *et al.*, in which different cochlear nuclei were anæsthetized [Takahashi84]. Numbing the nucleus magnocellularis changes the owl's neural response to ITDs but leaves ILD response intact, while numbing the nucleus angularis produces the converse effect.

Chapter 3

Computational Model

The preceding chapter suggested a number of low-level features used by the auditory system for scene analysis. Here I wish to find physiologically compatible algorithms to filter some of these features and to implement some of the algorithms on a computer.

Which Cues?

The first question of course is which features to extract. Common onset time is one of the most salient cues, as noted by Helmholtz:

> Now there are many circumstances which assist us first in separating the musical tones arising from different sources, and secondly, in keeping together the partial tones of each separate source. Thus when one musical tone is heard for some time before being joined by the second, and then the second continues after the first has ceased, the separation in sound is facilitated by the succession of time. [Helmholtz54]

Another salient cue is common location. Algorithmic work on this cue has, for the most part, used cross-correlation techniques to find interaural time differences as noted above. The problem is made difficult by reverberation in real-world listening situations. Also, auditory work on spectral cues for interaural level differences requires a physical head-and-ear sculptural model, making it difficult to do. For both of these reasons, I have chosen chosen not to study common location here.

Harmonicity and related pitch extraction work have been extensively studied, with quite a few pitch theories, extensions to theories, and implementations extant [Licklider51], [Terhardt72], [Goldstein73], [Wightman73], [Duifhuis82], [Scheffers83], [Moore85a], [Pierce90]. Though harmonicity is an important cue for scene analysis, fundamental frequencies in musical tones are often harmonically related. This leads to the coincidence of the partials of these tones, complicating the tasks of feature filtering and event formation. Also, the excellent work done by Cooke [Cooke91] contributed much to the use of this cue for separation, convincing me to leave it aside here. It is described more fully in chapter 6.

Common amplitude and frequency variation are also important scene analysis features. Cooke also studied amplitude variation fairly thoroughly; I have not duplicated his work here. The salience of frequency variation as a separation cue has been well-established by the psychoacoustic experiments mentioned above.

So the features to be modelled with implementation-level processes are common amplitude onset and frequency variation.

3.1 The Cochlear Model

The first step in the implementation of auditory processes is a model of the filtering and transduction performed by the ear, including the cochlea. Many ear and cochlear models have been proposed, as mentioned above, and they all have in common the processes described below.

First, a filter models the resonances and refraction of the outer ear and head and the impedance matching of the middle ear. Next, a bank of filters, either parallel or cascaded, extracts the energy present at each of the cochlear frequency channels, which are spread roughly logarithmically in frequency (some models, following the cochlea, have wider channels at low frequencies). Typically, each channel covers a frequency band of about 1/3 octave, roughly corresponding to a critical bandwidth [Scharf70, Moore83, Glasberg90], though channels overlap so that there are more than three of them per octave. Then a hair cell model imposes some sort of nonlinearity on the signal in each channel. This filter must at least make the signal non-negative, since the output represents the probability of neural firing at each time sample. For this non-linearity, different cochlear models and descriptions use variously half-wave rectification [Lyon82, Shamma85], an arctangent function [Seneff88], and other sigmoidal functions [Meddis89, Cooke91].

The ear/cochlea model used for the experiments here is the Lyon model, as implemented by Lyon and Slaney [Lyon82] [Lyon84] [Lyon86] [Lyon88] [Slaney88] [Slaney90]. This one was chosen because of its four-stage automatic gain control, which compresses its output signal to a dynamic range which is better suited to neural representation. Though some other models are more accurate at characterizing the dynamics of neural populations (*e.g.* Cooke's State Partition Model [Cooke91]), none of them models as many of the various mechanisms present in the ear for gain control as Lyon's. The Lyon/Slaney program also computes the autocorrelogram from the output of the cochlear filter channels, which was convenient.

3.2 Filtering Amplitude Onset

An amplitude onset is a rapid rise in sound amplitude over a short period of time. It is represented in a cochleagram by a rapid rise in neural firing rate. Such onsets can be seen in fig. 3.1, which is a half-second slice of time from a piano performance of a toccata by Frescobaldi. Onsets are easily seen, places where there is no sound energy in a channel at one time and then a harmonic piano partial begins. The problem is how the auditory system responds to such a change, and how to model it.

The method employed here is a simple cross-correlation operator. The cross-correlation of two linear functions f and k is defined by

$$c(t) = \int_{-\infty}^{\infty} k(x) f(t+x)\, dx.$$

The function $k(t)$, called the kernel, must be chosen carefully by the researcher to produce the desired result. Cross-correlation is physiologically compatible in that for fixed k its output may be thought of as a weighted sum of its inputs (or delayed inputs, if time were not spatialized), and such summation is within the functional capability of a neuron.

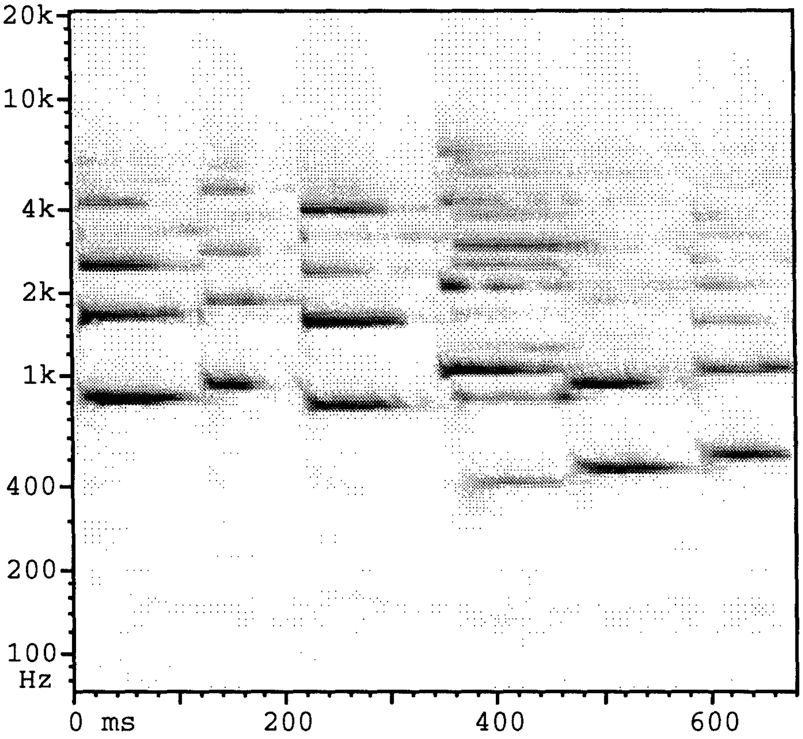


Figure 3.1: Cochleagram showing onsets at about 0, 120, 220, 340, 360, 470, and 580 ms.

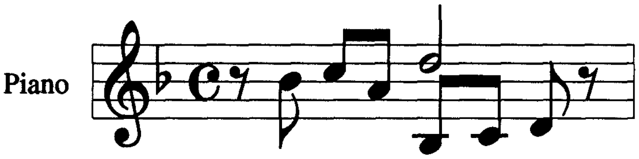


Figure 3.2: Score for the previous figure.

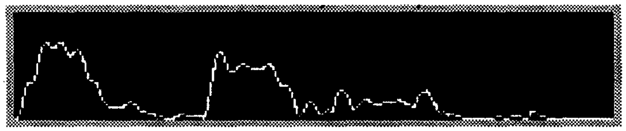


Figure 3.3: A cochlear channel with two pronounced onsets.

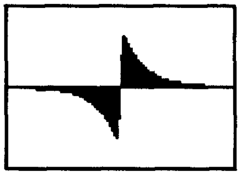


Figure 3.4: Onset kernel. Horizontal axis is time, vertical axis is the function value.

The cross-correlation function used here operates on a single channel of the cochleagram — a single horizontal slice in the image above. Fig. 3.3 is single horizontal slice from the image of piano music above. The kernel k of this correlation operator is chosen to filter onsets. It is the representation as a correlation kernel of a time-differentiation operator [Adelson86], and looks like fig. 3.4. In this figure, the horizontal axis is time and the vertical axis is the value of the kernel function k (the kernel is one-dimensional; the vertical dimension shows its value as it extends over the horizontal dimension.) In operation, this kernel is placed over the cochleagram at some point, then pointwise-multiplied and the resulting products summed to get the cross-correlation value. This multiply-add operation is repeated at each point in the cochleagram to get the onset feature map for a particular kernel.

The kernel has an upward spike which responds strongly to energy in the cochleagram, say at a time t_0. It also has a downward spike immediately preceding the upward one which inhibits response if there is any energy in the channel before the upward spike. This inhibition cancels out the response from the positive spike if sound existed in the channel immediately before t_0. In other words, the whole kernel responds only when there is no energy for a period of time and then energy appears

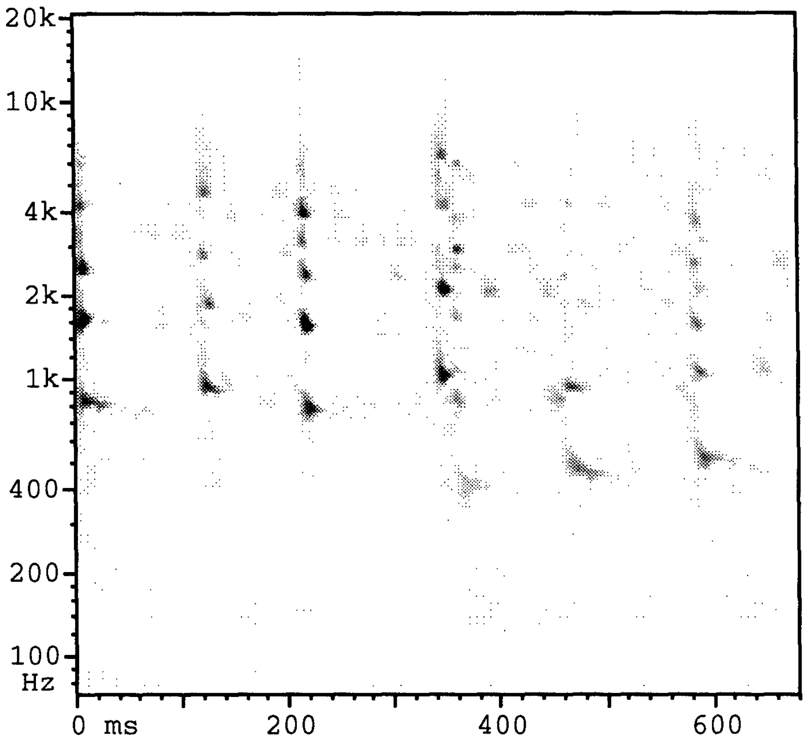


Figure 3.5: An onset map for the Frescobaldi piano excerpt.

— *i.e.*, at an amplitude onset.

A sample onset map, the result of this per-channel cross-correlation, is shown in fig. 3.5. (Actually this is but one slice, for a certain setting of the onset filter parameters, of a higher-dimensional feature map. I will use the term "feature map" both for a single slice like this and for the higher-dimensional object when it is clear from context which is meant.) This map shows strong responses at points at which a rapid rise in energy occurs in a channel. One can easily see the beginnings of partials in the piano music.

It should be noted that the onset filter described here responds to amplitude onsets but not *common* amplitude onsets. That is, it does not look for associations between different parts of the spectrum to find partials that have a coincident onset. All it

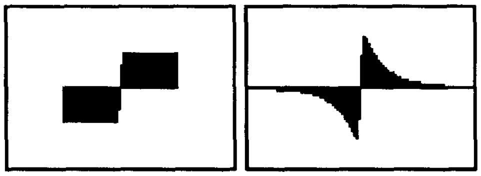


Figure 3.6: Onset kernels with rectangular (left) and exponential (right) functions.

does is find onsets in each spectral channel independently.

Onset Kernel Function

There are several characteristics of this onset filter, or operator, to investigate and tune for better response. One is the nature of the kernel function. The two kernels to be considered here both have a negative, inhibitory region followed by a symmetric positive, excitatory region; they differ in how they vary away from the central +/− spike. The simplest kind of kernel has just a rectangular response region that remains at a constant positive value for its duration. Another type has an exponential function decaying away from the center. Graphical views of both are in fig. 3.6, and feature maps computed with them (with other parameters otherwise constant) are shown in figs. 3.7 and 3.8.

A close examination of these two feature maps shows that fig. 3.8, made with the exponentially-decaying kernel, has sharper onset responses — the onset "blobs" are smeared across the image less. This sharpness persists for other values of the feature parameters that will be described shortly. Sharpness of response is a highly desirable property for maps that will be used by grouping mechanism, for the mechanism may need to distinguish events that happen near, but not at, the same time. The conclusion is that an exponentially decaying curve works best for onset feature filtering.

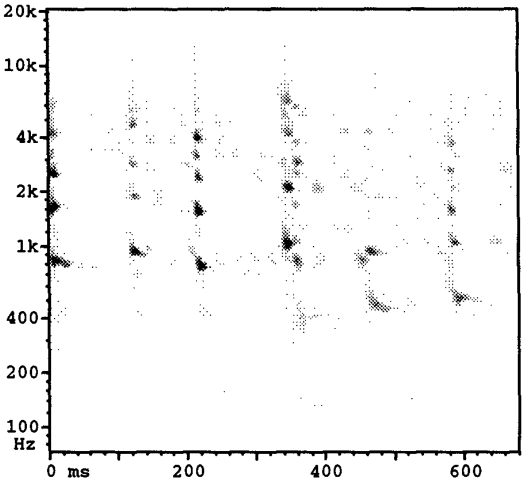


Figure 3.7: Onset map computed with rectangular kernel.

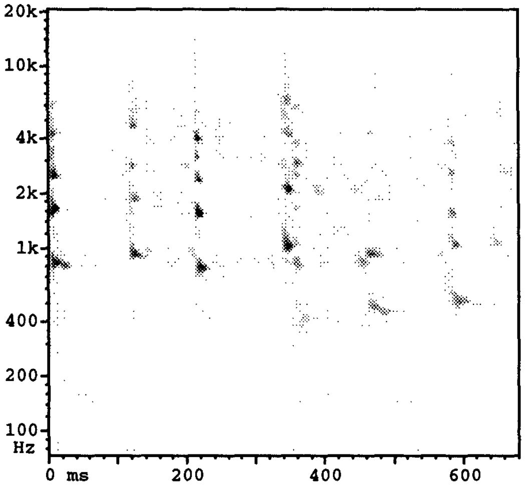


Figure 3.8: Onset map computed with exponential kernel.

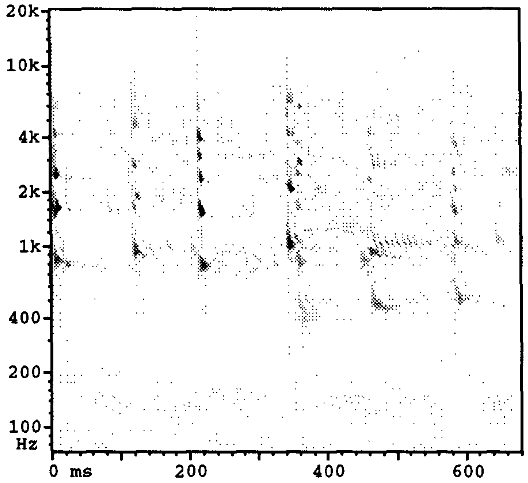


Figure 3.9: Onset map computed with 2.3 ms kernel.

Response Time

Another tuning parameter of the onset kernel is its response time. How quickly should the exponential curve decay away from its central +/− spike? Latency of onset responses in the cat auditory system vary from about 2 ms in the cochlear nucleus [Rhode86] to 20 ms in the auditory cortex area AI [Pickles88, p. 215]. Though a certain latency in a neuron does not strictly imply that the unit uses a characteristic decay of that amount of time — indeed, the use of an exponential curve with a characteristic decay is just a guess based on what works well — it does give a clue for which time constants are reasonable to try. Accordingly, I have used characteristic decay times (times to decay to $1/e$ of the peak spike height) from 2.3 ms to 18.1 ms. Resulting feature maps for some of the values in this range are shown in figs. 3.9-3.12.

Using Several Widths

Experiments with this onset filter have shown the value of using a range of widths. Different musical instruments, for example, have different attack times [Grey75,

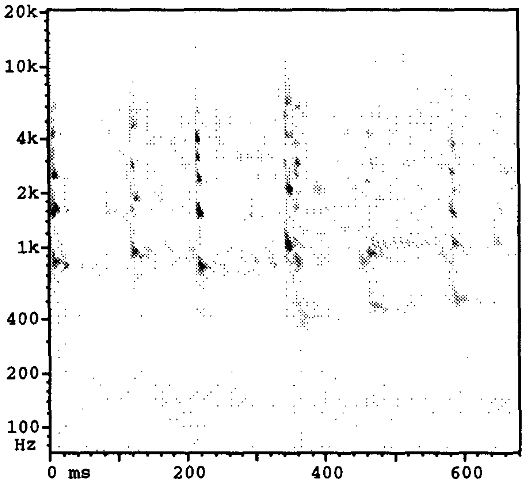


Figure 3.10: Onset map computed with 4.5 ms kernel.

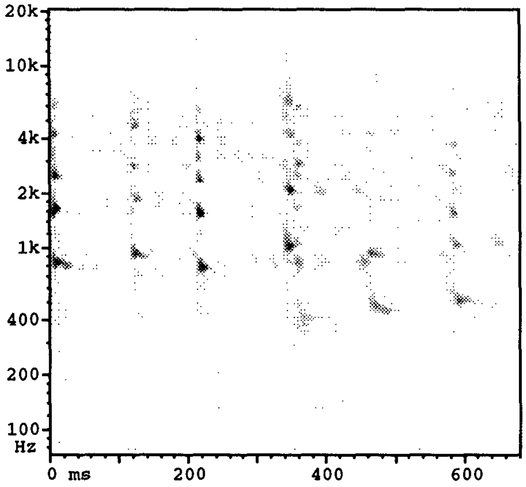


Figure 3.11: Onset map computed with 9.0 ms kernel.

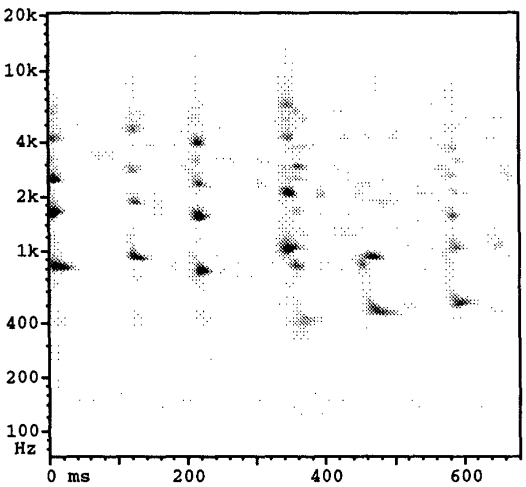


Figure 3.12: Onset map computed with 18.1 ms kernel.

Gordon84], and so respond best to different kernels. This, combined with the auditory system's range of onset response times, is a convincing argument that using several different characteristic response times is the best way to construct a filter to be used in scene analysis. Each characteristic response time is represented by a different kernel, so there are several kernels, one per response time. Accordingly, the onset feature map is a $2\frac{1}{2}$-dimensional, having a half-dimension ranging over a small number of different values (currently 4) of the characteristic decay time of the kernel.

One other tuning parameter affects the onset kernels. This parameter changes the relative height of the kernels. As seen in fig. 3.13, different decay values lead to different areas under the kernel curve. Without some kind of normalization, these areas will produce a wide dynamic range of responses, with the longest decay times producing the largest values. It turns out to be easier in the grouping algorithm described in chapter 5 to have onset response values with a smaller dynamic range. Accordingly, each onset kernel was normalized to have the same sum, as follows: The spike is sampled in time, so it consists of a discrete set of values. Since the spike is exponentially decaying, each value is a fraction of the previous; the ratio r between

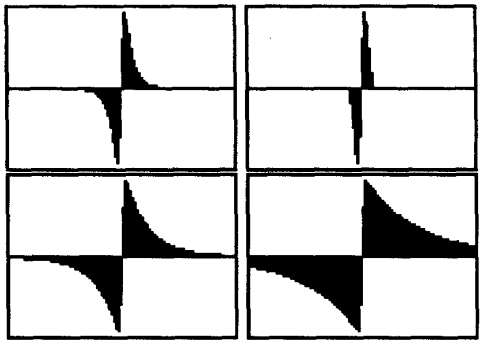


Figure 3.13: Onset kernels of different decay times.

points is given by

$$r = e^{-\frac{1}{d}}$$

where d is the number of samples in the characteristic decay time. The sum of values in the positive half of the onset kernel is then

$$\frac{m}{1-r}$$

where m is the value of the first sample. To make the values in the kernel sum to 1, m is simply normalized to $1 - r$.

Using this value of m, the output of the onset filter was still not well normalized; a longer kernel produced a higher peak values in its output map, even for fast percussive onsets. For this reason, an extra fudge factor of $\sqrt{d}$ was used. Thus m is defined by

$$m = \frac{1}{\sqrt{d}(1-r)} = \frac{1}{\sqrt{d}\left(1 - e^{-\frac{1}{d}}\right)}.$$

The maps in fig. 3.13 used this normalization.

An Artifact

This onset filter, using the above tunings, captures important onsets in the sound signal reasonably well. Unfortunately, it also picks out some features that do not correspond to perceptual onsets. This happens when a sound is varying in frequency,

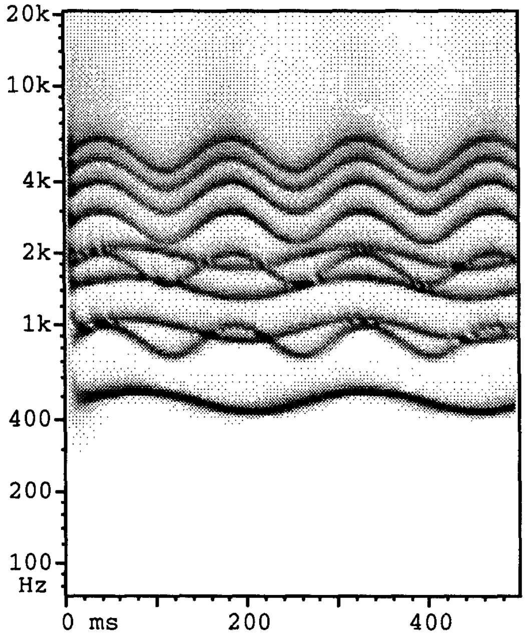


Figure 3.14: Two tones with different vibrato.

moving from one cochlear channel to another. As it enters a new channel, that channel has a sudden increase in intensity, an increase that will cause a strong response by the onset cross-correlation operator described above. The resulting response in the onset feature map may be seen in the example of fig. 3.14, which shows a sound made up of partials that are varying in frequency at various rates. A corresponding onset response is in fig. 3.15. As partials move up and down, they excite an onset response, despite the fact that the only onset noticeable perceptually is at the very start of the sound.

This spurious response can be considered either harmful or helpful. It is harmful because it may make the grouping mechanism, that is going to use this onset map as input data, believe that new sounds have appeared when in fact all that has happened is a change in frequency. This harmful response could be prevented by using a mechanism that causes inhibition of the onset response of a given channel

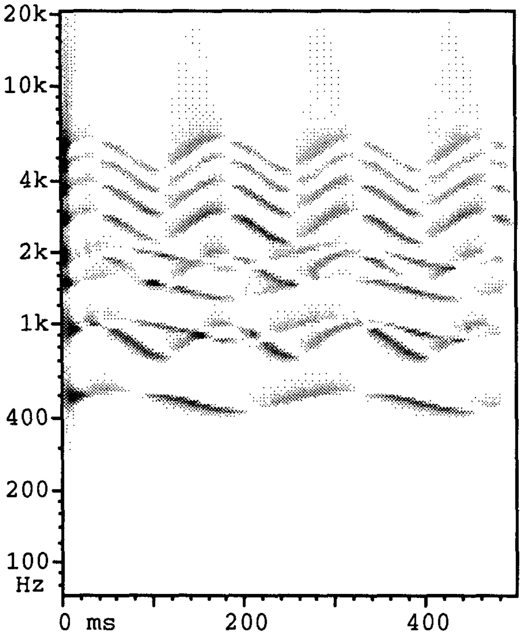


Figure 3.15: Onset map computed with 18.1 ms kernel.

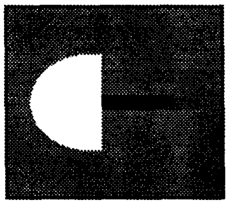


Figure 3.16: Kernel to eliminate frequency-variation artifacts. Horizontal axis is time, vertical is height.

whenever a neural response is present in either the same frequency channel or in nearby channels. In other words, the cross-correlation operator would no longer work on just a single neural channel at a time, but would have to examine channels with nearby frequencies for activity that inhibits the operator's response. This could be implemented fairly simply with a two-dimensional cross-correlation operator that has a negative region spreading over a few cochleagram channels, followed by a positive spike in the desired channel which responds to an onset. Fig. 3.16 shows a schematic idea of such a two-dimensional kernel. The white region on the left is a negative, inhibitory area, and the black and dark grey part on the right is a positive, excitatory region. The neutral grey in the background represents a zero weight.

It is possible that cochlear filters do not have a fixed center frequency as do the ones used here, but instead change with context to track moving frequencies. The output of such a filter would not usually exhibit the onset artifact described here. This idea is attractive for a mechanism for filtering frequency variation — the motion of the filter center frequency would give information about FV — but has not been implemented here.

Or Not an Artifact?

On the other hand, the spurious onset response caused by frequency variation may actually be helpful. The eventual use of the output of this onset filter is to tie together parts of the sound spectrum that have similar features. If two or more partials of a sound event are varying together in frequency, they trigger roughly identical responses at their respective frequencies from the onset filter. Such identical responses can aid in grouping these partials as a single sound event. This process is just the filtering of common frequency variation (see below) by filtering of common within-channel amplitude variation. On the other hand, two partials varying at the same rate in *opposite* directions in frequency also trigger such an amplitude response, leading to a common event grouping, when in fact these partials should probably not be grouped together.

The solution chosen here is to do nothing about spurious onset response to FV at this level, but to let the grouping mechanism described in chapter 5 make use of the available FV information to decide whether an onset really is spurious or not, and optionally to ignore it if so. This gives the grouping mechanism flexibility, because it is free to use FV-generated common amplitude onsets, or to ignore them; no information is lost at this level. (This is an application of Marr's "principle of least commitment" [Marr82, p. 106].)

Filtering of Amplitude Variation

By extending the kernel of this onset filter in time, it could be used as an amplitude variation filter for the sort of changes found in speech sounds and tremolo in musical ones. Indeed, there is no qualitative difference between an amplitude onset and amplitude variation. They are distinguished merely by the time scales involved, with onsets happening roughly within the time period found by Grey, about 40 ms, and amplitude variation happening at slower rates.

To filter amplitude variation, the kernel shape would probably need adjusting to make it respond to any variation rather than just a rapid spike of amplitude increase. In addition, speech sounds usually have components varying rapidly in frequency and amplitude at the same time. Such sounds would probably require a mechanism capable of tracking components that change in frequency, then applying the amplitude variation cross-correlator to those components.

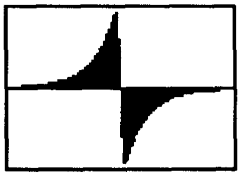


Figure 3.17: Offset kernel.

Offsets

Also, by inverting the onset kernel, the filter becomes an offset filter, as in fig. 3.17. Offset synchrony in scene analysis is not as important as that of onset synchrony; the example of fig. 3.18 may clarify why. This sound, the same few notes from the Frescobaldi toccata, has piano notes that were played without the sustain pedal and thus end nearly simultaneously when their keys were let up. Despite this clearly audible common offset, the offset feature map shows little evidence of the offset synchrony between the harmonics of the piano note; perhaps different harmonics of a piano note cease at different rates when the key is released. A small amount can be seen for the second note in the segment, which ends at about 190 ms, but not much of it, and almost none at all for the other notes. Kernels with other characteristic decays than the 18.1 ms one shown here produce maps that are no better, and are usually worse, at showing offsets.

Another example of difference between onsets and offsets may be heard when the Pierce example (fig. 2.5, p. 31) is played backwards. As the harmonics cease one by one, a slight change in the timbre of the tone complex is heard, but no tone stands out and the perceptual effect is generally much weaker.

Offset maps are not used by the grouping mechanism of chapter 5 that uses onset maps for event formation.

3.2.1 Results

Results of processing piano music with the onset filter were shown in figs. 3.9-3.12. Onset maps for several other instruments are shown in figs. 3.19-3.21. Each of these

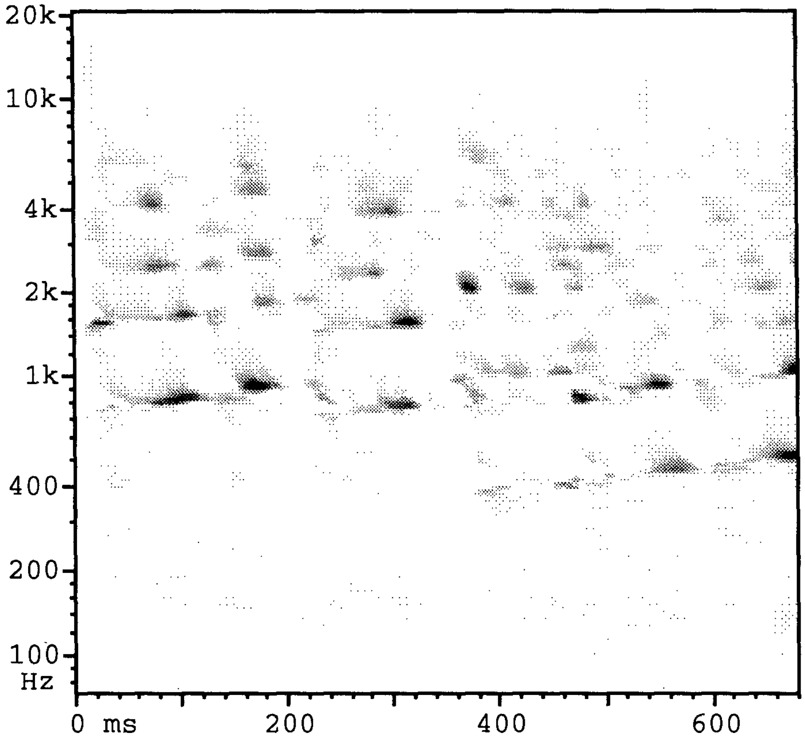


Figure 3.18: Offset map for the Frescobaldi sound.

figures shows results for two instruments, with a cochleagram and four onset maps for each one. All of these images are from a single tone played on the instrument; for the pitched instruments, it is middle C.

Of the five maps on each half-page, the upper left one is the cochleagram for the tone signal, then to the right are the 2.3 ms and 4.5 ms onset feature maps. The next line has the 9.0 ms and 18.1 ms onset feature maps.

The bowed string instruments (cello and violin) cause a strong response in the onset feature maps at tone onset, particularly with the longer-duration filter kernels. They cause strong onset responses at later times in the sound as well. Part of this response is due to vibrato in the tone, part of it to beating of the high harmonics within a single cochlear frequency channel, and part of it is unknown in origin. In particular, there are noise bursts in the 4-20 kHz range at various times in the violin sound that have yet to be explained. This tone was created by a single bow movement, so it is not due to a reversal of bow direction. Notable also is the low signal-to-noise ratio; partials, for instance, do not stand out as well for these instruments as for the others.

The brasses (horn and trumpet) have the clearest visible onset responses. The pitch of the played notes does not vary appreciably, so there are almost no onset responses due to the changing frequency of partials, and the instruments appear to have very little associated broadband noise; both of these factors contribute to a sharp onset response. The onset filters, down to a characteristic decay of 4.5 ms (for the horn) and 2.3 ms (for the trumpet), show strong responses in the output feature maps, providing a relatively precise time of onset. Also, these instruments have the best visible resolution of harmonics; one can count at least 10 of them in the cochleagram image.

The snare drum beat is, of course, a burst of noise across the spectrum, and as such generates some strong onset responses. The best responses to the snare are, surprisingly, at the longer decay times — 9.0 ms and 18.1 ms. The plucked string triggers response at even the fastest characteristic decay in the filter kernel; some of this response is probably due to the noise burst generated as a string is released.

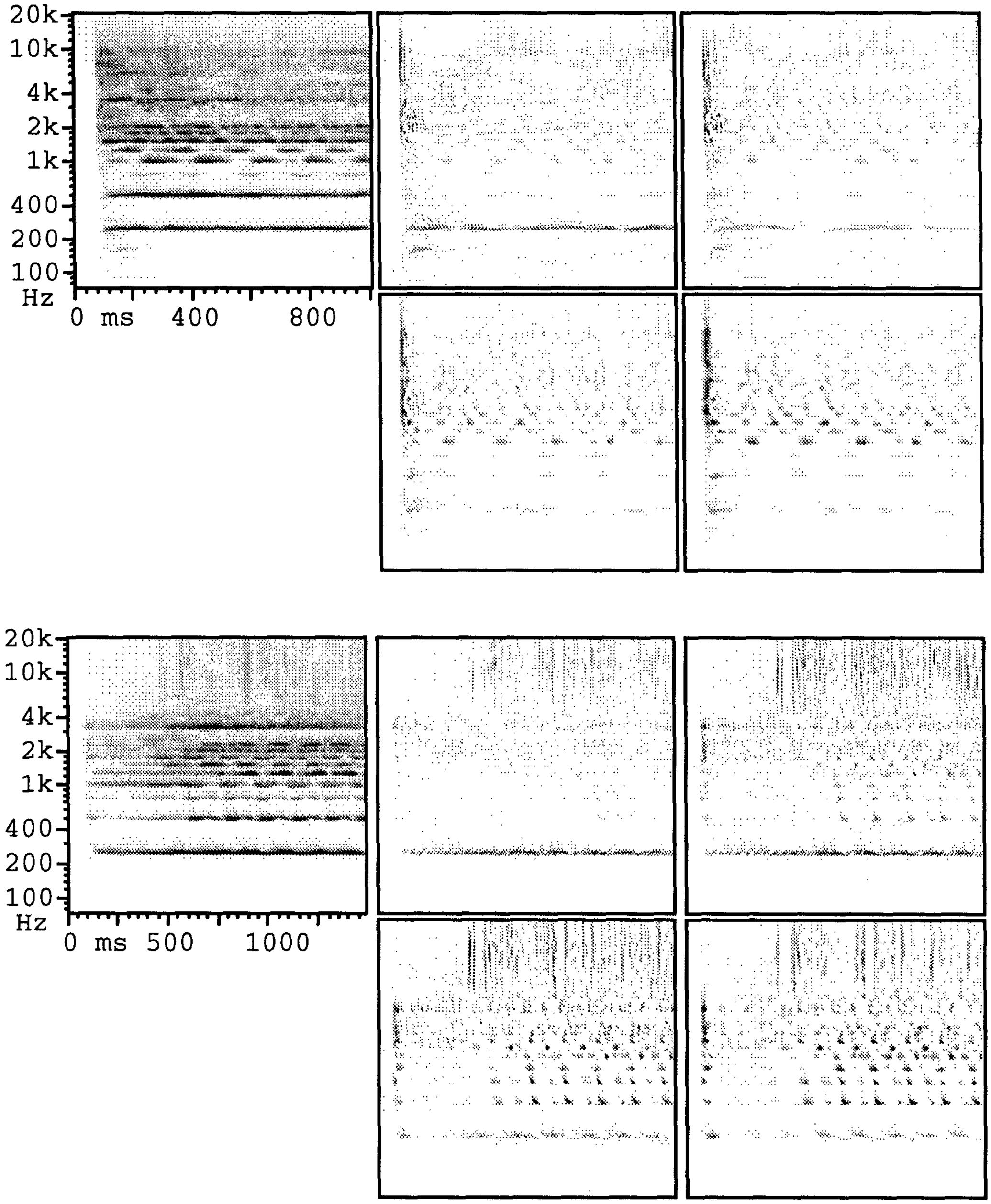


Figure 3.19: Cello (top) and bowed violin (bottom) onset responses.

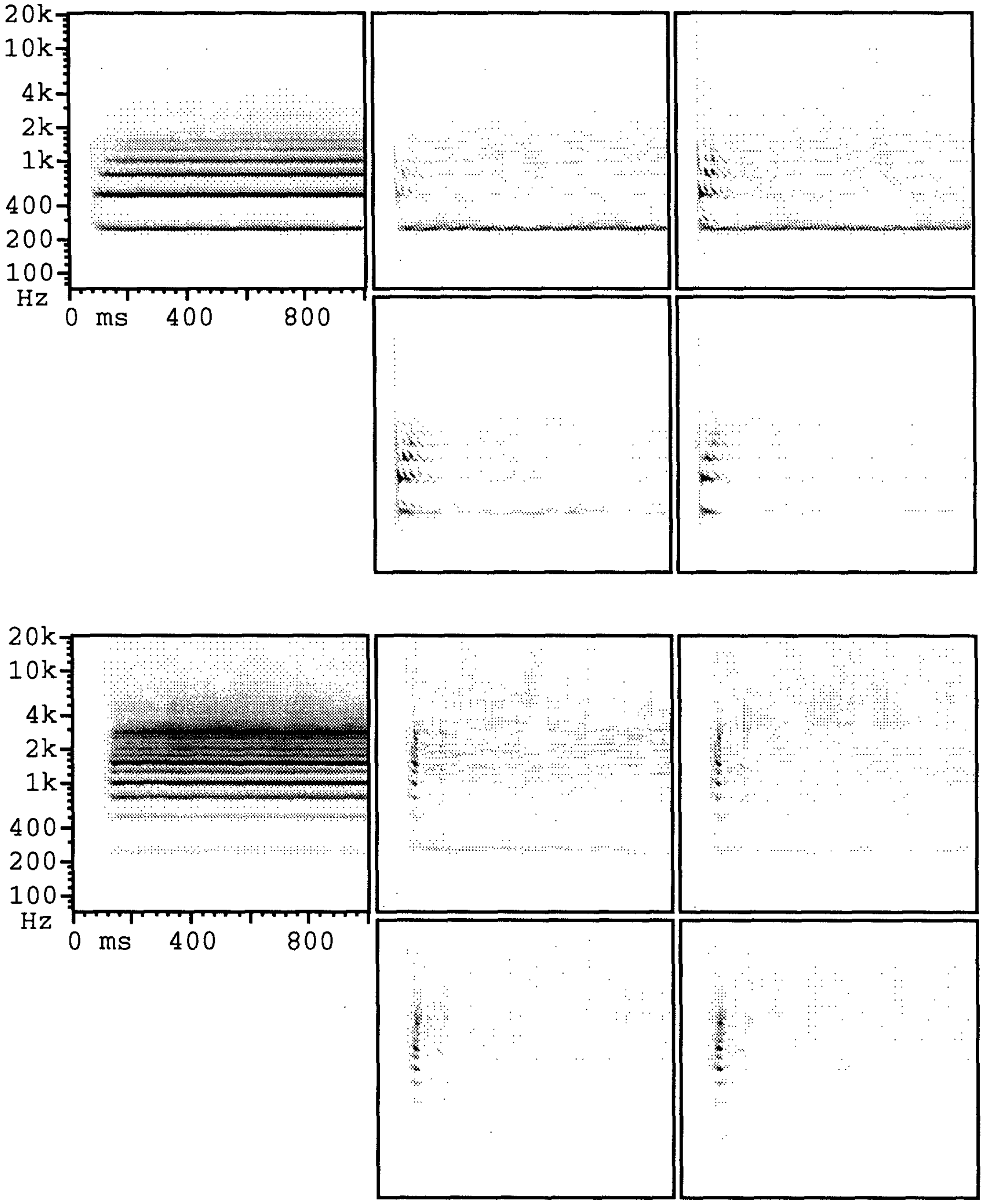


Figure 3.20: French horn (top) and trumpet (bottom) onset responses.

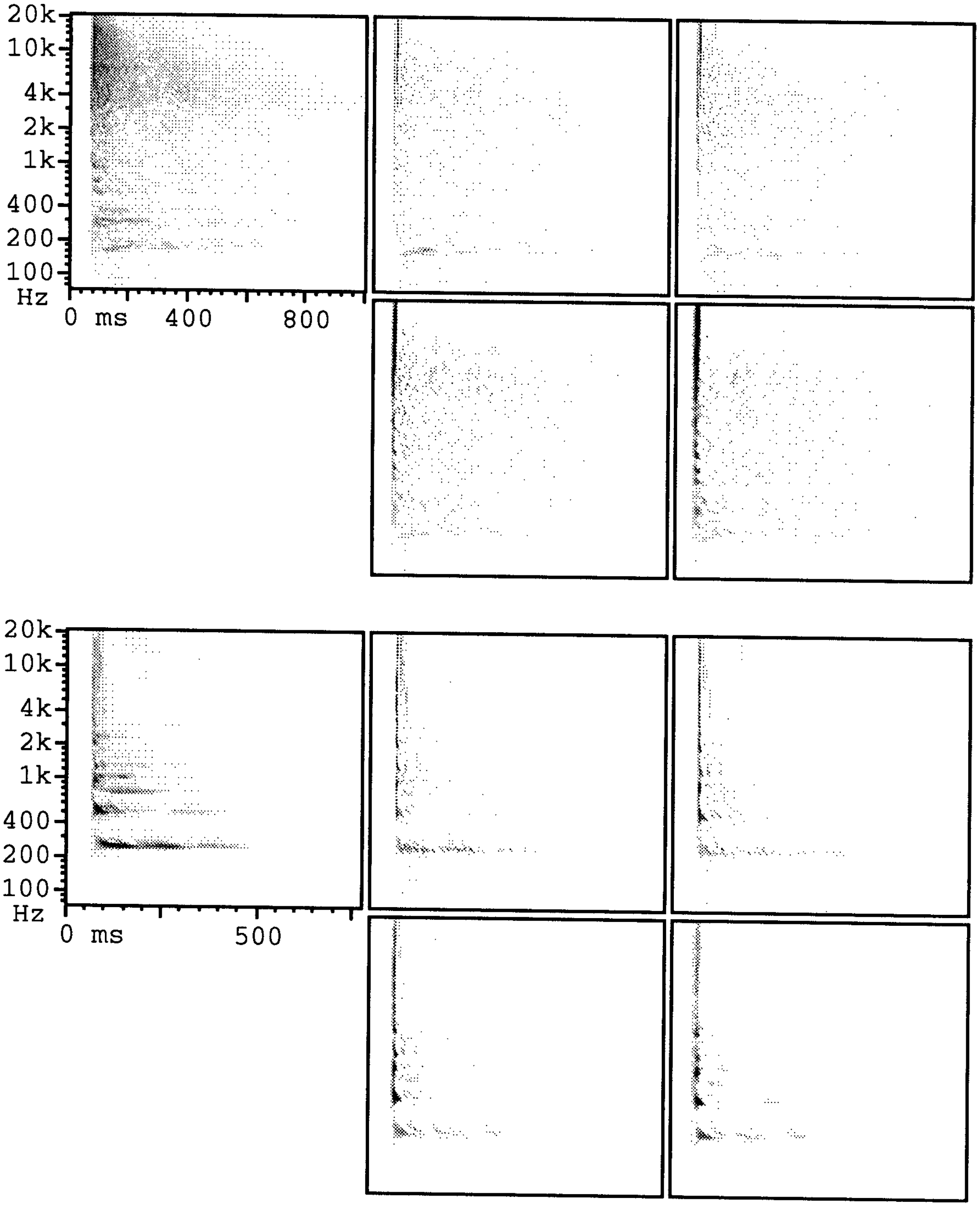


Figure 3.21: Snare drum (top) and plucked violin (bottom) onset responses.

Kernel shape:	zero-sum -/+ spike
Kernel function:	exponential decay
Characteristic decay time:	2.3 - 18.1 ms
Normalization:	for equal response

Table 3.1: Summary of onset kernel characteristics.

3.3 Filtering Frequency Variation

The second event-formation feature to be filtered in this implementation of the auditory model is frequency variation (FV); see section 2.7 for a discussion of its importance for scene analysis. FV is the movement of a partial in frequency and can be seen the cochleagram in fig. 3.22. Here, two simultaneous sounds move in pitch with different vibrato envelopes. The frequency variation can be seen as the up-and-down motion of their harmonics going across the picture in time. The lower sound, with four harmonics, has a slower and shallower vibrato than the upper sound, with six harmonics.

A source separator is ultimately interested in *common* FV, that is, variation in the frequency of several partials in the same direction at the same rate. Such partials have the same slope in the image at one instant — at one vertical slice. The goal is to find partials with such identical slopes.

One way to do this is to build a polyphonic pitch detector of some sort, then examine its output looking for changes in the pitch. This approach is complicated by two factors: First, polyphonic pitch detectors are fairly difficult to create: Most of the pitch algorithms mentioned above have been monophonic, or at least not explicitly polyphonic. Second, no one knows how pitch is extracted in the human auditory system. As far as I have been able to determine, no neurophysiological probe has yet encountered a map with anything like perceptual pitch as one of its dimensions. In contrast, neurons encoding FV have been found by the several researchers mentioned in chapter 2. For these reasons, I am avoiding a pitch detector in favor of more direct methods.

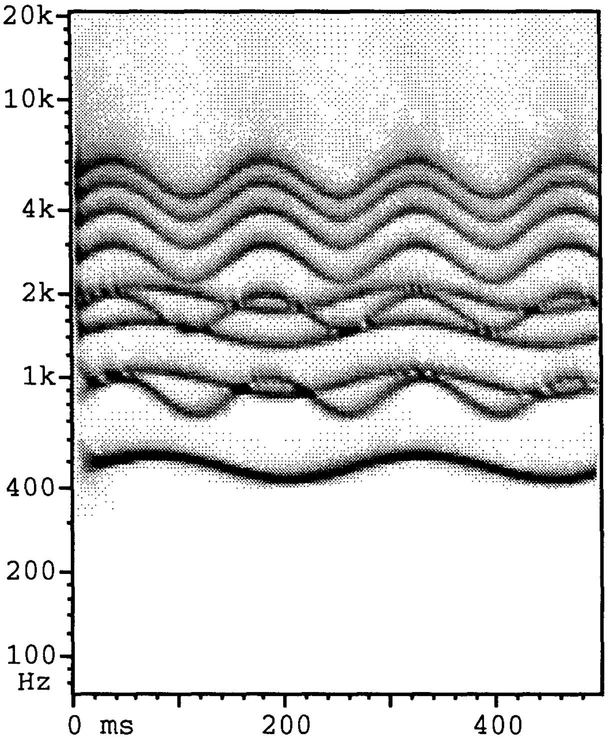


Figure 3.22: Two tones with different vibrato.

3.3.1 Frequency Variation in the Cochleagram

The method to be described here to filter a certain slope of FV operates on the cochleagram, and is similar to some methods that have been used for filtering in computational machine vision [Adelson86]. Like the onset filter of the last section, it uses a cross-correlator, though this time it is a two-dimensional one that operates on areas of the cochleagram rather than just horizontal frequency channels. The equation for the two-dimensional cross-correlation of a function c is just a two-dimensional version of the one before:

$$C(x,y) = \int\int_{-\infty}^{\infty} k(x_1,y_1)c(x+x_1,y+y_1)\ dx_1\ dy_1$$

where

$C(x,y)$	is the cross-correlation result,
$c(x,y)$	is a two-dimensional function, and
$k(x,y)$	is a two-dimensional kernel.

The kernel of this cross-correlation operator is chosen to filter FV at a certain rate — for example, at two octaves/s. The kernel is sheared from the horizontal to an angle corresponding to its specific rate of FV. The kernel in fig. 3.23, when cross correlated with the cochleagram of fig. 3.22, responds most strongly when it is centered on a partial that is varying in frequency at the kernel's characteristic rate, as is seen in the result map of fig. 3.24. Several factors combine to make this happen.

3.3.2 Logarithmic Scale

One factor is the use of height, or logarithm of frequency, as the scale in the cochleagram. This ensures that partials that are moving up or down in frequency at a certain rate *measured on a logarithmic scale* such as octaves/s, semitones/s, or cents/s, will have the same slope. Measuring frequency variation on such a logarithmic scale rather than a linear scale is appropriate because partials, when varying in frequency, maintain fixed ratios rather than fixed differences. For example, if the fundamental of a

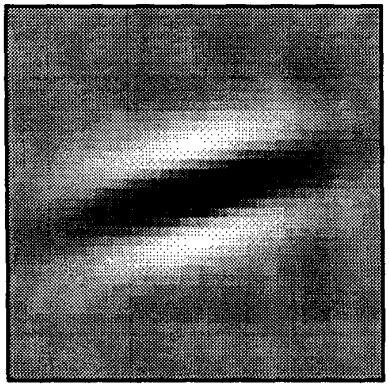


Figure 3.23: A correlation kernel for FV filtering. Time is the horizontal axis, height (log f) the vertical.

note goes from 400 to 420 Hz, then the third harmonic will go from 1200 to 1260 Hz rather than to 1220 Hz. Since the ratio of partial frequencies stays fixed, the difference of the heights stays fixed, and the partials will move with the same slope in the height scale. Note that the scale used for the frequency axis here is strictly logarithmic, rather than the scale of equal spacing in the cochlea that is skewed towards linear spacing at lower frequencies. This apparent slight variation of physiological compatibility can be excused by noticing that the computation being performed here may well exist in the auditory system. It is just the placement in space of the frequency channels that may be wrong in this model. That is, a process computing one point of a cross-correlation on this strictly logarithmic scale may use as input the feature-map elements over two semitones of range, whether at low frequencies or high. Neurons computing the same thing in the auditory cortex may receive the same inputs, from a two-semitone range, and perform the same calculation. The only difference is that a low-frequency neural unit may have a basilar membrane receptive field with a wider spatial extent than a high-frequency one, due to the increased linearity of frequency in the cochlear scale.

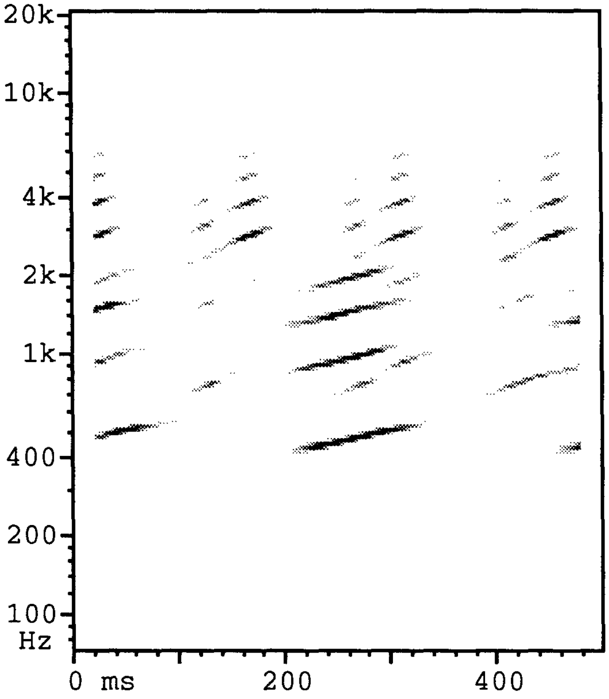


Figure 3.24: A FV feature map, computed with the kernel of the previous figure, showing partials rising in frequency.

Using a log-frequency scale, the cross-correlation equation becomes

$$F(h,t) = \int\int_{-\infty}^{\infty} k(h_1,t_1)c(h+h_1,t+t_1)\ dh_1\ dt_1,$$

where

h	is height, or $\log f$,
t	is time,
$k(h,t)$	is the two-dimensional kernel,
$c(h,t)$	is the cochleagram, and
$F(h,t)$	is the resulting frequency-variation feature map.

3.3.3 Kernel Shape

Another factor that makes this cross-correlation operator work is that the kernel is the Cartesian product of two one-dimensional kernels. In the time (horizontal) dimension, it is Gaussian in order to localize its effect to a short time period. This makes the correlation operator respond to short-time events in the cochleagram — so it can pick out instances of frequency change on this time scale.

Center-Surround Effects

In the height dimension, the kernel is center-surround. A center-surround function looks something like fig. 3.25, which has the form

$$(\sigma_1^2 - x^2)e^{\frac{-x^2}{2\sigma_2^2}}.$$

The reason for this function is to make the kernel respond to FV only at its characteristic rate. A function $f(x)$, when cross-correlated with a center-surround function, will result in positive values only where f itself has a locally high region. Flat parts of f will multiply with both the positive and negative parts of the center-surround function and lead to a zero result. Any partial varying in frequency at the kernel's characteristic rate aligns with the central axis of the kernel, intersecting

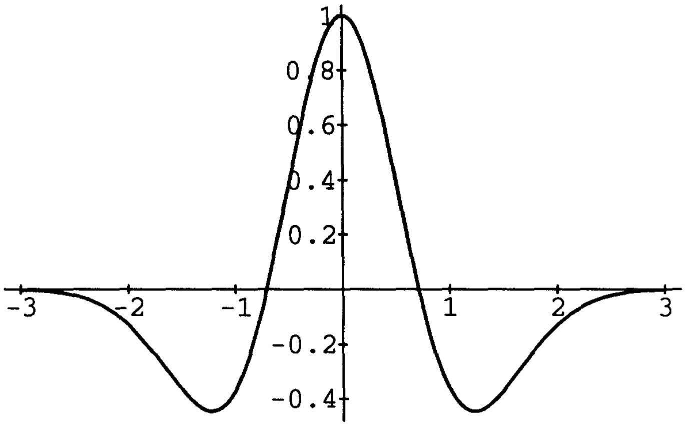


Figure 3.25: A typical center-surround function.

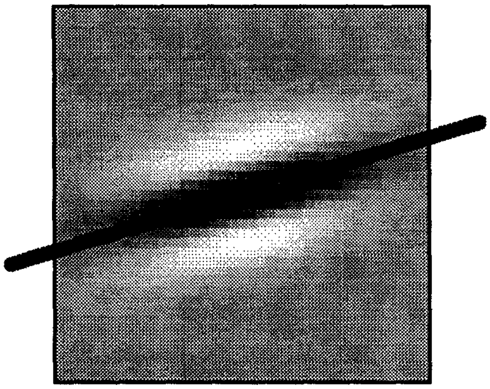


Figure 3.26: A partial aligned with the kernel axis.

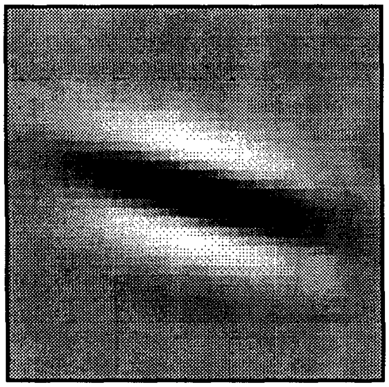


Figure 3.27: A kernel for filtering descending-frequency partials.

only its positive parts. This is shown schematically in fig. 3.26, which represents a partial by a dark grey line and the kernel with a rectangular area of black, grey, and white for positive (excitatory), negative (inhibitory), and zero areas respectively. The positive values of the kernel, when multiplied by the positive values of the partial in the feature map, produce a positive correlation value. A partial moving moving downward in frequency would be trigger a strong response by the kernel in fig. 3.27.

Crossing Partials

Other cases result in zero or negative responses, as shown in figs. 3.28, 3.29, and 3.30. In the first kernel, the partial intersects both positive (black) and negative (white) regions of the kernel, leading to a zero or very small result. In the next kernel, the partial intersects only negative points, producing a negative output. And in the last kernel, the partial intersects only zero- or very-small-valued points, producing negligible response. Since correlation is a linear operation, the kernel also filters FV at its characteristic rate with crossing partials, as depicted in fig. 3.31. The downward-moving partial contributes a zero response, while the upward one has a positive result, giving a positive overall sum.

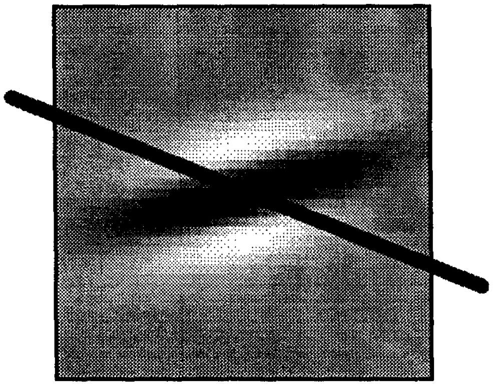


Figure 3.28: A partial crossing a kernel.

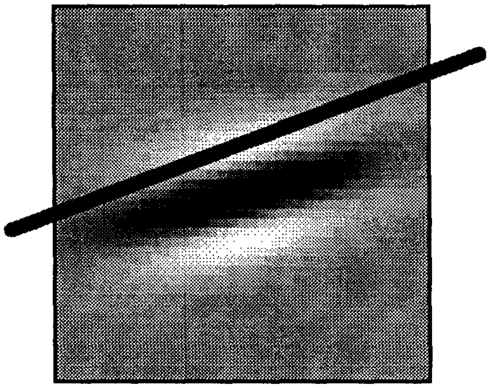


Figure 3.29: A partial aligned with a kernel but off-axis.

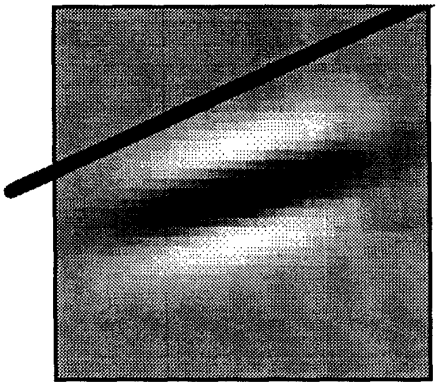


Figure 3.30: A partial aligned with a kernel far off the axis.

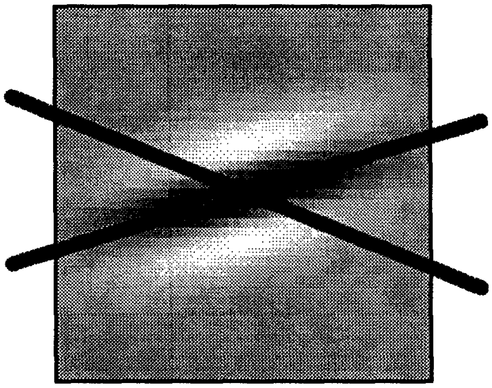


Figure 3.31: Two partials crossing a kernels at different angles.

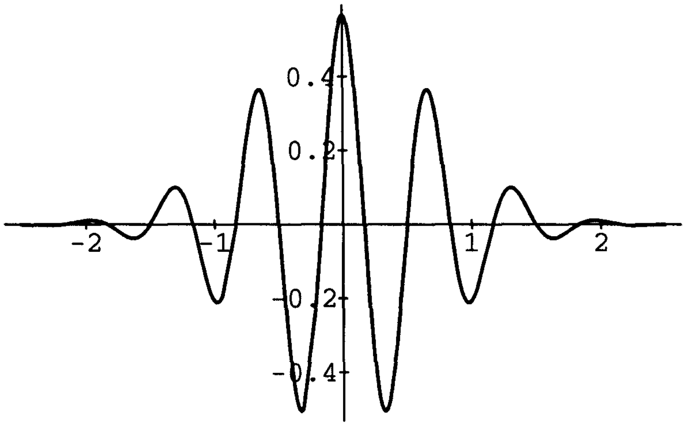


Figure 3.32: A Gabor function.

Center-Surround Functions

Also important is the choice of the center-surround function that is multiplied by the Gaussian function to make the kernel. The usual choice in machine vision work [Adelson86, Heeger91] has been a Gabor function (fig. 3.32), which is a sinusoid windowed by a Gaussian for localizing its effect:

$$g(x) = \frac{1}{\pi} cos(x) e^{\frac{-x^2}{2\sigma^2}}$$

A Gabor function has several excitatory and inhibitory regions — actually, an infinite number of them, though their values become so small far from zero as to be negligible. A Gabor function is best suited, in vision, for filtering out gratings — parallel stripes of light and dark alternating at the period of the Gabor function. A corresponding pattern in a cochleagram would be several equally-spaced partials — equally spaced on a height scale. This pattern is quite unusual, as the normal equal spacing of harmonics on a linear frequency scale translates to unequal logarithmic (height) spacing. The feature to be filtered here is a single partial with a specific FV rate. For this purpose, it seems that a center-surround like the one in fig. 3.25 with a single excitatory lobe and surrounding inhibitory lobes works best. (A Gabor function with the Gaussian spread selected to reduce the excitatory side-lobes to small values would do almost

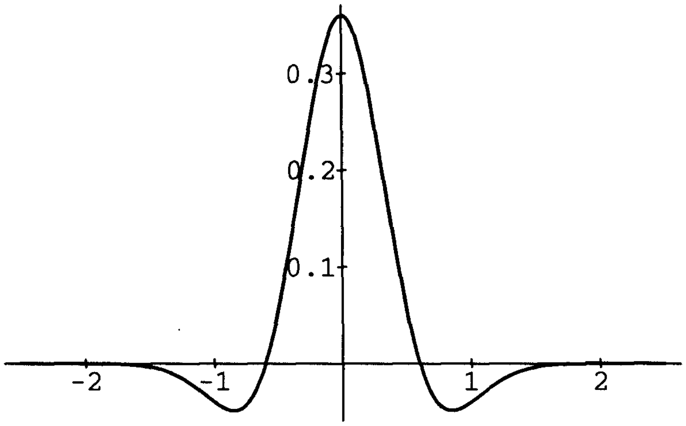


Figure 3.33: Positive-sum center-surround function.

the same thing.)

Zero-sum

Another requirement of the center-surround function f is that it sum to zero or a negative number: $\int_{-\infty}^{\infty} f(x)\ dx = 0$. If it sums to a positive number, like the function in fig. 3.33, then a crossing partial will elicit a positive response, causing smearing in the output image. Fig. 3.34, made with a kernel constructed with the function in fig. 3.33, illustrates the difference in output. Compared to fig. 3.24 above, made with a zero-sum kernel, the regions of positive response smear out much more and are noticeably less precise in localizing the places where FV happens.

3.3.4 Tuning the FV Kernel

As with the onset correlation kernel, there are several parameters that influence the operation of the FV kernel.

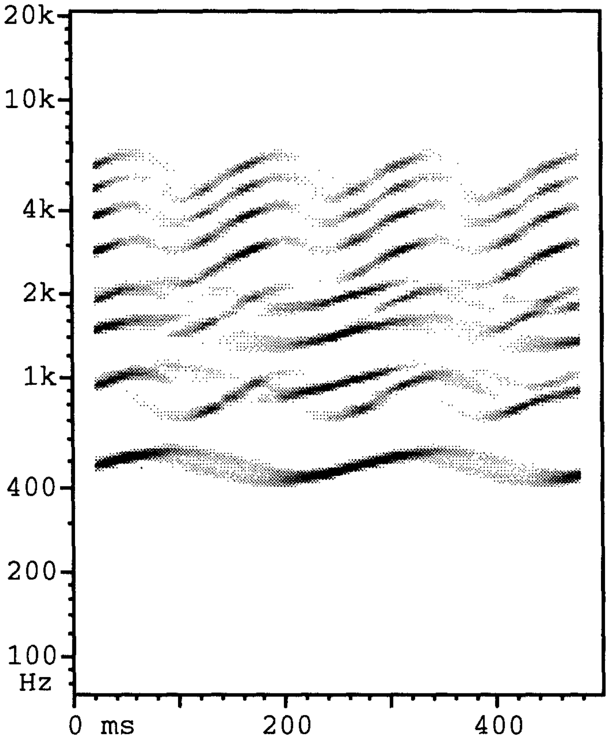


Figure 3.34: FV filtering with the kernel above.

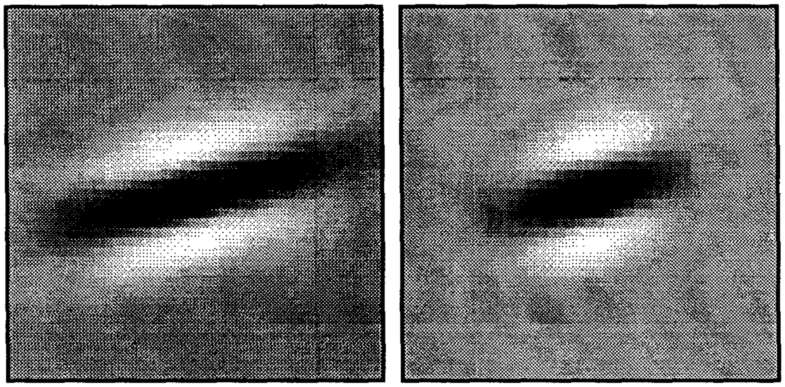


Figure 3.35: FV kernels of relatively long and short durations.

Time Spread

One of the most important of these is the extent of the kernel in time and height. There are two kernel parameters that govern the extent, one for each dimension.

The spread parameter in the time direction determines how wide the Gaussian function is for the kernel. Kernels created using two different spread factors are shown in fig. 3.35. The tradeoff between short and long kernels mirrors the familiar time-frequency tradeoff [Oppenheim75, Rabiner75] of the Discrete Fourier Transform. A short kernel uses information from a relatively small duration of the signal. This means that it can respond more quickly, filtering out relatively ephemeral frequency sweeps in the sound. A long kernel, while responding more slowly, is more accurate in filtering only FV at its characteristic rate. Imagine a partial moving through the centers of the kernels in fig. 3.35, at a FV rate aligned with the axis of the kernel. The one going through the shorter kernel has to change slope farther before it begins to intersect a significant part of the negative region of the kernel, compared to the one through the longer kernel. The short kernel is less rate-specific.

Another consideration enters the short/long tradeoff: response time. The longer a kernel is, the more time it takes to give its response after the sound signal has entered the auditory pathway. This computational model is being run on a non-real-time sound-processing workbench (see Appendix C) [Mellinger91], so response time has no effect there. It does matter, however, in the human auditory system, which must listen and respond to the real world. Shorter response times are preferred. Whitfield and Evans [Whitfield65] report, in measuring neural responses to frequency-varying tones in the cat primary auditory cortex, that "as the modulation rate increased, the phase of firing of the unit appeared to lag progressively with respect to the modulation waveform." This suggests a delay time that is approximately constant, as would be produced by an FV filter with a certain size of kernel. (The delays in question, about 20 ms, are generally much longer than the latency of response to tone onset.) Whitfield and Evans also report a falloff in unit response at a sinusoidal FM rate of 15 Hz. This rate matches the change in a perception of beating to one of roughness, perhaps because following the FM is not possible at higher rates. Since the sinusoidal modulator descends (or ascends) for only half of its period, this suggests a duration of $\frac{1}{2} \cdot \frac{1}{15}$ s, or 33 ms, for the kernel duration.

Many Time Spreads

It is entirely possible, even likely, that the auditory system uses several different lengths of time over which it collects information for FV filtering. This variation may be a separate dimension of some neural map, or it may vary consistently with some other parameter such as frequency of FV rate. There is no evidence one way or another yet. Figs. 3.36 and 3.37 show examples of maps computed with increasingly long kernels. Several lengths seem to work reasonably well; the length chosen here is 43 ms (seen in fig. 3.24), which is near to the 33 ms auditory-neuron response time mentioned above.

It should also be noted that choice of kernel duration affects the spacing of kernels across the range of FV rates, to be discussed below.

Frequency Spread

Another parameter determining a kernel's effect is the spread of its center-surround component in the frequency direction. This parameter should be chosen so that the positive region — the dark axis in the images — covers the width of a partial in the

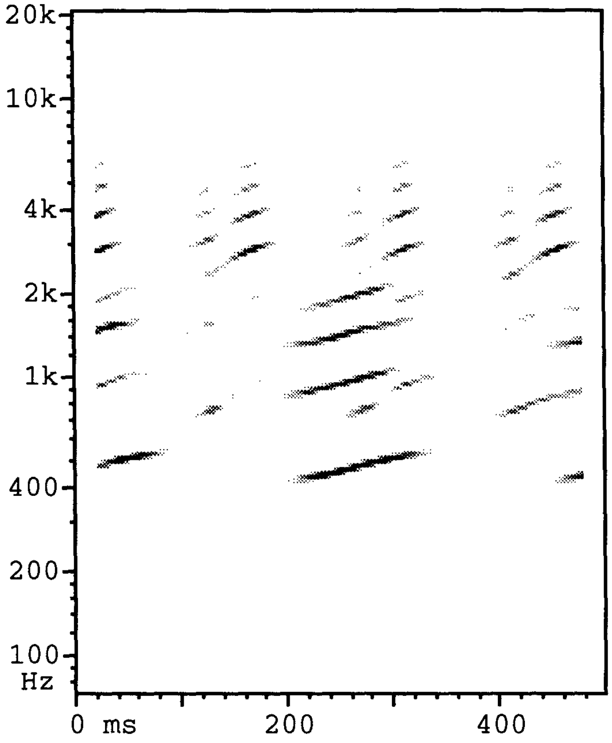


Figure 3.36: FV map computed from 23 ms kernel.

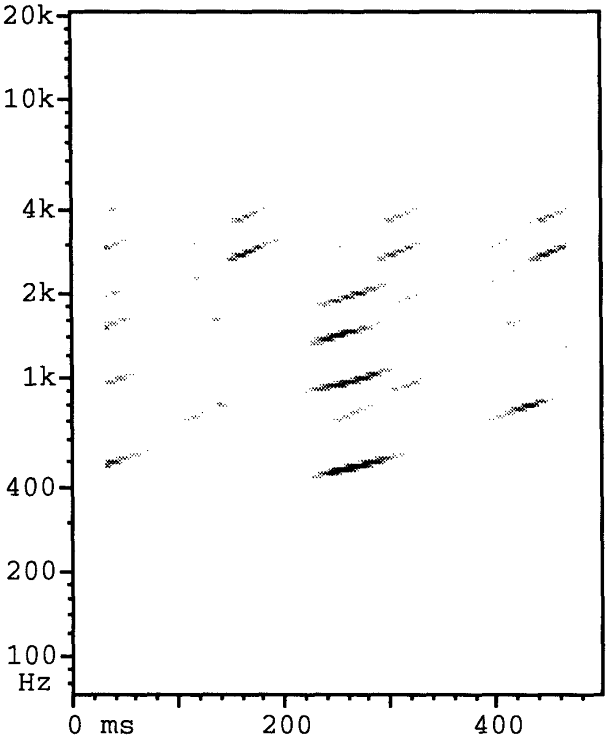


Figure 3.37: FV map computed from 57 ms kernel.

cochleagram. This makes the kernel respond best when it is aligned with the FV of the partial, and makes it stop responding most quickly when the FV rate of the partial no longer coincides with the kernel's characteristic rate. Pure-tone (sinusoidal) components in the cochleagram produced by Lyon's ear model have a $1/e$ size of about 4 cochlear channels, due to the spread of activation on the basilar membrane. This is equivalent to 2.2 semitones. Accordingly, the center-surround spread is chosen so that the positive part of the function is about 2 semitones high. This makes it work well for musical sounds, which have relatively narrow-band partials, though a different spread value or range of different spread values might be needed for speech which has wider-band partials.

Range of Rates

The most visible tuning parameter is the slope of the kernel's axis, which determines the rate of FV it filters, or, equivalently, the maximum point of its spatial frequency in time-height space. In order to do grouping by common FV, the filter must find FV at any of the rates that might occur in sound. FV can happen in smooth steps, as in glissandi and notes with vibrato, or in discrete steps, as in discrete scale runs steps and trills. The highest rate of FV found in examining a number of pieces of music has been about 4-5 octaves/s. This occurs in rapid whole-note trills and in deep, fast vibrato of the sort seen in fig. 3.24. (The definition of FV rate in the case of a trill is unclear, since discrete steps are involved. The measure used here is just the change in frequency divided by the rate of presentation of notes.) Accordingly, the limits of the range of FV filtering have been set at ±5 octaves/s. The work of Whitfield and Evans [Whitfield65] might have suggested an upper limit to FV rate in the auditory system, but unfortunately their system was limited to ±5% frequency change. With such a small excursion, FM became a change in timbre at rates fast enough to test maximum neural FV rate.

Spacing and Number of Rates

Another question is how many different rates are needed within this range, and what their spacing should be. Since the FV filter works at different characteristic FV

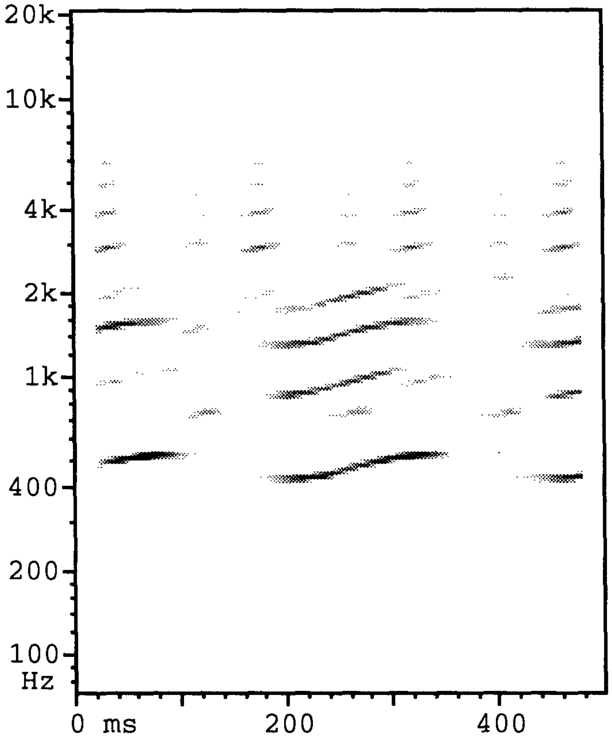


Figure 3.38: FV filtered at a rate of 2.5 octaves/s.

rates, the resulting feature map will be $2\frac{1}{2}$-dimensional, with one axis for time, one for height, and one half-dimension for FV rate. Using the tunings of parameters given above, the outputs of filters at successive FV rates begin to overlap when the rates are spread about 2 octaves apart, and overlap somewhat more when the rates are 1.25 octaves/s apart. For example, compare the two images in figs. 3.38 and 3.39. The responsive dark regions of the two images overlap at their ends, when the partials are making a transition from one rate to the other. Some amount of overlap is desirable, as it ensures that all rates of FV will have a response in one of the filter outputs, and that the sum of the outputs of all of the different FV-rate filters is roughly constant for each partial. Accordingly, FV rates spread at 1.25 octaves/s apart are used. So the kernels are at the rates 5, 3.75, 2.25, ..., -3.75, and -5 octaves/s.

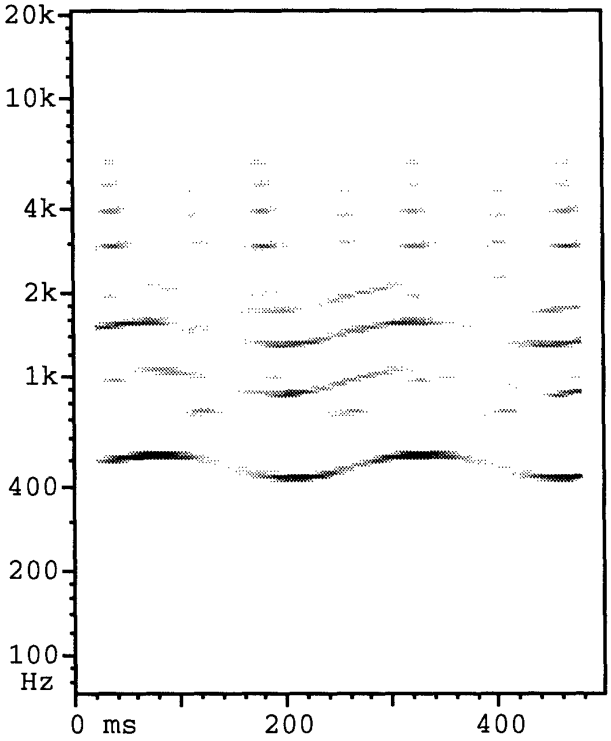


Figure 3.39: FV filtered at a rate of 1.25 octaves/s.

The best spacing of FV rates depends on the size and especially the time-spread of the kernels. As mentioned above, a longer-lasting kernel is more specific about the FV rate it filters than a shorter-lasting one. If the rates were more highly specific, each one would cover less of the desired range, so there would need to be more, more-closely-spaced kernels than described in the last paragraph. Conversely, shorter, faster-responding kernels, being less rate-specific, could be spaced farther apart, with fewer of them altogether.

FV rate:	-5 to +5 octaves/s
Spacing:	1.25 octaves/s
Frequency shape:	Zero-sum center-surround
Time shape:	Gaussian
Frequency spread:	2.2 semitones
Time spread:	43 ms characteristic decay

Table 3.2: Summary of FV kernel characteristics.

Summary of Kernel Characteristics

The kernel type and parameter values used in the computational model are summarized in table 3.2.

3.3.5 Evaluation

How well does this cross-correlation FV filter work? Does it filter FV sufficiently well to be of use to a scene analysis mechanism? What are its strong points and weak points?

An illustration of its strengths may be seen in figs. 3.40 and 3.41-3.43, which show the action of the technique on the sample input of two synthesized pitched notes: first the input sound, then the result of computing with the cross-correlation FV filter tuned to various rates. (This vibrato is relatively deep, deeper than is used commonly in music.) The FV filter works fairly well — the output maps show the various places where partials are moving up and down at specific rates.

However, the kernels do not do as well on the sound shown in fig. 3.44. This is a short segment out of a sound created by Stephen McAdams for Roger Reynolds' computer-music piece Archipelago [Reynolds83]. The sound starts out with a set of partials that make it sound like an oboe playing with slight vibrato. This perception doesn't last, however, as the even and odd harmonics of the note have independent, unequal vibrato, imperceptible at first but growing in depth until the sound seems to split in two. The odd harmonics continue sounding like an oboe, or hollow oboe, while the evens become the voice of a soprano singing \a\ one octave higher. The

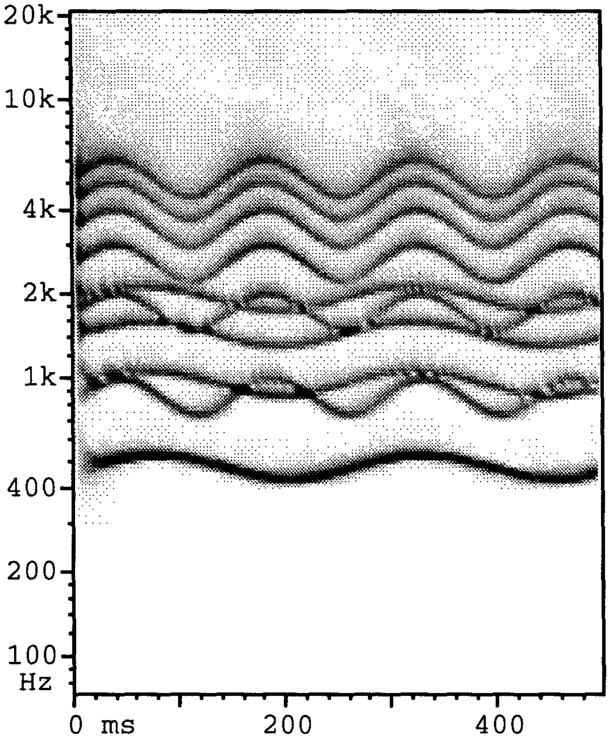


Figure 3.40: Two tones with different vibrato.

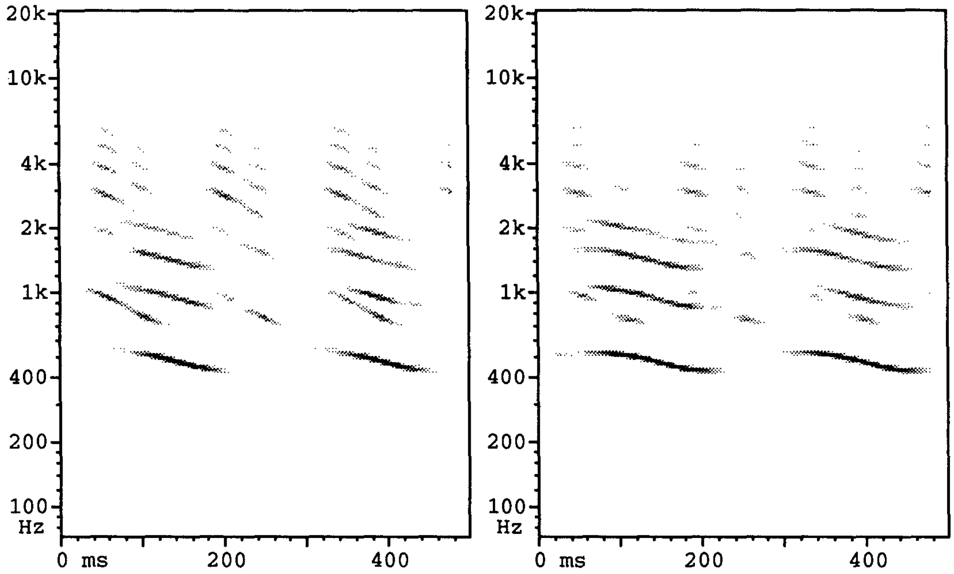


Figure 3.41: FV maps at -5 and -2.5 octaves/s.

half-second slice of sound shown here is centered at the point where the oboe appears to split into two tones.

With the McAdams/Reynolds oboe sound, the FV filter does not do very well at attaining a clear separation of the even and odd harmonics. Even using two kernels with FV rates that exactly match the two sets of partials, the results are only as good as shown in figs. 3.45 and 3.46. The differences are not much for a source separator to work with. Perhaps something could be done with the 0.5 octave/s image, but the grouping algorithm implemented in chapter 5 was not able to capture the harmonics effectively into the two natural groups. Accordingly, another technique for filtering motion information from the cochlear output was investigated. Most likely, the FV filter was simply not sensitive enough to extract the relevant information. For this reason, another method of filtering FV information from the sound signal was tried.

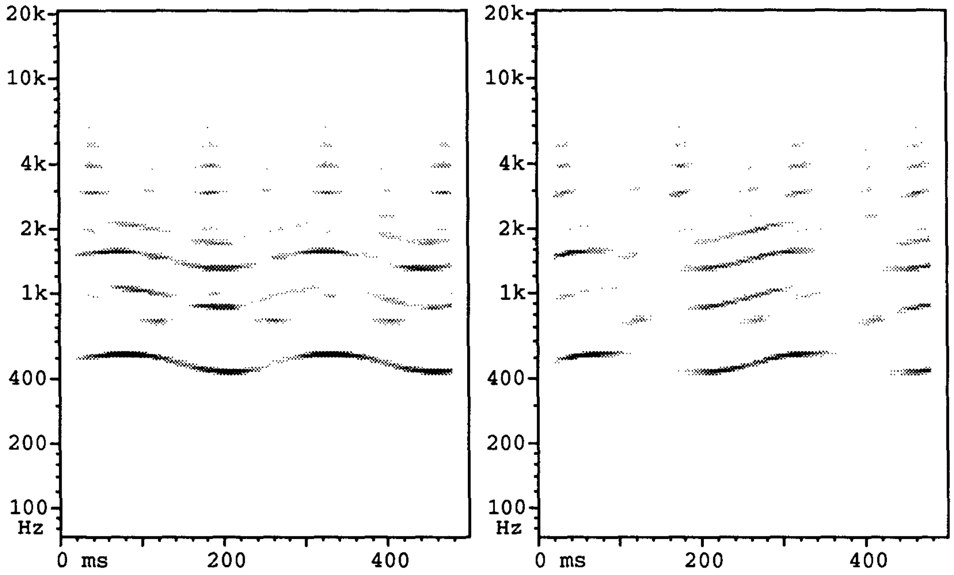


Figure 3.42: FV maps at 0 and 2.5 octaves/s.

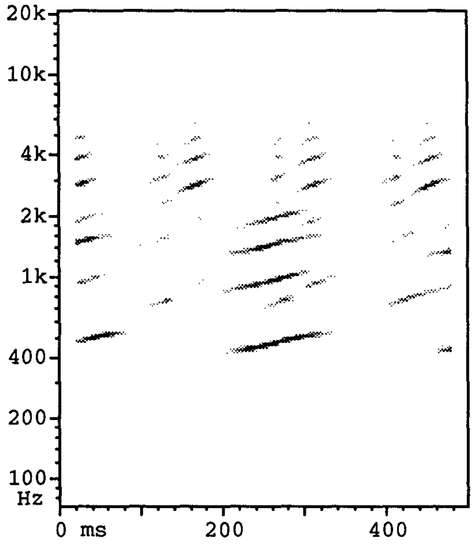


Figure 3.43: FV map at 5 octaves/s.

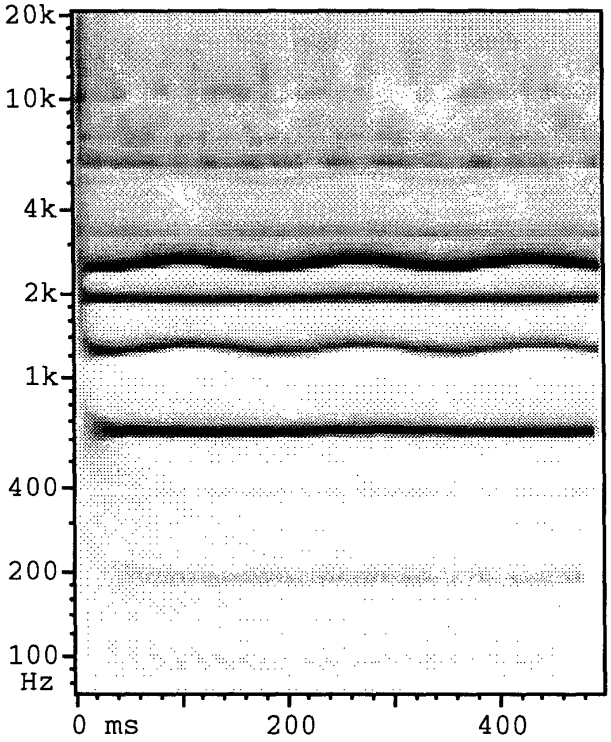


Figure 3.44: The McAdams/Reynolds oboe sound.

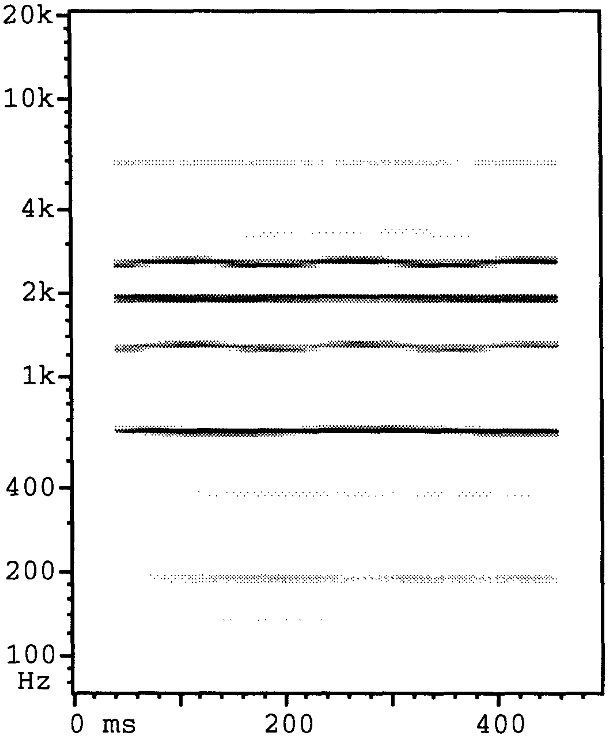


Figure 3.45: The McAdams/Reynolds oboe, FV rate 0 octaves/s.

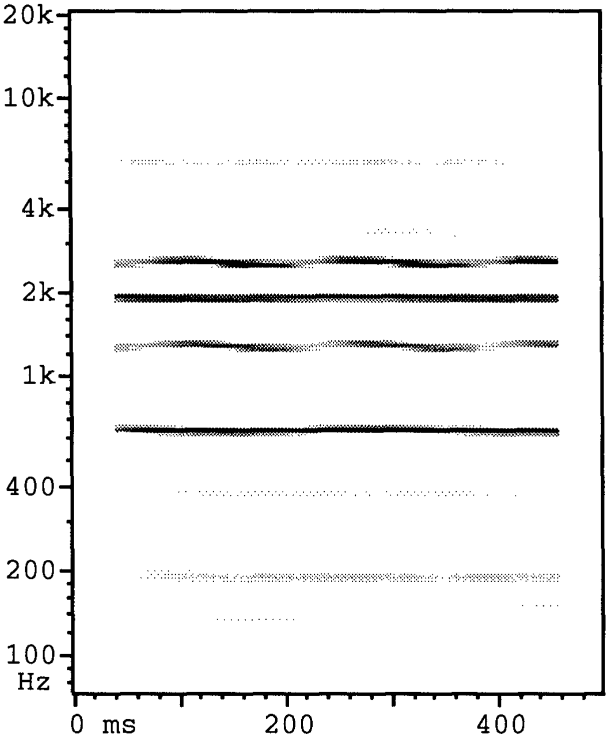


Figure 3.46: The McAdams/Reynolds oboe, FV rate 0.5 octaves/s.

3.4 Frequency Variation in the Correlogram

A second method for filtering FV in a sound works on the three-dimensional correlation feature map computed by the Lyon/Slaney ear-model program. It is based on the fact that small changes in frequency show up as relatively large changes in the correlogram image.

In the correlogram, a change in the frequency of a partial — say, upward — has two effects. First, the intensity spots corresponding to that partial move upward in the image, since they are increasing in frequency. The amount that they move up is the same as in the cochleagram, so there would be no improvement in sensitivity over the method of section 3.3.1 if that were the only effect. The spots also group more closely to the left, since an increase in frequency corresponds to a decrease in period, and this decrease is represented by a correspondingly smaller lag value in the correlogram. The space between spots on any horizontal line (any frequency channel) in a correlogram frame is just the period.

This effect can be seen in the images of the two tones in fig. 3.47. The pitch difference between the two channels is fairly small, as seen by the small difference in height between the bottom row of spots in the two images. However, the position of the spots in the lag dimension (horizontally) varies quite a bit more. In examining, for example, the fourth spot from the left in the bottom row of each image, one can see that the spots differ in position quite a bit. Even a small change in frequency, producing much smaller vertical motion of the spots than can be seen easily, causes a noticeable horizontal motion that is quite noticeable visually. Section 3.5 compares the sensitivity of cochleagram and correlogram motion filters.

Pitch Lines in the Correlogram

This correspondence between frequency and lag change means that the spots in an autocorrelogram are constrained to stay on a series of curves, with each curve corresponding to one period of the waveform. The leftmost line is where the autocorrelation function is matching successive peaks in the waveform within each channel, the next line to the right is where it matches peaks two periods apart in the waveform,

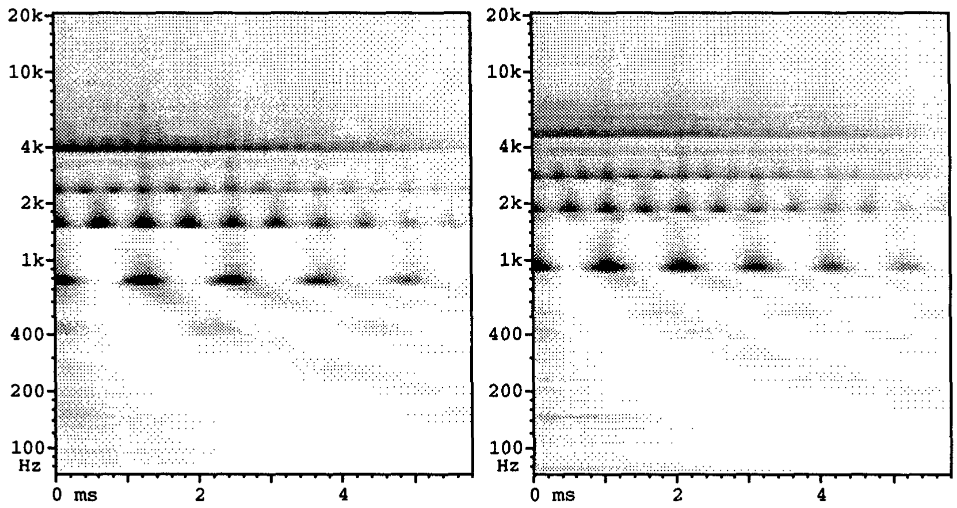


Figure 3.47: Correlogram frames for lower (left) and higher pitched tones.

then peaks three periods apart, and so on. To filter FV in the autocorrelogram, we need to respond to motion of the spots along these lines. The advantage of doing such filtering in autocorrelation space is that even small changes in the frequency can produce large changes in the period. At, for example, the third line from the left in the image, a certain change in frequency produces three times the change in period, making it three times as easy to filter out the motion. (We haven't gotten rid of the time/frequency tradeoff, of course: If the frequency changes significantly within those three periods of the wave, the spot smears out and determination of its motion becomes harder.)

A Correlation-Like Operator

The method used to filter motion of partials in the correlogram is based on a technique closely akin to two-dimensional cross-correlation. Again, it involves pointwise multiplication of a two-dimensional kernel by the two-dimensional signal in question. The only difference is that the kernel is not the same for each position in the signal

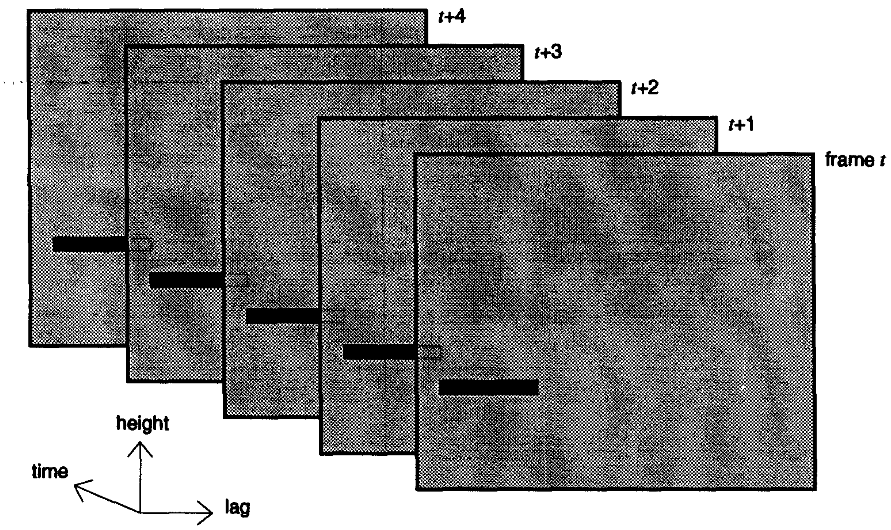


Figure 3.48: Two-D slice through the 3-D correlogram.

space. Instead, it varies from place to place in the map, its exact formulation depending on the position in the correlogram at which it is applied. Why this is necessary will be explained shortly.

The pointwise multiplication is applied to a two-dimensional slice through the three-dimensional correlogram. The aim is to take correlogram slices which contain the movement of a spot and operate on these slices with an appropriate kernel for filtering out the movement. The slice is nearly parallel to the lag and time axes of the feature map. Fig. 3.48 shows one such slice, which will be pointwise multiplied by a kernel, in schematic form. Each correlogram frame — that is, each time slice — is represented by a large grey rectangle, and the black area shows the slice through successive frames. Each horizontal black strip represents one cochlear channel, or more exactly, a section out of one cochlear channel. (It is only a section of a channel because the kernels have finite extent in the lag dimension — kernel values far away from the center of the kernel are close enough to zero that they are not shown.) The aggregation of these black strips over many frames of the correlogram is a two-dimensional area. The kernel is placed over this area and corresponding points are

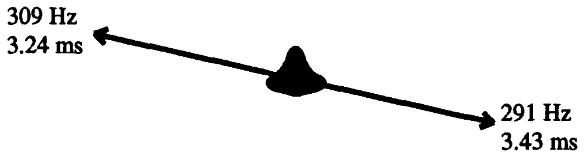


Figure 3.49: Spot motion in a correlogram.

multiplied and summed to get one point in the output feature map. The equation for this operation can be written as

$$V(h,l,t) = \int \int_{-\infty}^{\infty} k_{h,l}(l_1,t_1)C(h,l+l_1,t+t_1)\; dt_1\; dl_1$$

where

h	is height ($\log f$),
t	is time,
l	is lag,
$k_{h,l}(l,t)$	is the two-dimensional kernel,
$C(h,l,t)$	is the correlogram, and
$V(h,l,t)$	is the resulting frequency-variation feature map.

As mentioned above, the limits of integration in any implementation are not $\pm\infty$, since the kernel has significantly large values only within a finite distance of its center.

Vertical Motion

The kernel is designed to respond to motion of spots in the correlogram. The spots move primarily back and forth horizontally for slight changes in the frequency of a partial. However, there is also an up-and-down component to a spot's motion caused by change in frequency. For instance, think of a spot for a 550 Hz partial which has a frequency modulation depth of half a semitone. As shown schematically in fig. 3.49, the frequency will change between 534 and 566 Hz, corresponding to lags at the third period of 5.61 ms and 5.30 ms, respectively. So while the spot is moving horizontally between the 5.61 ms and 5.30 ms lag points, it will also have a vertical motion between channels at 534 and 566 Hz. The standard correlogram tuning makes this a motion of about 1.8 frequency channels vertically and about 14 lag bins horizontally.

In taking a slice of the correlogram, this additional vertical motion must be taken into account in taking a slice to be multiplied with a kernel. This is fairly easily done by adding a correction to the frequency channel at which a slice is taken, a correction that is proportional to the FV rate and to the distance in time from the center of the kernel: It depends on the FV rate because higher rates make for faster vertical motion, necessitating a larger correction factor. And it depends on the distance in time from the center of the kernel because a larger time corresponds to more frequency motion.

The equation with the correction factor $c(t)$ now becomes

$$V(h,l,t) = \int \int_{-\infty}^{\infty} k_{h,l}(l_1,t_1)C(h+c(t),l+l_1,t+t_1)\ dt_1\ dl_1$$

The dependence of k and c on the FV rate is not shown.

Kernel Shape

The above discussion is only about the nature of the correlogram slice; it still remains to be decided what kernel will be used to operate on this slice. The kernel here is very similar to the one used in the cochleagram FV filter of section 3.3.1, as many of the same considerations apply to its design.

As before, the scale in the frequency direction is logarithmic. This is not strictly necessary, as frequency channels are no longer required to be spaced at constant ratios, but it matches the output of the cochlea approximately and is much easier to work with than a varying-ratio scale. As before, the kernel is the Cartesian product of two one-dimensional functions, and as before, it is Gaussian in time in order to localize its effect. The time spread factor is also approximately the same as before, both because it roughly matches the response time of neurons and because it seems to work well.

The kernel differs along its other (non-time) dimension, however, because that dimension is no longer frequency but is now lag. This fact has two consequences.

The first is due to the fact that spots are repeated across a single frequency channel in the correlogram. For this reason, it is appropriate to use a function that has multiple excitatory regions, not just the one region used before. A Gabor function fits this requirement. Also, the width of the central positive hump of the function, and therefore of all other humps, is determined by the width of the correlogram spot for the kernel's frequency. Since each correlogram spot is half the period of the wave at that frequency, there is no tuning necessary for this parameter — it is determined solely by frequency. This is why there must be different kernels for different frequencies: The width of the hump is dependent on frequency.

Lag Decay Rate

There is still a tuning parameter for determining how quickly the Gabor function should decay away from its center. This is chosen so as to make nearly all of the response area of the kernel be within its central excitatory hump and the two surrounding inhibitory humps. The area outside these humps, while non-zero, has only a small effect on the kernel. The kernel, as before, is normalized to have a non-positive sum to handle mis-aligned and crossing partials.

FV Rate Tuning

A given kernel is also tuned for a specific rate of FV. A given FV rate will cause correlogram spots to move at a certain speed, and the kernel for this rate must be designed to pick out spots with just that speed. As before, the axis of the kernel is tilted by an amount that depends on FV rate. However, there there is another factor that must be considered: At a given horizontal frequency line in the correlogram, the several periods of the waveform move back and forth at different speeds. The leftmost spot (not counting the half-spot on the very edge of the frame) moves at one speed, the next one to the right at twice the speed, the next one at three times the speed, and so on. All of these different speeds are for the same FV rate. Accordingly, there must be, for a given FV rate, a different kernel for each lag (period) value. This is why the kernel function is written as $k_{h,l}(t,l)$: because the kernel design depends on both the frequency position and lag position it will be placed over in correlogram frames for the pointwise multiplication.

Range and Spacing of FV Rates

Finally, the range and spacing of FV rates must be determined. The range is as before, from -5 to +5 octaves/s. The spacing can be closer, since this filter is more sensitive to rate differences. Unfortunately, it has been somewhat difficult to experiment with the spacing, mainly because the process is quite slow and storage-intensive. I have chosen a spacing of about .5 octave/s for the filter, a value that seems to capture FV information reasonably well.

3.4.1 Cross-Correlation Output

The filtering process described above, using kernels pointwise-multiplied at locations in the correlogram, produces for each FV rate a result map of the same dimensionality (time × height × lag) and size as the correlogram. The result of running this filter on the McAdams/Reynolds oboe sound, tuned for the FV rate of -0.5 octave/s, can be seen in fig. 3.50. These images are two frames from the filter's output map which show times of maximum and minimum frequency change in the even partials, which are at frequencies of about 1300 Hz and 2600 Hz. The images show no response above about 6500 Hz because the filtering process cannot operate at higher frequencies. Correlation spots above this limit are so small that their centers are less than two array positions apart, causing an absence of separation between adjacent spots. Lack of distinct spots makes it impossible to respond to spot motion.

The images reveal some noticeable differences between the even and odd partials. In the left image of fig. 3.50, the autocorrelation spots for the even partials are darker than in the right image of the figure, showing the filtering of FV at the filter's characteristic rate. This cross-correlation technique for FV filtering works, at least to the extent of the difference in darkness of the spots.

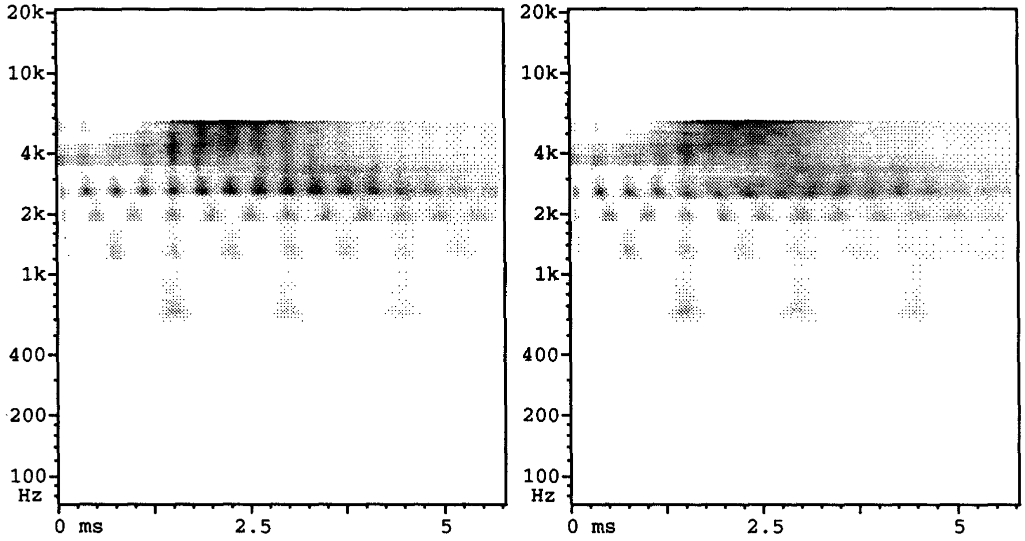


Figure 3.50: 0.5 octave/s filter output at maximum (left) and minimum response.

3.4.2 Summing Across Lags

The event formation mechanism to be detailed in chapter 5 will work best if it has two-dimensional maps with axes of time and height. This is the format of the cochleagram, of the output of the onset filter of section 3.2, and of the output of the cochleagram FV filter of section 3.3. The feature map described above is of dimension height × time × lag; this can be reduced to the desired two dimensions by summing across the lag dimension. Such summing loses information anout individual periods in the feature map, but preserves the information needed by the event formation mechanism, namely the FV data. In equation form, this summing can be expressed as

$$V_s(h,t) = \sum_{l_0 \leq l \leq l_1} V(h,l,t)$$

where

h	is height ($\log f$),
t	is time,

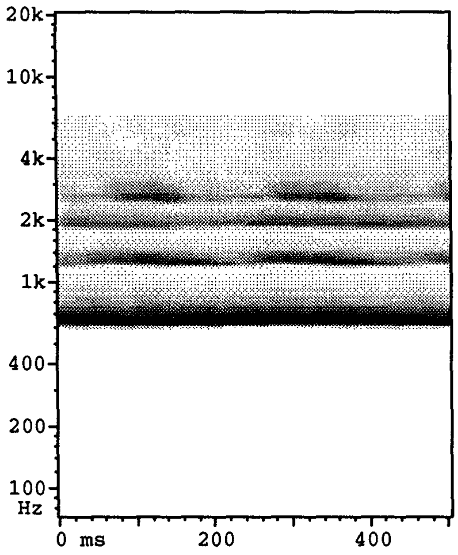


Figure 3.51: Summed filter output for -0.5 octave/s FV rate.

l	is lag,
l_0 and l_1	determine the summation range,
$V(h, l, t)$	is the FV output map described above, and
$V_s(h, t)$	is the resulting summed map.

The range of values summed, which is determined by l_0 and l_1, is chosen to extract a strong response from the filter. Such a response happens farther to the right in the image (at higher period numbers), where the spots move the most for any given FV rate. This is where the cross-correlation operator has the most information for distinguishing moving spots from unmoving ones, or from ones moving at the wrong rate. The values of l_0 and l_1 used here effect a summation in roughly the right half of the image: l_0 = 5.4 ms and l_1 = 11.6 ms.

Two feature maps produced by this summation are shown in figs. 3.51 and 3.52. These images are for two different FV rates, -0.5 and +0.5 octaves/s. Though the images are somewhat blurry, the parts of them showing FV of the even partials are

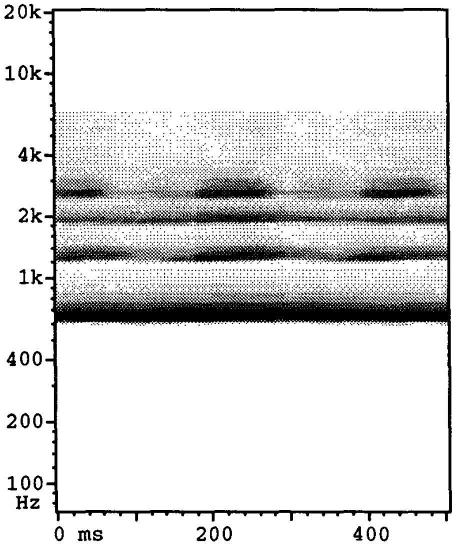


Figure 3.52: Summed filter output for +0.5 octave/s FV rate.

clear.

3.4.3 Lateral Inhibition

The difference in intensity of the FV response of figs. 3.51 and 3.52 is noticeable, but is not very large. This difference might be enough for an event-formation mechanism to work with, enabling it to separate the even and odd harmonics of this sound, but it would be better if the difference were more pronounced.

One way to to this is to enhance the contrast of the images. The problem with the above figures is that although the maximum response is noticeable, a fair amount of smearing has occurred, obscuring the desired data to some extent. The response can be brought out more clearly with some form of contrast enhancement.

Lateral inhibition is a useful and physiologically compatible technique for contrast enhancement. In this context, lateral inhibition means that the value at one T×H position of a feature map can reduce (inhibit) the values of neighboring positions.

The effect of this inhibitory contribution is to bring out any contrasts in the map.

The type of lateral inhibition performed here operates on each frequency channel of the image independently. In other words, it is a one-dimensional operation on each horizontal slice of the image. The operation is to subtract, from each point in the feature map, the minimum value of any point in a local neighborhood around the point. This is expressed in equation form as

$$V_i(h,t) = V_s(h,t) - \min_{t-\sigma \le t_1 \le t+\sigma} V_s(h,t_1)$$

where

σ is a width constant,
V_i is the laterally-inhibited result map, and
other variables are as above.

The value of σ must be large enough that a given segment of FV in the sound is not removed from the feature map. That is, if σ is too small, then the lateral inhibition will cause the center of a long-duration FV response (a long dark area in the image) to be reduced to zero. This is because all of the neighbors of a position in the middle of the response band will be approximately the same value as the given position, so the minimum value of this neighborhood is the same value, and the subtraction done in computing V_i will make the result zero. The value of σ used here is 58 ms, which was chosen simply because it seems to work well and is in the range of delay times found in the auditory system.

The results of performing this lateral inhibition on the feature maps of figs. 3.51 and 3.52 are shown in figs. 3.53 and 3.54. Though the images are still somewhat blurry, the up-and-down motion of the even partials stands out clearly — upward motion in fig. 3.54, downward in fig. 3.53. This amount of difference should be sufficient for the event-formation mechanism to separate out the even and odd partials.

Note: The lateral inhibition performed at this step cannot be performed with the multiplicative kernel used in previous steps because it depends on a minimum operator rather than just linear operators.

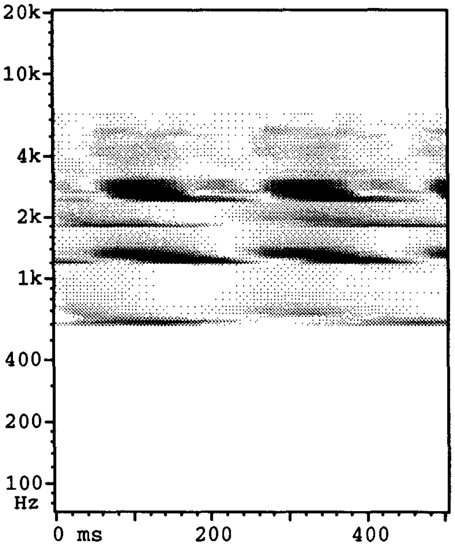


Figure 3.53: Laterally-inhibited, summed filter output for -0.5 octave/s FV rate.

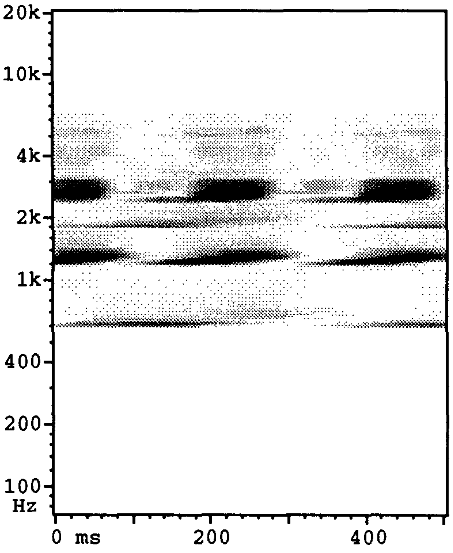


Figure 3.54: Laterally-inhibited, summed filter output for +0.5 octave/s FV rate.

3.5 Comparison of FV Filters

Two different techniques for filtering FV information from a sound signal for use in auditory scene analysis have been presented. How do they compare? That is, how sensitive are the two to different rates of FV in a signal?

A rough measure of the relative FV sensitivity of the two filtering techniques — the one that uses the cochleagram and the one that uses the correlogram — can be made by relating the sharpness of a peak in the representation to the distance the peak moves for a given amount of FV. In the cochleagram, a peak is a spectral peak; in the correlogram, a peak is a periodicity spot in a frequency channel (a peak in a horizontal slice). That is, we wish to measure how much frequency change is needed to produce a significant motion of the peak. Sharper peaks need less motion to be filtered out than broader peaks, and the desired measure should take this into account.

The method used here computes a "quality" measure for the two filtering techniques. It is based on measuring a peak in the output functions of the two filters — a spectral peak in the cochleagram and a periodicity peak in the correlogram. It computes the relative amount of frequency change needed to move a peak in the function between the two points, one on either side of the peak, at which the function has half of its peak value. This measure, called Q because of its similarity to the quality measure from filtering theory, divides the frequency of the peak by the frequency difference between the half-height points on either side of the peak.

$$Q = \frac{f}{f^{+} - f^{-}}$$

Here f^{+} is the higher frequency where the function has half its peak value, and f^{-} is the lower. Fig. 3.55 shows a peak of the function with f^{+} and f^{-} marked.

3.5.1 Cochleagram

The Q value in a cochleagram can be determined by measurement. Let w be the width of a peak, in cochleagram channels, at half the height of the peak. Since the cochleagram, as used here, is constant-Q, w is the same at all frequencies. Let c be

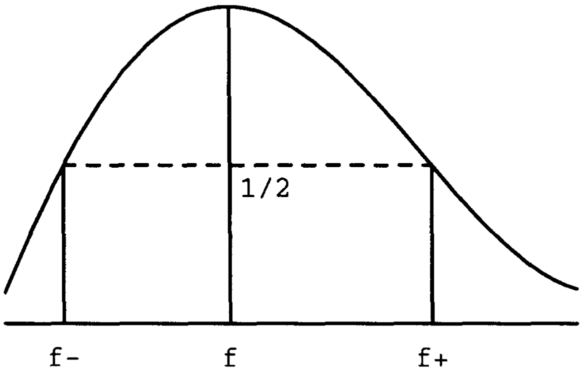


Figure 3.55: A peak in a function, with f^+ and f^- marked.

the number of frequency channels per octave. Then the ratio between f^+ and f^- is $2^{w/c}$. If a peak is at frequency f, then

$$f^- = \frac{f}{2^{\frac{w}{2c}}}$$

and

$$f^+ = f \cdot 2^{\frac{w}{2c}}.$$

So then

$$\begin{aligned} Q &= \frac{f}{f^+ - f^-} \\ &= \frac{f}{f \cdot 2^{\frac{w}{2c}} - f \cdot 2^{\frac{-w}{2c}}} \\ &= \frac{1}{2^{\frac{w}{2c}} - 2^{\frac{-w}{2c}}}. \end{aligned}$$

With the cochlear tuning used here, $c = 21.8$, and by measurement, $w \approx 2.93$. So $Q \approx 10.7$ in the cochleagram.

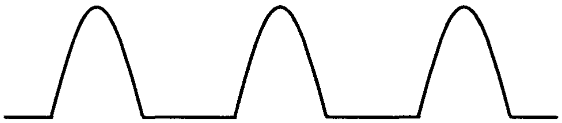


Figure 3.56: Halfwave-rectified sine wave.

3.5.2 Correlogram

The correlogram is slightly more complicated. A horizontal slice through one frame in the correlogram is the autocorrelation function of one frequency channel of the cochleagram. This function is the one whose peaks we should compare to cochleagram peaks, because horizontal slices are where the dot-product operator for FV filtering is applied. The signal in a cochleagram frequency channel is, for a pure tone, very close to a halfwave-rectified sine wave, as in fig. 3.56. To compute Q for the correlogram function, we need to find the sharpness of peaks in the autocorrelation of this function. This is done by two methods, an analytical one and an empirical one.

Analytical Method

One way to find the sharpness of peaks in the correlogram is with a mathematical analysis of correlogram output, by assuming that the function output by the cochleagram for a pure tone is exactly a halfwave-rectified sine wave.

Let $\lceil \sin x \rceil = \max(\sin x, 0)$ represent a halfwave-rectified sine function. Then the autocorrelation of this function over one period is

$$C(l) = \int_0^{2\pi} \lceil \sin x \rceil \lceil \sin(x + l) \rceil \; dx.$$

The restriction to one period of the function only makes the math easier. Since a sine wave is periodic, any number of additional periods would just multiply the result by some integer constant, not affecting the Q measurement.

The multiplication of the two functions for a fixed value of l is illustrated in fig. 3.57, with $\sin x$ as the top curve and $\sin(x + l)$ as the bottom. In the range $[0,\ 2\pi)$, both curves are non-zero only between 0 and $\pi - l$. Actually this is true

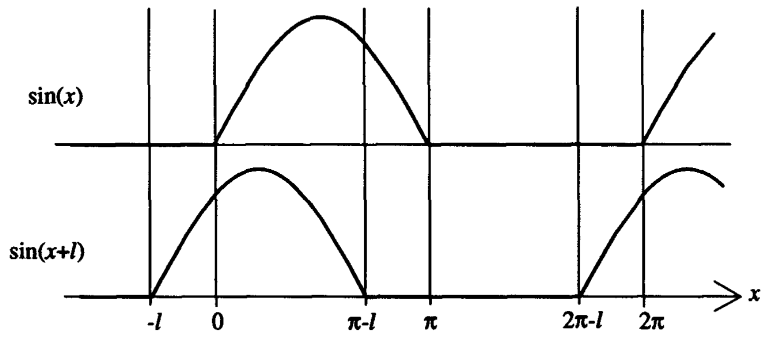


Figure 3.57: Computation of autocorrelation value for fixed l.

only for $0 \leq l \leq \pi$, or more generally for $0 \leq l \pm 2\pi m \leq \pi$ with m any integer. An analogous result holds for $\pi \leq l \pm 2\pi m \leq 2\pi$, making the function symmetric about $x = 0$.

So the integral becomes

$$
\begin{aligned}
C(l) &= \int_0^{2\pi} \lceil \sin x \rceil \lceil \sin(x+l) \rceil \; dx \\
&= \int_0^{\pi - l} \sin x \sin(x+l) \; dx \\
&= \frac{1}{2} \int_0^{\pi - l} \cos l - \cos(2x+l) \; dx \\
&= \frac{1}{2} x \cos l - \frac{1}{4} \sin(2x+l) \Big|_{x=0}^{\pi - l} \\
&= \frac{\pi - l}{2} \cos l - \frac{1}{4} \sin(2\pi - l) + \frac{1}{4} \sin l \\
&= \frac{\pi - l}{2} \cos l + \frac{1}{2} \sin l
\end{aligned}
$$

One period of $C(l)$ is plotted in fig. 3.58, along with one period of a cosine for comparison. The maxima of $C(l)$ occur at $\pm 2\pi m$, where it has the value $C(0) = \pi/2$.

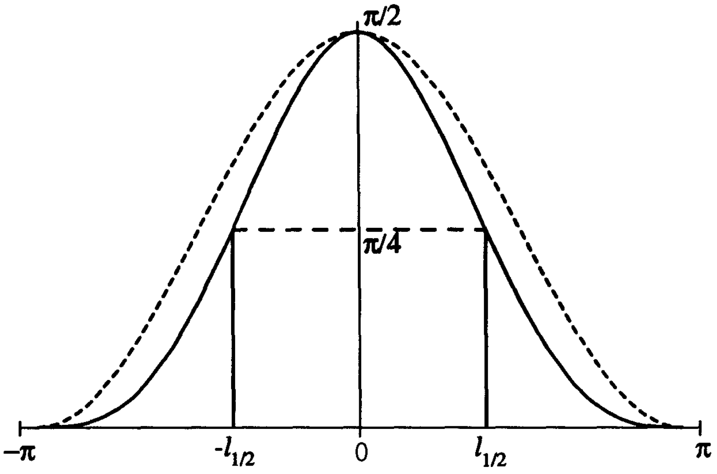


Figure 3.58: Autocorrelation function $C(l)$ (solid curve) with cosine wave (dotted curve) for comparison.

To compute Q, we need the width of the peak at half of its height. A numerical solution to $C(l) = \pi/4$ gives $l_{1/2} \approx 1.2359$. So the width of the peak at height $\pi/4$ is $2 \cdot l_{1/2} \approx 2.4718$.

The above calculation assumes a halfwave-rectified sine with a period of 2π. For one with period p, the half-height width is $2 \cdot l_{1/2} \cdot \frac{p}{2\pi} = \frac{l_{1/2}p}{\pi}$. A spectral peak at frequency f has a peak in the correlogram at the lag value of $p = 1/f$ as well as repeated peaks at lag values of $2p$, $3p$, and so on. For a given peak, let n be the "peak number" — the lag value divided by p. So the peak at p has $n = 1$, that at $2p$ has $n = 2$, etc.

To compute Q, we need expressions for f^+ and f^-. Let $p^+ = 1/f^+$ and $p^- = 1/f^-$. At the n^{th} peak,

$$
\begin{aligned}
p^+ &= np - \frac{l_{1/2} \cdot p}{2\pi} = p(n - \frac{l_{1/2}}{2\pi}) \\
\text{and } p^- &= np + \frac{l_{1/2} \cdot p}{2\pi} = p(n + \frac{l_{1/2}}{2\pi}).
\end{aligned}
$$

n	Q_n (calculated)
1	2.4436
2	5.0347
3	7.5931
4	10.1432
5	12.6901
6	15.2353

Table 3.3: Analytical Q values for the leftmost few spots in a correlogram.

Then

$$\begin{aligned} Q_n &= \frac{f}{f^+ - f^-} \\ &= \frac{\frac{1}{np}}{1/p^+ - 1/p^-} \\ &= \frac{\frac{1}{np}}{\frac{1}{p(n-\frac{l_{1/2}}{2\pi})} - \frac{1}{p(n+\frac{l_{1/2}}{2\pi})}} \\ &= \frac{\frac{1}{2\pi n}}{\frac{1}{2\pi n - l_{1/2}} - \frac{1}{2\pi n + l_{1/2}}}. \end{aligned}$$

Some numerical values of Q for different n are listed in table 3.3. Further insight can be obtained with an approximation to Q. Let $\epsilon = \frac{l_{1/2}}{2\pi n}$. Then

$$\begin{aligned} Q_n &= \frac{1}{\frac{1}{1-\epsilon} - \frac{1}{1+\epsilon}} \\ &\approx \frac{1}{(1+\epsilon) - (1-\epsilon)} \\ &= \frac{1}{2\epsilon} \\ &= \frac{\pi n}{l_{1/2}}. \end{aligned}$$

Thus Q increases approximately linearly with n.

Empirical Method

Another way to determine the sharpness of correlogram peaks is to measure them. At a sampling rate of r, frequency of f, and measured peak width (in lags) of w, the periodicity numbers are

$$\begin{aligned} p^{+} &= \frac{1}{f} - \frac{w}{2r} \\ \text{and } p^{-} &= \frac{1}{f} + \frac{w}{2r} \end{aligned}$$

So then

$$\begin{aligned} Q &= \frac{f}{\frac{1}{p^{+}} - \frac{1}{p^{-}}} \\ &= \frac{f}{\frac{1}{\frac{1}{f} - \frac{w}{2r}} - \frac{1}{\frac{1}{f} + \frac{w}{2r}}} \end{aligned}$$

At a sampling rate of $r = 44100$ Hz, a tone at $f = 500$ Hz generates spots in the correlogram that are $w \approx 25$ lag bins wide; this gives the Q_n values shown in table 3.4. These Q values are slightly better than the values calculated by the mathematical analysis. A possible reason for this discrepancy is that the function used in the analytical computation, which was assumed to be a halfwave-rectified sine, may have slightly sharper peaks in a real cochleagram than a sine function. This could be caused by the automatic gain control in Lyon's cochlear model.

The conclusion is that for $n > 3$ (by the analytical method) or $n > 4$ (by the measurement method), the quality measure Q is higher in the correlogram than in the cochleagram, making the correlogram method better able to respond to the changes in frequency to be used later for auditory scene analysis. However, the difference may not be as marked as analyzed here, since the amount of smearing of correlogram spots, hence the reduction of peak sharpness, also depends on the rate of frequency

n	Q_n (measured)
1	3.4571
2	7.0206
3	10.5604
4	14.0943
5	17.6258
6	21.1562

Table 3.4: Measured Q values for the leftmost few spots in a correlogram.

variation and on the period number n. That is, as the FV rate gets farther from 0, correlogram spots smear out more, and for a given non-zero FV rate, the spots smear out more the farther to the right they are (the higher their n value). This smearing is worse for quickly-varying frequencies, suggesting that filtering FV in the correlogram works best for lower FV rates. Perhaps the auditory system uses something like correlogram FV filtering for low rates of FV where greater resolution is needed, like frequency jitter in instrument tones, and something like cochleagram FV filtering for higher rates.

Chapter 4

Event and Source Formation

> The human auditory system has a system for segregating the "neural spectrogram" and grouping its primitive components and features into a set of separate streams, each of which is the internal representation of distinct external acoustic features.
>
> Albert S. Bregman

The previous two chapters cover the feature filtering stage of the auditory model. This chapter outlines the higher-level processes in which feature information is used to decide what events are present in the sound signal, and to group these events into sources. Most of this chapter is an overview of the factors that influence these processes, while the next one contains an implementation of an event-formation algorithm that incorporates some of these factors. This chapter is but a brief overview of the important aspects of event and source formation, as that subject is broad and deep. Much of Bregman's 700^+ page book is devoted to it, and only the surface of the subject is touched here. Nonetheless, the hope is to describe enough of these processes that the implementation in chapter 5 will make sense.

This chapter begins with a review of some of the characteristics of source perception: How sources are organized in time and in level of detail, how different perceptual organizations compete with one another, what constraints they must obey. It goes into the processes of event formation and then source formation, describing some of the factors that influence each. The chapter draws throughout on the works of Moore [Moore89] and especially Bregman [Bregman90].

4.1 Characteristics of Event and Source Formation

The terms *formation* and *separation* refer to the same processes but with different emphases. Formation, either of events or of sources, means the associating of different parts of the time-frequency image to make a single object. This is the complement of separation, which emphasizes the portion of the process that distinguish parts of the image as belonging to different sources. Neither formation nor separation can exist without the other, but formation emphasizes the unifying part of the process while separation emphasizes the discriminating part.

Simultaneous *vs.* Sequential

Another important distinction is that between *simultaneous* and *sequential* grouping, also called [Bregman90, p. 30] vertical and horizontal grouping respectively. Simultaneous grouping is the process by which we hear different parts of the spectrum, potentially widely-spaced parts, as arising from the same source. Harmonicity is a kind of simultaneous grouping, and would be covered in this section if it were not for the periodicity analysis believed to underlie it. Simultaneous grouping is associated closely with event formation: Though events last over time, and require features filtered at different times to be integrated, much of the task of event formation can be thought of as finding which of the partials that exist at any instant belong together.

Sequential integration is the part of the scene analysis process that forms groups over time. Of it, Bregman says,

> This is the kind of integration that forms the melodic component of music. It is the process that connects events that have arisen at different times from the same source. [Bregman90, p. 30]

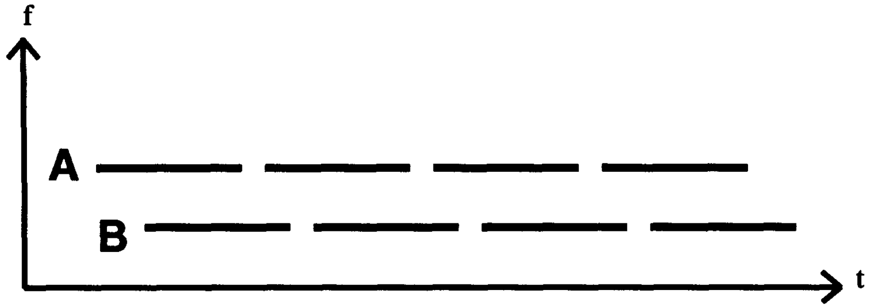


Figure 4.1: Tones in Bregman-Rudnicky experiment.

(Note: Though Bregman does not define "event," his meaning is probably close to the one used here.) In a piece of music with several parts — one or more melodies and harmonies — each part may be thought of as a separate source.

Sequential and simultaneous organization happen concurrently, influencing each other's progress. Simultaneous organization clearly happens first in most musical hearing, for we hear quite easily such simultaneous cues as common onset, common offset, and harmonically-spaced partials, even when several instrument notes are competing for attention. Yet sequential organization can affect simultaneous organization as well, as demonstrated by an experiment of Bregman and Rudnicky [Bregman75, Bregman90, pp. 213-215]. They played two simultaneous sequences of repeated sine tones, each sequence at the same repetition rate but at a different frequency. See fig. 4.1. The initial perception was of a single complex timbre made up of both tones A and B. However, the stimulus rate was fast enough that after a few repetitions of the pattern, the two sequences broke up and became separate sources. At this point, each sequence was heard as a pure tone. In other words, the perception of timbre — a simultaneous feature that includes at a minimum grouping of harmonic tones — was affected by the perception of sequential sources.

4.1.1 Hierarchical Perception

Grouping is a hierarchical process. To some extent this fact is reflected in the structure of this thesis, in which low-level features are grouped into events, and events into sources. But sources themselves have hierarchical structure as well. Think of hearing an orchestra. At the largest grouping scale, the entire orchestra is a single source. At successively finer scales of analysis, one can hear the string section as a source, the first violin section as a source, or perhaps even a solo violinist as a source. Each of these levels of grouping is contained within the previous; the organization of sources is hierarchical.

This hierarchical organization is partly directed by attention. That is, one can focus attention on a particular level of the hierarchy to make apparent the source present at that level. The levels are not always as clear-cut as in the example just given, as when a piece of music has many interweaving elements that come into and go out of existence rapidly. Indeed, probably everyone has had the experience of hearing a piece of music differently after repeated listening. Often the difference is due to perceptual re-organization of the sources, with musical events placed into different streams after the deeper analysis enabled by familiarity.

4.1.2 Competing Organizations

As the above experiment of Bregman and Rudnicky demonstrates, there may be several possible descriptions, or organizations, of a sound signal. A visual analogy is the Necker Cube illusion, in which the visual system can impose either of two organizations on a set of line segments, producing percepts with two different front corners. In either the auditory or visual case, there may be several different factors that tend to favor one organization over another. Which one wins is determined by the number and strengths of the factors for a particular sound.

Many of Bregman's experiments have exploited this competition between grouping factors. By inventing sound stimuli that can be altered so as to favor one grouping cue that leads to one perceptual organization, or another cue that leads to another organization, he has discovered a wealth of information about various types of grouping cues. For instance, the Bregman and Pinker experiment of fig. 2.6 (page 33) exploited competition between a cue for common onset that led to grouping B and C together, and a cue for nearness in frequency that led to grouping A and B together. The experimenters were further able to change the frequency of A to bring it nearer to or farther from that of B, and the frequency of C to bring it into or out of a harmonic relationship with B. Another example is the experiment of Bregman and Rudnicky seen in fig. 4.1. Here, the onset asynchrony of tones in the two sequences competes to make two separate streams against the tones' harmonicity, which tends to make a single stream.

4.1.3 Allocation and Accounting

Two complementary principles that figured in Gestalt thinking on vision are important in auditory scene analysis. The first, known to Gestaltists as "belongingness" and in auditory scene analysis research as the principle of exclusive allocation, says that "a sensory element should not be used in more than one description at a time" [Bregman90, p. 12]. In other words, if there is some element of the incoming sound — a unit of energy in a spectrogram, for instance — then this element should be assigned to only one source; it cannot do double duty. A complementary principle, known as accounting, requires that all incoming sounds be assigned to one source or another. If a sound is heard that cannot be assigned to any existing source, then it becomes a source by itself, able to have other sounds be grouped with it later. This principle could be applied additively: if one source could account for only part of the energy present at some frequency, then this part could be subtracted from the total to leave a residue that would have to be accounted for by some other source.

Allocation and accounting happen at all the levels of organization mentioned in chapter 1. All of the features responded to must be accounted for by being associated with events, and each unit of intensity of a filtered feature is generally associated with but a single source (but note the exceptions below). Furthermore, all of the events produced by an event-formation process must be accounted for as part of some source, and generally only one source.

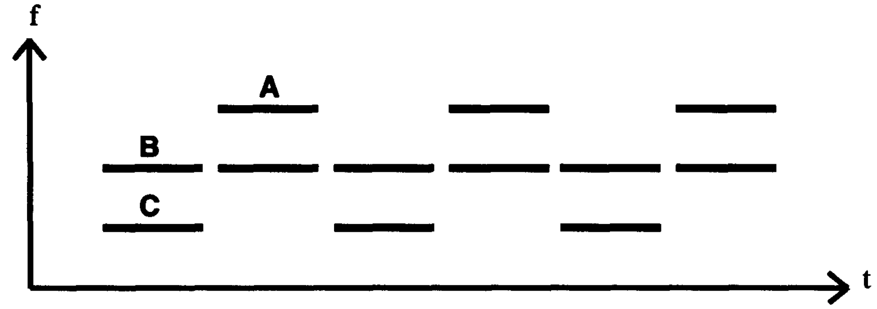


Figure 4.2: Exclusive allocation in Bregman's experiment.

Evidence and Explanation

Allocation and accounting are two aspects of another principle of perceptual organization: the interplay between evidence and explanation. A sound signal provides evidence of events in the world, evidence that is explained by the auditory system. One of the forms this explanation takes is grouping sound elements into sources; the explanatory process continues working until it has accounted for all of the sound present at the ear.

Evidence For and Against Exclusive Allocation

Evidence for the principle of exclusive allocation comes from several experiments that supplement common sense. Some of them have been seen already, as in the Bregman-Pinker experiment (fig. 2.6). There, tone B could belong to only one of the streams at a time; one cannot hear it simultaneously as part of both possible streams. A similar example is Bregman's demonstration with the sound sequence shown in fig. 4.2 [Bregman90, p. 337]. Here, the middle-frequency tone cannot contribute simultaneously to the two possible organizations, that of a complex-timbred note sounding alternately at low and high frequencies, and of a of a single pure tone B with pure tones C and A alternating below and above.

A number of cases have been found in which allocation fails, causing one stimulus to contribute to two or more events or streams. This phenomenon, known as duplex perception [Ciocca89], was first pointed out in sound by Rand [Rand74]. Gardner and Darwin [Gardner86] found it present in a study of vowel harmonics, in which they modulated the frequency of a single harmonic of a vowel sound. The modulated harmonic stood out from the rest of the complex as would be expected from its frequency-modulation cue, but the timbre of the vowel did not change as it did when the harmonic was eliminated entirely. In other words, the harmonic contributed to the perception of both the vowel and the separate source.

A similar effect may be found in McAdams's oboe sound (fig. 3.44, page 97). When the even harmonics split apart from the complex sound to become a separate soprano-like source, the remaining odd harmonics continue sounding like an oboe. The oboe sounds slightly hollow, but not nearly as much so as when the odd harmonics are played alone, when they sound much like a clarinet.

The above instances both derive their duplex effect from frequency-modulated sounds, but it apparently works for harmonicity and pitch perception as well. Moore *et al.* [Moore85a] mistuned one harmonic of a harmonic complex by 3% or more and found that it stands out from the complex as a separate entity. At the same time, it contributes to residue pitch, changing the perceived pitch slightly.

Though these and other instances show that the principle of exclusive allocation is sometimes violated, duplex perception seems to be the exception rather than the rule. It also seems to happen more with speech sounds than other types of sounds [Moore89, p. 270], which may reflect the operation of a separate speech processing module. As Moore says,

> It appears that the speech perception mechanism sometimes groups acoustic elements together even when the acoustic properties of the elements suggest that they come from different sources. [Moore89, pp. 270-271]

4.2 Natural Constraints

The principles of accounting and exclusive allocation in perceptual systems derive from constraints on the nature of sounds the auditory system is exposed to. Several other such constraints affect the ability to hear and separate sources.

Common Fate

One general class of constraints, called *common fate* by Gestaltists, holds that different parts of the perceptual field — of vision, audition, and probably other senses — that are associated with a source usually have common properties. These common properties are reflected in the auditory system, which has mechanisms to use these commonalities as cues for scene analysis. Many of these have already been discussed: common onset and offset, common frequency and amplitude variation, and harmonicity or common fundamental frequency. (Note: The latter commonality, not involving motion or change, would not fit Gestaltists meaning of common fate.)

Continuity

A general class of constraints, also reflected in auditory scene analysis mechanisms, is that of continuity or similarity over time. This constraint arises from the fact that most sounds tend not to change in character rapidly. A clarinet is not likely to suddenly sound like an automobile. A piano is not likely to drop its third harmonic halfway through a note. The principle of continuity, like accounting and allocation, applies at many levels of organization, from the continuity of partials in a note to the continuity of subject matter in speech. It is explored further below under event formation (section 4.3) and source formation (section 4.5).

Closure

Another important principle from the Gestalt school is *closure*, the idea that incomplete sensory stimuli can sometimes be closed, or made complete, by the perceptual addition of non-existent parts. An example of this may be heard in speech interrupted by brief noise bursts, in which the speech is perfectly intelligible even though some phonemes are lost. Indeed, subjects generally find it difficult to tell where the interrupting noises occurred; perceptual completion works so well that we are not even aware of the place of its effect [Miller50, Dirks70]. Visual instances of closure abound, as for example the effect in which fig. 4.3 is unintelligible because the obscuring element is missing. In fig. 4.4, closure is possible and the underlying pattern is

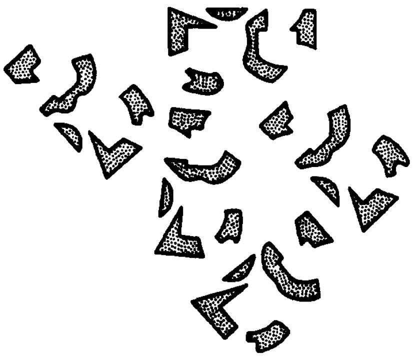


Figure 4.3: What is the underlying pattern? (Used by permission from Bregman.)

plain [Bregman73]. Perhaps the perceptual system can perceive of part of an object as being either present, absent, or obscured; fig. 4.3 leaves parts absent, while fig. 4.4 makes them obscured and hence amenable to closure.

Closure is an important part of a scene analysis mechanism because it helps preserve continuity of partials across interruptions, and helps at a higher level to preserve the continuity of events in streams.

Information from Other Senses

Cross-modal information, particularly visual information, undoubtedly has some effect on source perception. Moore points out [Moore89, p. 223] that this effect used to be exploited in movie theaters, which generally put a single speaker behind the screen. Though auditory localization information reaching the ears indicated only one source, voices seemed to come from people in different parts of the screen because of the overriding weight of visual cues. Another example is ventriloquism, in which the movement of the dummy's mouth makes us incorrectly localize the sound to it rather than to the true source [Massaro87, p. 84].

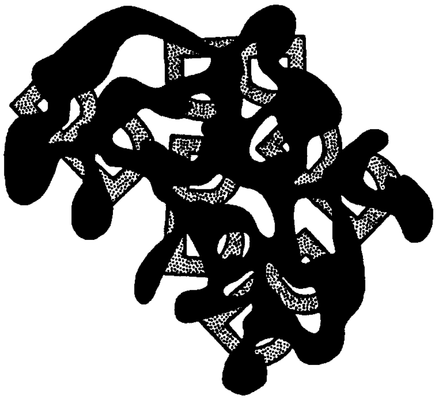


Figure 4.4: The underlying pattern is clear. (Used by permission from Bregman.)

4.3 Event Formation

The auditory system brings together information about a sound signal from various perceptual mechanisms to decide which events are present. This process, event formation, associates sound from across the time-frequency sensory field, grouping it into events as they occur for use by higher levels of the auditory system.

Event formation is primarily a spectral organization, placing sound energy represented by neural firings into the correct groups at each instant. This is not to say that the process is instantaneous, examining the spectrum and cue filters at each instant to decide on the correct grouping. The process clearly uses knowledge from the recent past, as for instance in Pierce's demonstration (fig. 2.5, page 31) where information about a recent onset contributes to the perceptual organization for a period of time after the onset occurs. Event formation even uses "knowledge from the future", as in this demonstration by Warren [Warren82]. In the top of fig. 4.5, a pure tone is played up to the start of a noise burst, then starts up again when the noise ends. The auditory system performs closure to give the perception of a single tone that continues through the noise. However, if the sound at the bottom of the figure is played — a tone that stops when the noise begins and does not re-start — then the

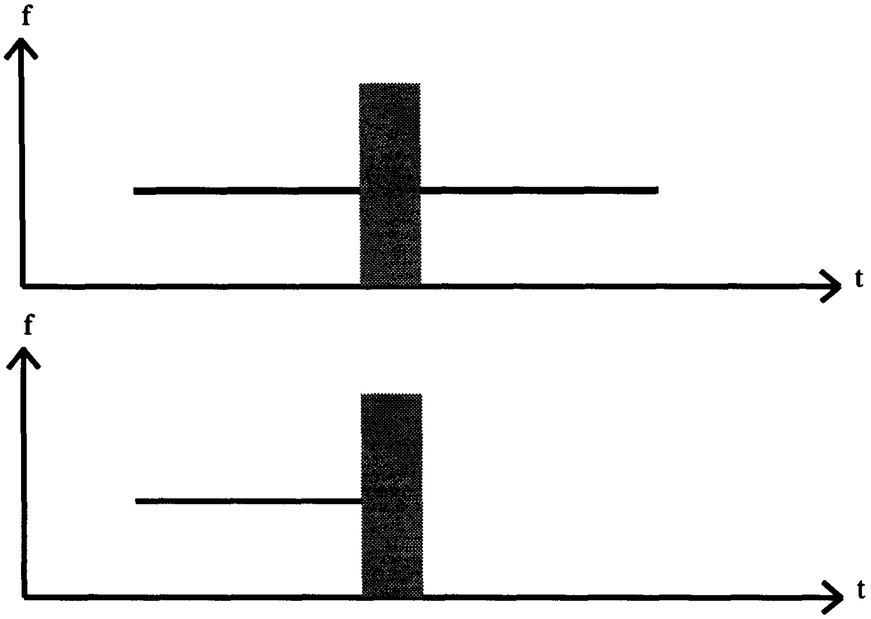


Figure 4.5: Masking experiment (after [Warren82]).

tone is heard as stopping at the time the noise begins. In other words, whether or not the tone is heard as continuing through the noise depends on what comes later. In this case, the process that is forming the tone event must occur after the noise stops to decide about its event's termination.

Grouping Partials

Much of event formation involves grouping partials together. Partials usually obey the common-fate constraints mentioned above, in that common onset, offset, frequency variation, and amplitude variation are all evidence that several partials belong to the same event. Partials also obey the continuity constraint: a partial that exists at one frequency at one point in time is likely to be at the same frequency, or a nearby one, a short time later. Partials do not flash into and out of existence randomly; they vary smoothly in frequency. Bregman and Dannenbring [Bregman73] studied this effect in an experiment diagrammed in fig. 4.6. They tested whether the high and low tones H_1 and L_1 split apart into separate sources when the sequences were played at various rates (100-225 ms for the steady-state portion of the sounds). There was a marked difference between the discrete case and the ramped and semi-ramped cases, with listeners tending to segregate H_1 and L_1 into separate events in the discrete case. The auditory system uses the continuity of the partial in the ramped case, and its implied continuity in the semi-ramped case, to group all of the available sound into one event.

4.4 Affinity Groups

How does the auditory system perform event formation? One possible way is by tracking partials and keeping some kind of evaluation of how well they associate with each other. In this conception, partials are placed in *affinity groups* — collections of associates partials — and remain there as long as evidence for their belonging is strong enough. For instance, if a partial has FV at a rate that matches that of the other partials in its affinity group, then this match constitutes evidence that the partial belongs to the group. Evidence can go against keeping a partial in a group,

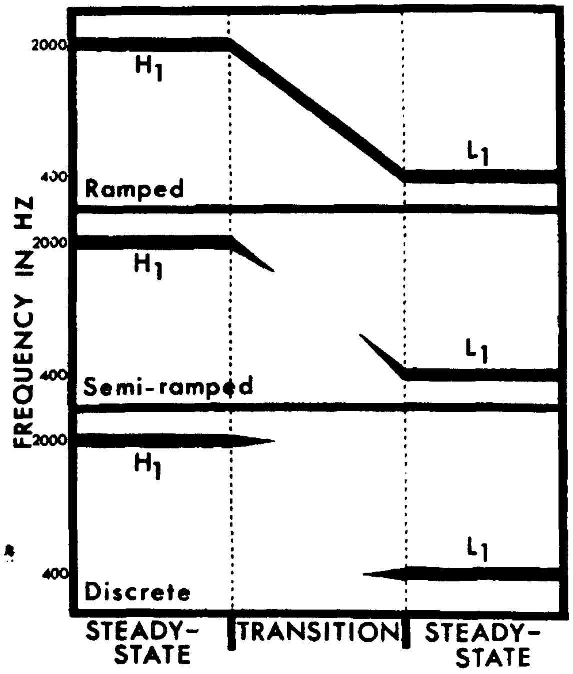

2000
H1
L1
400
Ramped
2000
H1
L1
400
Semi-ramped
2000
H1
L1
400
Discrete
FREQUENCY IN HZ
STEADY-STATE
TRANSITION
STEADY-STATE

Figure 4.6: Continuity of partials. Used by permission from Bregman.

too: if a partial begins at a different time than the members of an existing affinity group, then this onset mismatch is evidence that the partial does not belong to the group.

In accumulating evidence for the creation and maintenance of affinity groups, one must pay attention to the absence of cues in addition to their presence. For instance, the absence of harmonicity for a given partial tends to make it stand out as a separate source [Moore85b]. The absence of any sound at all drives the decision about when an event ends in the bottom part of fig. 4.5 above.

Contribution Strength

Each cue has a different contribution strength for its partial's presence in an affinity group, a strength that depends on the context of the event. Common onset seems to be one of the stronger cues, as Hartmann notes (see quote p. 31). Parsons [Parsons76], Hartmann [Hartmann88], and Cooke [Cooke91] all note the importance of harmonicity as a grouping cue. Common FV may be one of the weaker cues [Carlyon91], but it is important musically in that other cues are often missing and it is the only one available to hear out multiple voices. Cue strength may vary depending on sound characteristics, as when an onset in one frequency channel is perceptually unimportant because it arises from a partial that is varying in frequency.

Context for cue strength can come from learned patterns. For example, in listening to piano music, one listens for the percussive onset of each note. Violin notes have attacks that are much more spread out in time, providing less of an onset cue; in this case, the harmonicity and pitch separation cues may be acting most strongly.

Hysteresis

A final characteristic of affinity groups is that they display hysteresis. That is, an affinity group tends to stay together in the absence of evidence that it should split apart, and even for some period of time after the arrival of evidence that it should do so. Conversely, different events will fail to merge for a brief duration despite evidence that they should do so. This effect can be heard in Pierce's delayed-onset example of chapter 2.5, reproduced in fig. 4.7. In the later half of the diagram, as each partial

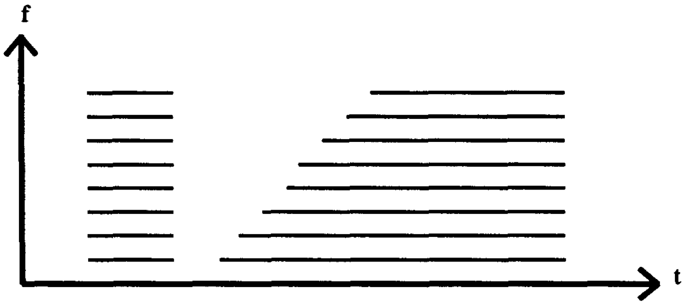


Figure 4.7: Pierce's delayed-onset example.

starts, it is heard as a separate source from the existing set of partials because it has a distinct onset. If hysteresis were not operating, then common onset would be the only principle in operation that separates each new partial from the mixture. In this case, as each partial entered, one would hear an onset but no isolated tone appearing as a separate source, for harmonicity would immediately capture the tone for the complex. This is not the case, however; each partial stands alone as a separate event for a brief period of time, roughly half a second, before merging with the existing partials. This independence stems from the definition of by hysteresis: Evidence must accumulate for the new partial's membership in the complex over some period of time until it becomes sufficiently strong to cause fusion. Bregman also observed this effect, noting that sounds tend to be heard as a single stream when they start but to split apart after several seconds of accumulating evidence [Bregman78a].

4.5 Source Formation

Source formation is the process in which events are grouped over time into coherent sources, each group perceived as a complete stream. Like event formation, it is a complex dynamic operation depending on many factors beyond the scope of this thesis. A few of these factors are listed here as a brief overview of the principles that could be used in a system that goes beyond the event-formation stage implemented in the next chapter. These factors are generally cues that furnish the auditory system with evidence for or against an event belonging to a source; whether or not a mechanism like the affinity groups proposed above would work remains to be seen. Also, the statement from chapter 1 bears repeating: Event and source formation are closely related processes, neither existing without the other.

4.5.1 Pitch Separation

Two successive notes near each other in frequency tend to be placed by the auditory system into a common source, and conversely [Dowling78, Fucks62]. This fact is salient in melody: Ortmann [Ortmann26] studied a large number of melodies, tabulating the pitch intervals between adjacent notes. He found that the intervals used in melodies tended to be small, with approximately a reciprocal ($1/df$) relation between interval in semitones and frequency of occurrence.

J. S. Bach used pitch separation in a technique called virtual polyphony. This can be heard in the C major solo violin sonata, in which a single instrument plays interleaved notes that are widely separated into three disjoint ranges of pitch. If the piece is played rapidly enough, the notes in each range are heard as a separate stream. In other words, one instrument can become several sources by virtue of pitch separation. This example has been thoroughly analyzed by Erickson [Erickson82].

Pitch separation has been used extensively as a grouping cue in psychoacoustic experiments. It was used in Bregman and Pinker's classic experiment (fig. 2.6, page 33) to induce tones A and B to group together in a single stream. Another example is the Bregman-Rudnicky experiment [Bregman75] diagrammed in fig. 4.8, in which the frequency separation of tones A and B separates them from the sequence of the X tones. Bregman points out, noting work by himself and van Noorden, that both pitch similarity and spectral similarity may be used as grouping cues [Bregman90, pp. 83-92].

Deutsch's scale illusion [Deutsch75] shows how strong the pitch similarity cue can be. This illusion consists of two simultaneous musical scales covering the same octave, one descending and one ascending. The first note of the descending scale is played in the left ear and the first note of the ascending scale in the right. Thereafter, notes

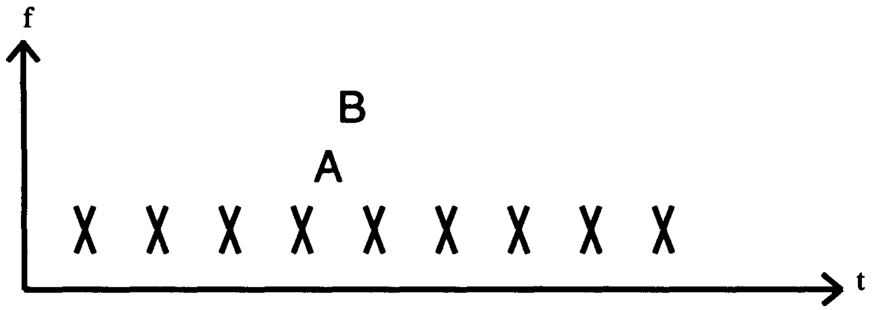


Figure 4.8: A different part of the Bregman-Rudnicky experiment.

in each scale alternate from ear to ear. The perception Deutsch reports is not a descending scale switching back and forth from ear to ear, or similarly an ascending one, but rather an illusory experience in which all of the higher tones are heard in one ear and all of the lower tones in the opposite ear. In other words, the auditory system is keying on the frequency of the tones, rather than the 180° shift in location induced by the ear-to-ear interleaving of the scales.

4.5.2 Timbre

Timbre is one of the most important characteristics by which we associate notes with separate instruments. Without it, we would often have no way of knowing which sequences of notes to follow when pitch lines of two instruments cross one another. A musical instance of the use of timbre for scene analysis is found in Robert Erickson's piece LOOPS. In it, he uses timbre contrasts and pitch range differences to play with the listener's idea of the extant sources.

Wessel has a convincing demonstration of the salience of timbre for source grouping [Wessel79], shown schematically in fig. 4.9. In the figure, X represents one timbre and O a different one. When the sequence is played slowly, the notes sound like three repeated ascending tones that change in timbre. At a high enough repetition rate, the notes of each timbre separate out into separate streams, with all of the X notes in one stream and O notes in the other. Since the X's alone form a descending sequence,

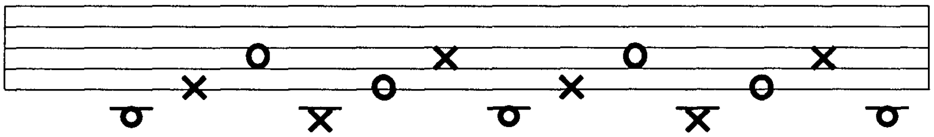


Figure 4.9: Wessel's timbre illusion.

as do the o's, the perception is of two descending sequences playing at once.

Bregman and Levitan also report an effect of timbre for scene analysis [Bregman90, pp. 86-90]. They played a series of tones that alternated relatively bright and dull timbres in a repeating sequence of four notes. Alternate pairs of notes had different pitches, forcing listeners to identify the tones in sources by either timbral similarity (spectral brightness) or pitch similarity. The results showed that whenever the timbres were sufficiently similar — when the "formant peaks" in the spectrum were at least as close in frequency as the pitches — then subjects heard the timbral grouping more easily than the pitch grouping.

4.5.3 Rate of Repetition

The effect of repetition rate on event grouping, like the effect of harmonicity on partial grouping, is so well established that it is frequently used as the independent variable in streaming experiments. Fast sequences of tones split apart into multiple sources more easily. The experiments and demonstrations already covered which use repetition rate for streaming include the ones in these references: [Bregman78b] [Bregman73] [van Noorden75] [Lakatos91] [Bregman75] [Erickson82] [Wessel79].

4.5.4 Number of Repetitions

The number of times that a stimulus is repeated influences the perception of sources in it. Bregman [Bregman78a] cites van Noorden as the first to notice this effect and describes his experiments with it. Van Noorden made up a set of tones comprising two higher tones H_1 and H_2 and a lower tone L, played in the sequence L-H_1-H_2-L. If these tones are played once, they are heard as a single four-note melody. Playing the sequence increasing numbers of times made the tones split apart into two streams. The experiment slowed down the speed of repetition to get a measure of the tradeoff between repetition number and rate of presentation as cues for grouping, showing that repetition number did indeed affect source perception.

Anstis and Saida also found a similar result [Anstis85]. They played a continuous sequence of tones that alternated between two pitches and asked subjects to judge whether they heard one stream or two. They found that the likelihood of hearing one stream was best at the beginning of the sequences and declined steadily with repetition.

4.5.5 Loudness

Sound pressure level has two effects on source perception. The first is that low signal intensities promote fusion. Hartmann [Hartmann88] reviews some of the experiments that use this effect. One of these is Houtsma's [Houtsma79], which tried to induce a sensation of pitch for harmonics spread fairly far apart in the spectrum. Since it is often difficult for the auditory system to hear such scattered frequencies as a single source and fuse them into a pitch, Houtsma used a sensation level of 20 dB SPL to promote fusion.

The other effect of loudness on source formation, somewhat disputed, is that sounds of equal or nearly-equal loudness tend to be grouped together. Van Noorden [van Noorden77] played a repeating sequence of 1-kHz tones, with alternate tones at different loudnesses. Near 35 dB SPL, at sufficiently high repetition rates, he found that only 3 dB of loudness difference was enough to cause fission of the sequence into two sources. (With a high rate of repetition and a greater loudness difference, he was also able to produce another perceptual effect that he dubbed roll: hearing the loud tones normally but the quiet tones at twice the rate.)

Bregman points out [Bregman90, p. 126-7] that van Noorden had his subjects *trying* to hear separate streams, which undoubtedly gave a lower threshold than would be found for involuntary fission. Indeed, the loudness difference needed for involuntary fission would almost certainly introduce forward and backward masking, greatly complicating the task of finding out whether loudness differences alone were responsible for fission. Bregman concludes,

> I do not think that loudness differences can be used to segregate signals in the same powerful way as frequency or timbre differences can. Loudness differences, however, are not irrelevant. I think sudden loudness *changes* do introduce boundaries into signals, but these boundaries are heard as being beginnings or endings of the louder event not the softer one. [Bregman90, p. 127]

Shepard has a demonstration in which stream separation is based on a loudness difference, leading to the paradoxical effect that a decrease in amplitude can cause an increase in salience [Shepard91].

4.6 Summary of Event and Source Formation Mechanisms

Event and source formation are part of a hierarchical structuring that the auditory system imposes on incoming sound stimuli. In constructing sound percepts, the auditory system must perform allocation and accounting, ensuring that a rough correspondence is maintained between the sound energy present at each frequency and the perceptual objects that explain the sound.

The auditory system uses several principles to make the correct association between sound stimuli and auditory objects. Common fate, including common onset and offset, common frequency variation, harmonicity, and common amplitude variation, enables the auditory system to make groupings of different frequency components. Continuity ensures that brief interruptions or changes do not terminate partials, events, or sources. Closure "fills in" obscured parts of perceptual entities. Information from other senses can also contribute to auditory grouping decisions. Hysteresis ensures that groups tend to stay together for a short period of time after the appearance of contradicting evidence.

Several cues affect the grouping of events into sources. In tonal sequences, events near in pitch tend to be grouped together, as do events with a similar timbre. Faster occurrence tends to split sequences of events apart into separate streams, as does repeated presentation. Quiet sounds tend to be grouped into a comomn source, while the effect of different loudness levels on grouping processes is unclear.

Chapter 5

Algorithm for Event Formation

This chapter presents an algorithm for event formation that incorporates some of the principles put forth in the last chapter. This algorithm is limited in scope, as its main goal is to demonstrate that the cue filters of chapter 3 produce results that really can be used for event formation. Another goal of this algorithm is to show how principles for event formation can be implemented. This algorithm does no source formation — that is, once events are detected, there is no attempt to place them in sequential streams of the type mentioned in section 4.5. That process could operate on the output of this one, but it is complex enough that I have not treated it here.

This algorithm satisfies the principle of using physiologically compatible processes at only a very high level of abstraction. It obeys some of the same constraints and expresses some of the same principles as the human auditory system, but the processes and data representations used in the two systems are probably not very similar. For instance, the use of thresholds for triggering or inhibiting actions during partial tracking, and a triangular matrix for affinity values, are undoubtedly not found in the brain.

This algorithm is also limited in that it uses for input only information from the cochleagram plus the features covered in chapter 3, onset and frequency variation. The other features mentioned in chapter 2 — harmonicity, amplitude variation, and location in space — are ignored. Any complete event formation algorithm would have to incorporate all of these cues to have a valid claim of being a model of human auditory scene analysis. Perhaps such an algorithm could be the natural outgrowth of this one.

5.1 Overview of Operation

> "The brain is a kludge."
>
> Marvin Minsky

Partials and Events

The algorithm operates by keeping track over time of two important descriptions of the sound: the set of partials and the set of events. A partial is simply a peak in the cochleagram over some number of time steps. Partials are found by noticing when they start, as revealed by the onset feature map that is one of the inputs to the algorithm, and then tracking them over time until they terminate. This tracking uses cues from the cochleagram, to track partial peaks (spectral peaks) over time and to notice when they disappear, and from the FV feature map, to track changes in the frequency of partials and to keep partials separate when they cross one another.

Events are formed of groups of related partials. Partials are put in close relation by having a common onset when they first start up, or by having common FV as they continue in time. Events can capture and lose partials over time, though this is relatively rare. The input feature maps can contribute evidence both for and against the association of a particular partial with a particular event, as well as contributing evidence both for and against the tracking of a particular partial to particular frequency channels.

Output

The output of the algorithm is the set of events that were found during the processing. Each such output event is stored in a separate feature map, which is the same size as the cochleagram map representing the input. Each such output map has a zero value everywhere but at those time/frequency points where a partial for this event existed;

at such a point, the value is the corresponding value from the cochleagram, showing the analog of intensity of neural firing rate for that T×F point. At this level, events can be *detected*: each one found by the algorithm is produced as a distinct output.

5.1.1 Cycle of Processing

In more detail, here is how the algorithm works. This description does not yet have all of the particulars, but covers enough of them that the fuller description to come will make sense. Unclear meanings will be explained more fully in succeeding sections.

The process moves through the duration of the input sound signal, one step at a time. A step in this case is a single time step at the sampling rate of the input maps; in the examples in this chapter the feature maps have been downsampled to a rate of 441 Hz, so each time step is $\frac{1}{441} = 2.27$ ms. This unit of time is called a *time slice* or a *tick*.

At the start of the processing cycle for each time slice, there may be some existing partials and events. (This is of course not true at the very beginning of the sound, where no partials or events exist yet, but it is true in the general case.) Existing partials are updated by finding what frequency they move to in the new time slice. They are also checked to see whether their peak cochleagram firing intensities have fallen below a minimum threshold for existence, and are terminated if so.

Next, onset information is processed. The onset map is checked to see if there are any noticeably strong onsets; onsets that may be due to FV of an existing partial are ignored. If there are onsets, the peak values of the onset map are found and turned into new partials. New partials that begin at the same time slice, or nearly the same slice, are grouped in one event on the evidence of common onset.

Next, the algorithm processes FV information. For each event, the FV values of the partials that compose the event are compared. If a partial has FV sufficiently different from that of the other partials in its event, it is separated from the event and placed into its own newly-created event. Also, if partials in two different events have sufficiently similar FV, then the two events merge, assembling their partials into one unified event.

Finally, information about whatever events currently exist, if any, is recorded in the output feature maps. For each partial, the current time slice and the partial's current frequency are used to index a T×F point in the output map for whichever event the partial belongs to, and this point is marked to provide a record of which partials were part of the event.

When the entire sound has been processed, tick by tick, the resulting feature maps, one per event, are saved away and the process is done.

5.1.2 Affinity Groups

Partials and events, two of the most important data structures of this algorithm, have already been introduced. Another important one links these two types: the affinity group. An affinity group is simply a set of partials that are closely enough related — have a high enough affinity for each other — that they are considered to belong to the same event. At any time, each event has its own affinity group that consists of those partials that are currently members of the event. Partials can move into and out of an affinity group as a consequence of the algorithm steps outlined above, but the event-affinity group relation is fixed and one-to-one. A partial can belong to at most one affinity group — an expression of the principle of exclusive allocation — but it is not required to: a partial can be *floating*, without being attached to any affinity group, for a brief period just after creation (and in a few other cases described below) until enough information accumulates to assign it to an affinity group — or to form its own affinity group if none of the existing ones are acceptable.

Affinity Function

Closely related to the affinity groups is the affinity function $\mathit{aff}(p_1, p_2)$, which keeps track of how strong the belief is at any moment that two partials p_1 and p_2 belong to the same event. $\mathit{aff}(p_1, p_2) = 1$ represents certainty that p_1 and p_2 belong to the same event and $\mathit{aff}(p_1, p_2) = 0$ represents certainty that they should be separate. This function aff is symmetric — *i.e.*, $\mathit{aff}(p_1, p_2) = \mathit{aff}(p_2, p_1)$. The values of the aff function for a partial p are normally initialized to 0.5 when p is created (*i.e.*, $\mathit{aff}(p, p_i) = 0.5$ for all partials p_i). The values of aff are changed by various processes over time, as will be described below. Appendix A describes an efficient method for storing and accessing a symmetric function such as *aff* that needs to grow to arbitrary size. (Note: $aff(p_1, p_2)$ should probably be written as $aff(p_1, p_2, t)$, as its value changes over time. However, time is left implicit, with the understanding that the time value is the current time slice.)

5.2 Tracking of Partials

The last section presented a chronological view of the steps executed at each time slice by this event formation algorithm. This section presents, in greater detail, the same process for another point of view: the partial's. Here is described when and how a partial gets created, how it is updated from tick to tick, how it might diverge into two partials or merge with another and disappear, and finally how it may cease to exist. The next section has a similar description for events.

Notation

Many of the processes described here make use of numeric parameters for thresholds, time durations, frequency spreads, and so on. To call attention to the fact that such a parameter comes into play, I use the notation `[param]`. Here `param` is my own cryptic name for the parameter in question; the exact name does not matter much, as the important matter is that it exists as a numerical quantity. These parameters are summarized in Appendix B.

5.2.1 Creation of Partials

Onset

Partials are created when an onset occurs. The exact details of what constitutes an onset are given in section 5.4.1, but suffice it to say here that a partial is created when there is enough intensity in the onset map at a particular frequency channel. Creation of a new partial at a given frequency channel is inhibited if an existing partial has FV that is moving it into that channel. The extent of this inhibition is controlled by spread factors in the frequency `[onFvLoSize,onFvHiSize]` and time `[onFvLook]` dimensions, and its activation by a threshold of FV intensity `[onFvStop]`.

Divergence

Partials can also be created by divergence. Occasionally, a single partial will split in two, so a new partial must be created for the split-off one. More often, small noisy variations in the intensity of nearby frequency channels will give rise to two or more spectral peaks that are all reasonable continuations of a given partial. In this case, the algorithm must track all of the possible continuations until it is clear which one is the real partial. It does so by creating new partials.

Affinity Values of New Partials

A new partial which was created by an onset has its *aff* values — that is, values of $\mathit{aff}(p, p_i)$ for all other partials p_i — initialized by commonality of onset. If p and another partial started within a certain time `[onParRecent]` of each other, then the *aff* value is set to 1. If not, it is set to 0. If p is associated with another partial this way, and the other partial is part of an event, then p joins that event. Otherwise, p becomes a floating partial.

A floating partial created in this fashion can exist for only a limited amount of time `[maxNewAge]`. After the time limit has expired, p undergoes *event forcing*, in which it is matched up against each event in turn to determine which one to join. For each event, the average affinity between p and that event's partials is computed; whichever event has the highest average is the winner, and p joins that event. If none of the averages is high enough `[minMeanAff]`, p is placed in a new event.

A partial created by splitting has its *aff* values initialized to 0.5 to signify that it is not yet associated with any event. Such a partial is floating, and also has a maximum time `[maxGroupAge]` that it can remain so before undergoing event forcing. The process is actually a little more complex than this, and will be elaborated in section 5.2.4.

The reason for having floating partials is that a new partial often has insufficient evidence to be placed into any event; some time must pass before there is any evidence available. Partials cannot, however, remain floating forever; they need to become associated with some event, just as the auditory system seeks to "explain" what it hears (or at least hears clearly) by associating it with some source.

5.2.2 Continuation of Partials

After a partial is created, it is tracked from tick to tick. Each partial is centered on a peak among all the frequencies of the current time slice of the cochleagram — in other words, a local maximum along a vertical strip. The principle of continuity mentioned in chapter 4 is applied by having each partial, in making the transition from tick t to tick $t+1$, move to the nearest peak in $t+1$ to its peak in t. If a partial must change in frequency in the transition from t to $t+1$ — if its peak at t is no longer a peak at $t+1$ — it follows the local gradient to the first peak. So a partial moves only a little in frequency, preserving continuity.

Termination and Continuation

The mechanism that updates partials from tick to tick also watches out for partials that come to an end. If the neural firing rate represented by cochleagram value drops below a certain threshold `[earThresh]`, then the partial dies. Actually it is not quite this simple, as there is a mechanism to implement closure. A partial is permitted to remain below threshold for a period of time `[earRecent]` provided it later goes supra-threshold. This period of time is brief enough (about 5 milliseconds) that it enables rapidly repeated notes to be identified as separate events. This mechanism ensures continuation of a partial through intrusive noise, or through the destructive interference that sometimes occurs when partials from two different events cross or come very close in frequency. Such interference can be seen in the circled region in fig. 5.1.

Updating Floating Partials

Another function performed at each tick is to check whether floating partials have accumulated enough evidence to place them in an event. Various steps of the algorithm

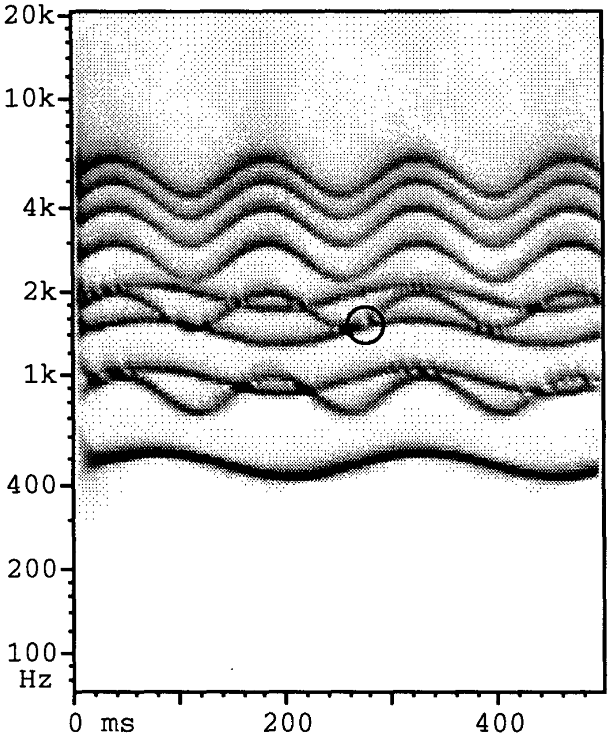


Figure 5.1: Destructive interference of partials.

described below modify the *aff* values reflecting, the affinity of a partial p for each of the other partials. Here, a floating partial is matched against each of the existing events to determine whether its average *aff* value with the partials in the event is high enough `[minMeanAff]`. If so, the floating partial is placed into the event; this avoids event forcing later on.

5.2.3 Merging of Partials

Sometimes, two partials merge together and become one. More commonly, two partials cross over one another, enduring a period of time in which the two partials' peaks are at the same frequency; this can be seen often in fig. 5.1. The partial tracker must handle these cases correctly, removing a partial in the former case and remembering the partial's affinity groups for eventual reuse in the latter.

Merging a Floating Partial

The simplest action is performed when a floating partial merges with another partial. Floating partials show up just after an onset or just after a partial splits. In this case, the merging partial is almost certainly an artifact, caused by noisy data inducing a temporary local maximum in the cochleagram. For this reason, such partials are terminated upon merging. I have encountered no cases in which terminating a floating partial removed a "real" partial, or one that was other than a short-lived artifact.

Merging Other Partials

The case of two partials which are both associated with events coming together deserves more care. In this case, the algorithm must remember the events associated with both partials, in case the partials later move apart again and must be re-united with their events. This remembering is an application of the principle of continuity, for it ensures that the perceptually relevant elements, the partials, continue through interfering sounds. The partials, which now have a common peak, are put into a *neighbor group*, a collection signifying that the partials have different events, were once separate, and should become separate again if the peak splits. Neighbor groups will be explored more thoroughly in the section on diverging of partials, below.

Merging and Exclusive Allocation

The principle of exclusive allocation is applied to merging partials by having a limit `[maxGroupAge]` to the amount of time that two partials may stay merged. The idea behind this is that each partial should be associated with only one event at a time. If two partials merge, they become in effect a single partial, and this single partial has two associated events as discussed above. Having two events is acceptable for a short period of time — indeed, it is necessary to handle crossing partials — but exclusive allocation demands that this situation not continue indefinitely.

In ensemble music, instruments playing notes that make up a chord have partials that coincide in frequency. In this case, it would be better to have some mechanism to associate one frequency with more than one event. Such a mechanism would work best along with the harmonicity cue to identify which frequencies are likely to make up a given tone. Since my implementation does not include a harmonicity filter, no attempt was made to handle coinciding partials.

The merging of partials could be handled in a more accurate way by comparing the relative intensities of the two partials before merging and the one partial afterward. If the intensity of the one partial is the sum of the other two, then the situation is recognizable as two partials which have merged. It would be even more accurate to take into account the amplitude variations of the other partials in the merged partials' events, since any such variation would affect the summed amplitude as well. A factor complicating all of this is that individual partials are nearly sine waves. When two sine waves of the same frequency (*e.g.*, two merging partials) and same amplitude are summed, the amplitude of the result can vary from twice the amplitude of the sine waves (for a phase difference of 0 between the two sines) through the same amplitude (for a phase difference of $\frac{2\pi}{3}$) to zero amplitude (for a phase difference of $\frac{\pi}{2}$). Also, since the ear model's gain control is non-linear, the value in the cochleagram of the sum of two intensities is not the sum of their independent cochleagram values, so a simple linear additive model is insufficient.

None of the factors and mechanisms from the last paragraph were used in the implementation here, because of the complications listed and because they weren't needed for the test sounds used.

Averaging of Affinity Values

One other thing that happens when two partials merge is that their *aff* values are averaged. This is done because the two have now effectively become one partial, so they need to have a unified set of values that relate the partial to other partials. Before this step of the algorithm was implemented, when partials retained their *aff* values after merging, crossing partials did not split properly. This will be covered further below.

5.2.4 Diverging of Partials

Partials sometimes diverge into two or more partials. This can result from several causes. One is the rare case in which a partial really does split and become two. A much more common case happens when two or more partials have merged, then diverge again after a brief period of time. Another common case is a *phantom split* — a brief burst of noise in the data that temporarily looks like a new partial splitting off.

Each of these cases can be seen in the circled regions of fig. 5.2. Two partials diverge near letter **a**; in this case there really are two partials near the beginning of the sound, but the noise burst caused by the onset of sound pretty much obscures them. Two cases of partials merging and diverging are seen at letters **b** and **c**. The former is an instance of two partials crossing, the latter, two partials which come together, touch, and diverge again without crossing. Phantom splits can be seen at letter **d**, where interference from two nearby partials plays havoc with the sound. The algorithm must deal with all of these cases.

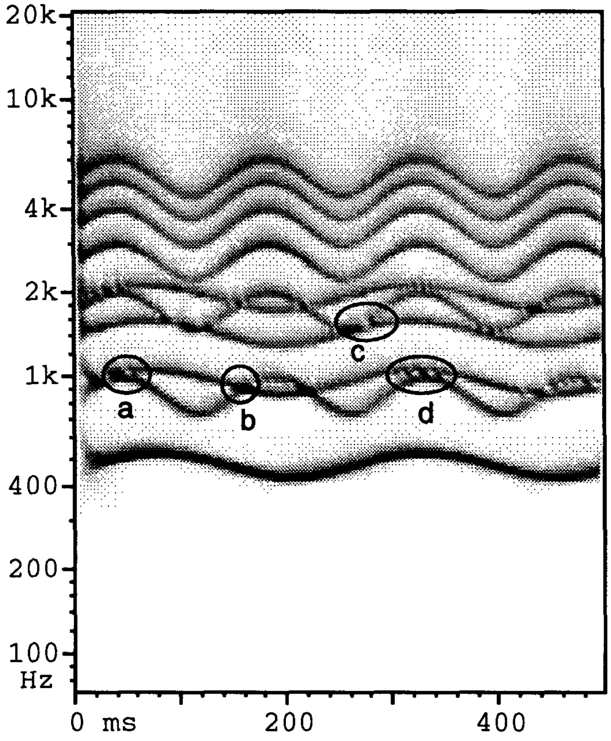


Figure 5.2: Partial divergence phenomena.

Phantom Splits

The algorithm handles phantom splits in several different ways. The simplest, mentioned above, is that newly split partials die easily. That is, any partial that splits off from an existing partial is kept in a neighbor group with the original partial, and if it merges back in with the original (or another one), the newly-split partial is immediately terminated.

Also, split partials are made difficult to create by the mechanism which looks for splitting partials. This mechanism, at each step, checks the neighborhood of each partial peak `[splitNear]` to determine whether there are any other nearby peaks that may be the beginning of a partial splitting off. Such a nearby peak is ignored if it is already the peak of another partial. As a precaution against phantom splits, the partial is also ignored unless there is a valley, of some minimum fraction of the peak height `[valleyDepth]`, between the main partial and the new peak. If this valley is not present or not deep enough, the peak is ignored.

Neighbor Groups

Neighbor groups, mentioned above, are groups of partials which are near or at the same frequency. They are created when two partials merge and when two diverge.

The principal function of a neighbor group is to keep track of which events may be associated with the partials in the group. The relationship between neighbor-group partials and events is a bit more fluid than that between normal partials and events, in that partials in a neighbor group can swap their events around relatively easily. For example, suppose that two partials p_1 and p_2, with events e_x and e_y respectively, come together and merge. They form a neighbor group. A short while later suppose the peak of the merged group splits; a reasonable assumption is that p_1 and p_2 are diverging again, perhaps because they just crossed each other or touched briefly without crossing.

The problem now is that it's not clear which of the two split partials (*i.e.* which peak) belongs with which event. Should it be p_1 that is assigned to the new peak, associating the peak with e_x, or p_2, associating it with e_y? The solution adopted in this algorithm is to make a guess, but to allow easy correction in case the guess is wrong. The guess is made based on similarity of frequency variation. That is, the FV values at the main and newly-split peaks are compared with the average FV values of all the partials in e_x and e_y. Whichever arrangement — e_x associated with the new peak and e_y with the main one, or vice versa — produces a better match is used.

This is but a guess, however, and it can be overturned easily. The newly-split partials are kept in a neighbor group. They initially have the same *aff* values, for they were recently merged, but as time goes by their *aff* values change independently as described below. At each tick, both partials are checked to see if either one has accumulated enough evidence to be certain that the guess was right — *i.e.*, if the mean *aff* value between the partial and the other partials in its event is sufficiently high **`[minMeanAff]`**. If so, the partial breaks out of the neighbor group and becomes solidly attached to its event again. (If there had been exactly two partials in the neighbor group, as in the example above, then the neighbor group now has only one member, and effectively ceases to exist.)

Event Forcing for Neighbor Groups

After a partial p has been in a neighbor group for a certain amount of time, it is forced into an event **`[maxGroupAge]`**. This is done by checking the match between p and each of the events in p's neighbor group. Whichever event has the best affinity for p , as measured by average *aff* value of p and the event's partials, claims p from the group. If p had belonged to a different event because of a bad guess, then this event is remembered in the neighbor group on the assumption that some other partial in the group probably belongs in that event. If none of the events has high enough affinity for p , a new event is created and p is placed in it.

Unassociated Partials

If a divergence happens to a partial p which has no neighbor group — for example, a partial that had not recently merged with another — then a new partial is created for the split and both partials are placed in a neighbor group. This new partial has no associated event, though like any partial in a neighbor group it can claim p's event if it gains enough affinity with the partials of that event. Remember also that such a partial will disappear upon merging with any existing partial. If the new partial lives long enough, it and *p* must decide which events to join (or create) as in the previous paragraph.

Remembering Peaks

A partial in a neighbor group is considered to have an uncertain association with its event, it any. For this reason, such a partial does not record its peak in the event's feature map at each tick as normal partials do. Instead, it simply remembers its succession of peaks until it is removed from the neighbor group and becomes definitely associated with an event. At that time, it records its past peaks in the event. This "writing in the past" is the only place in the event-formation algorithm at which information propagates backward in time, or equivalently that the mechanism looks into the future to decide about the present. The time scale involved, a maximum of 18 ms, is fairly short, though somewhat longer than the time scale of backward masking in the auditory system in which information also "goes backward in time" [Elliott79].

To sum up: A partial *p* splits when a new peak appears near *p*'s peak with a deep enough valley in between. The new partial, or an existing one if *p* had recently merged with another partial, is placed in a neighbor group with p, with a guess about which partial belongs with *p*'s event. If enough evidence accumulates to conclude definitely that either partial belongs with its event, then it is removed from the neighbor group. Otherwise, after a certain amount of time has elapsed, the partials are forced into events and the neighbor group vanishes.

5.2.5 Termination of Partials

The algorithm runs a termination process on a partial for one of three reasons: The partial is a newly-minted one that merges with an existing partial; it is not newly minted, but merges with another partial for too long a time `[maxGroupAge]`; or its peak firing rate in the cochleagram drops below a minimum threshold `[earThresh]` for too long a time `[earRecent]`. In the computer program, the terminator also deals with partials at the end of processing a sound.

Upon termination, a partial is removed from any neighbor group and event it belongs to. It also records its past peaks, as described in the previous section, if it was in a neighbor group and had not recorded the peaks previously. The event the partial records in is just a best guess based on mean *aff* values of the partials in each event.

5.3 Event Handling

An event is created when there is a partial p that needs to join an event, but no existing event has partials with high enough affinity for p . This happens often at the onset of a note in music, for then a number of partials commence at or very near the same time. (As mentioned in section 2.6, notes in ensemble playing are typically spread out enough in time that they can be heard as separate events.) These new partials have high affinity for each other because of their common onset, but low affinity for other existing partials. A new event is needed to contain them. A new event can also be created when partials diverge, as described above: If a new partial exists for a long enough time, it is forced into an event. At that time, if none of the existing events has a high enough affinity for the partial, a new event is needed for it.

Whenever a partial initiates a new event, it can bring other partials with it. If a partial starts a new event because it was just created by an onset, then it captures other partials with the same onset time as well as partials with an onset within a short time `[onParRecent]`. If a partial has split off from another and eventually initiates a new event, then the algorithm tries to enlist other partials in the new event. It does this by going over the other partials one by one, seeing whether the mean *aff* value between it and the new event is higher than the mean *aff* value between the it and the event it is currently in. If so, the partial joins the new event.

5.3.1 Order Dependence

This brings up the principle that the algorithm's operation should not be dependent on the arbitrary order in which partials and events are stored in the computer's memory. The above description of bringing partials to a new event *is* order dependent, in this way: The description says that each partial is tested for a high enough mean *aff* value between it and the new event, joining the event if so. If this were really the case, then partials that join the event early would affect the testing of partials that might join later, since their membership in the event would affect the mean *aff* value. In other words, the order in which partials are tested could affect which ones join the new event.

One aim of this algorithm is to avoid such *order dependence* — operation dependent on the arbitrary order of storage. In this case, it is avoided by marking partials that are to move on one pass through the list of partials, then actually moving them on a second pass. This process is iterated until no more partials move.

5.3.2 Event Continuation

There is fairly little work to do in maintaining an event from one tick to the next. For each event, the algorithm walks through the partials associated with the event and records the peak frequency of each partial. The value stored in the output feature map is the firing intensity value from the cochleagram. This makes the displayed output feature map reflect the input, easing the task of seeing that one was derived from the other. In addition to the peak frequency, values from nearby adjacent frequency channels `[loSide,hiSide]` are copied to the output map, provided they exceed a certain minimum `[checkThresh]`. This ensures that the entire width of the partial is captured in the output map.

5.3.3 Event Diverging

Events sometimes diverge. This can happen in musical sound because something thought to be one sound eventually breaks in two, as in the McAdams oboe sound in fig. 3.44 (p. 97), or because a guess that some partials that start together belong to one event is later overruled, as in the sound of fig. 5.2. These are really two instances of the same underlying process: The auditory system uses whatever information it currently has — here, common onset and FV — to make one decision about extant events, only to override that decision later in the face of additional accumulated evidence.

FV Drives Event Diverging

Since onset and FV information are all that the algorithm has to work with, and since onset information applies only at the beginning of a partial, the only information source that can induce event splitting is the FV feature map. At each time slice, the FV map updates the *aff* value between each pair of partials in a way detailed below. Later, each partial which is in an event is checked to see if its mean *aff* value with the other partials in the event has fallen below a minimum threshold `[eDivThresh]`. If so, the partial is considered no longer closely associated with the event, so it is removed from the event and placed in a new one. It also grabs any other partials it can in the process described above.

Symmetric Diverging

A newly-diverged event of course has a feature map for keeping a record of the partials in it. Note that the old event — the one diverged from — and the new event are perfectly symmetric, in that there is no reason to distinguish one as 'old' and one as 'new'; each has the same history, the same set of partials up to the time of split. For this reason, the new event's output feature map is initialized from the old's so that they show the same history. One could think of going back to fix up the previous partial tracks, cleaning up the output maps in light of the new association of partials and events. This is not done because it involves too great a step into the past, making it a distinctly non-auditory process.

No attempt is made here to handle duplex perception. This algorithm is admittedly incomplete in this respect; as stated above, its main purpose is to illustrate the effectiveness of the filters in chapter 3.

5.3.4 Event Merging

Events in music can also merge. This happens in Pierce's sound (fig. 4.7) where each harmonic is a new event when it begins but soon joins the rest of the harmonics in the tone complex. Notes in unison, or sometimes in chords, can also merge in a similar fashion — the notes are initially heard as separate events, but if they persist long enough their individual identities are lost and they become one event. This merging could be implemented fairly easily by comparing *aff* values between the partials of pairs of events. It has not been implemented here for the simple reason that I haven't yet tried to process any sounds that require event merging.

5.3.5 Event Termination

An event ends when it has no more partials. Partials leave an event either because they join another event or because they come to an end. In either case, the partial is removed from its event so that it no longer contributes to the output feature map.

There is no application of the Gestalt closure principle here. That is, an event that has no partials is not left alive for a brief period in case the lack of partials was merely a temporary phenomenon and the event event really should continue. The reason for this is simple: An event has no identity of its own; its expression of its rôle in the sound is entirely through the set of partials it contains. Once its partials are all gone, there is no way for it to become associated with existing partials. To put it mechanistically, the only time an event acquires a partial is when the *aff* values between the partial and the event's partials are high. If the event has no partials, it can't have *aff* values with another partial, and so can't acquire the partial.

5.4 Onset and Frequency Variation Information

Up to now, details about exactly how onsets are detected and how FV information is used have been glossed over. This omission is corrected in this section. The processes and methods described here are driven even more than those previously by practical considerations. That is, if it is unclear why a particular method is used or a particular threshold applied, the most likely reason is that it was necessary to make the program work.

A note on terminology: In this section, the term *feature map* refers to a T×H two-dimensional image as produced by one of the kernels (onset or FV) of chapter 3. Thus, 'onset map' does not mean the array of such images produced by the entire set of onset kernels; it means only one of these images.

5.4.1 Onsets

An onset is first noticed when the sum of values in a time slice of one onset feature map rises above a threshold `[onThresh]`. The requirement is actually slightly stricter than this: the time-slice sum must stay above threshold for some number `[onConsec]` of consecutive time slices, a number that is proportional to the time spread of the onset kernel. When this consecutive-slice requirement was not present, noise in the input data triggered too many spurious onsets with the short-time onset kernels.

Onset Frequencies

The above process detects when an onset occurs. Next the particular frequency channels at which partials should be created must be found. The algorithm looks at all of the frequency channels, finding those which are local maxima in the cochleagram and which have values above a threshold `[onLowThresh]`. This threshold is necessary to eliminate peaks that represent nothing more than random fluctuations of near-zero firing intensities. Each peak must not be in an inhibitory period, as will be explained below. Peaks must also pass the tests of not having a partial at the same frequency that recently `[onTime]` started, and of having valleys of a minimum relative depth `[valleyDepth]` between the peak and the nearest partials on the high- and low-frequency sides. This requirement prevents noisy data which makes a small local peak near another large one from generating a spurious partial.

A New Partial

Once such a peak is found, it becomes a new partial. In order to prevent an onset at the very next tick, such a partial inhibits further onsets until the onset feature map value falls below a threshold `[onStartThresh]` and then for a short while longer `[onTime]`. This refractory period prevents small dips in the onset value over time caused by noise from re-triggering an onset. The refractory period actually depends on the onset kernel size, with longer kernels having longer refractory periods.

As mentioned above, new partials are grouped by common onset into the same event if they start within a certain time limit `[onParRecent]`. The requirement is actually slightly broader than this: After one partial begins, a second one must fall within the time limit after the first partial falls below a threshold `[onStartThresh]`. This keeps slower onsets together in the same event.

Preventing Order Dependence in Onsets

To prevent order dependence, new partials at any slice are processed in order from the highest peak value in the cochleagram to the lowest. Actually there are two arbitrary orderings which could give rise to order dependence here. One is in the set of local maxima across frequencies of a time slice that gives rise to new partials; these are the peaks that are processed from highest peak to lowest. Order dependence can arise here because if two peaks near each other in frequency do not have a sufficiently deep valley between them, then whichever peak becomes a partial first will prevent the other from becoming one. The other way order dependence can creep in is in the order that the onset feature maps for different kernel sizes are examined. This dependence is prevented by checking all of the onset maps at once for supra-threshold values; only after all frequency channels of all of the onset maps have been collected do partials start getting created.

Onset Inhibition by Frequency Variation

One last factor affecting onset detection is inhibition due to FV. Whenever there is a partial at a certain frequency (*i.e.*, its peak is at that frequency), and there is FV evidence that the partial could be moving into nearby channels and causing onsets to be triggered, then an inhibitory period `[onFvLook]` begins for that frequency, a period that depends on the size of the onset kernel. The FV evidence is simply a value in the FV map above a threshold `[onFvStop]`. The partial is assumed to have a certain spread in frequency `[onFvLoSize,onFvHiSize]` over which it could be triggering such spurious onsets. Note that inhibition is frequency-specific so that FV at one frequency need not affect potential new partials at other frequencies, and onset-kernel-specific, since longer-duration kernels respond longer to FV and so need longer inhibition.

When a new partial is created, its *aff* values with other existing partials are initialized by deciding whether the partial has a common onset with each of the others. If so, the *aff* value is set to 1; if not, it is set to 0.

5.4.2 Frequency Variation

The handling of FV information is quite a bit simpler than that of onset information. Aside from the onset inhibition mentioned above, the only effect FV information has is to update the affinity values between each pair of partials. Other processes described above use these updated *aff* values to group partials together in events, or to split them off from events.

The process of updating *aff* values is fairly simple. First, only FV maps at which there is significant activity `[fvSumThresh]` in the current time slice are attended to. For each partial pair (p_1, p_2), and for each FV map f, if the FV values in f at p_1's peak and p_2's peak are within a threshold `[fvDiffThresh]` of each other, and are non-zero, then $\mathit{aff}(p_1, p_2)$ is incremented by a small amount `[fvIncr]`. If they are not within this threshold, then $\mathit{aff}(p_1, p_2)$ is decremented by a small amount `[fvDecr]`. The increment and decrement values are fairly small, since they can potentially be added to or subtracted from $\mathit{aff}(p_1, p_2)$ once for each of the nine FV maps, and this is done on every tick. Having them small induces hysteresis, in that accumulation of evidence for event splitting or merging takes a little while; events don't split or merge instantly when they have common FV.

There is no order dependence in the processing of FV maps: Since the only effect of such processing is the incrementing or decrementing of the *aff* values, and since addition is commutative, the order of operations is irrelevant.

5.5 Evaluation

How well does this algorithm work, given that the cue filters from chapter 3 are providing the input feature maps? This section presents the results of running it on some test sounds.

The last few sections have mentioned the parameters that control operation of this event formation algorithm. A few of these are varied from the different sounds; this is necessary mainly because of level differences in the input signal. The gain control stages that are part of Lyon's ear model take care of most of these level differences, but the ear model is designed to leave some of the difference still present in its output. It is this difference that is made up for in the changing thresholds from sound to sound.

The first sound to test is in fig. 5.3, reproduced from chapter 3. Recall that this is the start of a toccata played on the piano. The challenge here is to separate out each piano note, capturing its partials as a separate event. There are seven piano notes in the fragment — two of them happen nearly simultaneously at 340 and 360 ms.

The resulting events from running the algorithm are the images in figs. 5.4-5.7. As you can see, the algorithm found each piano note as a separate event, though it did not always find all of the partials for each note. The durations of the notes are usually about right, though some of the partials are cut off short.

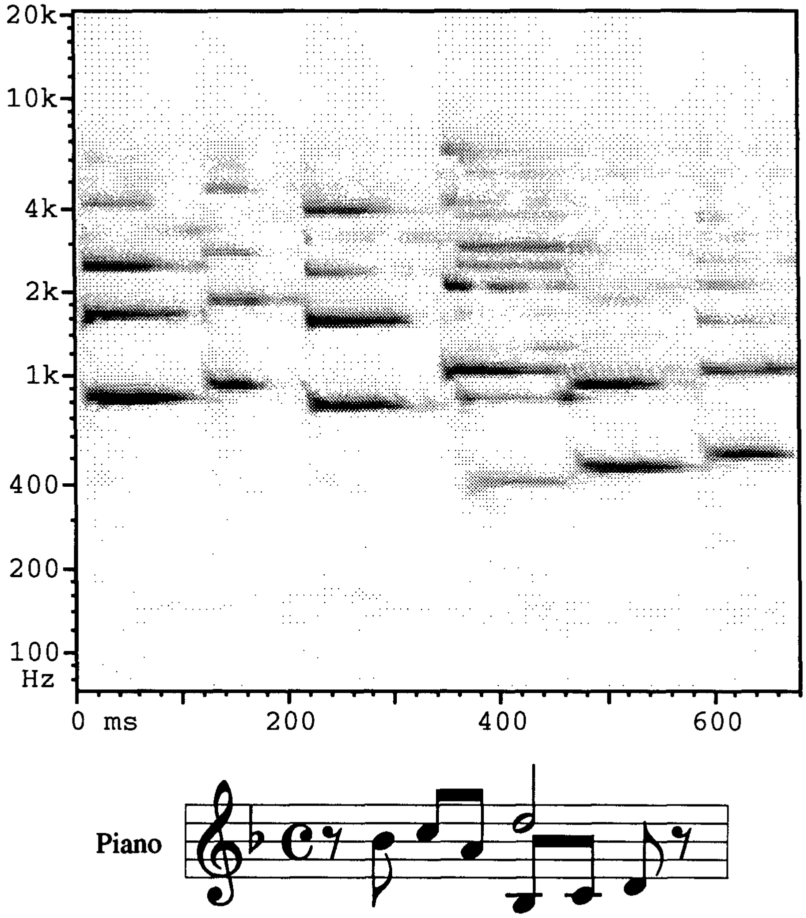


Figure 5.3: Frescobaldi toccata played on a piano, and score.

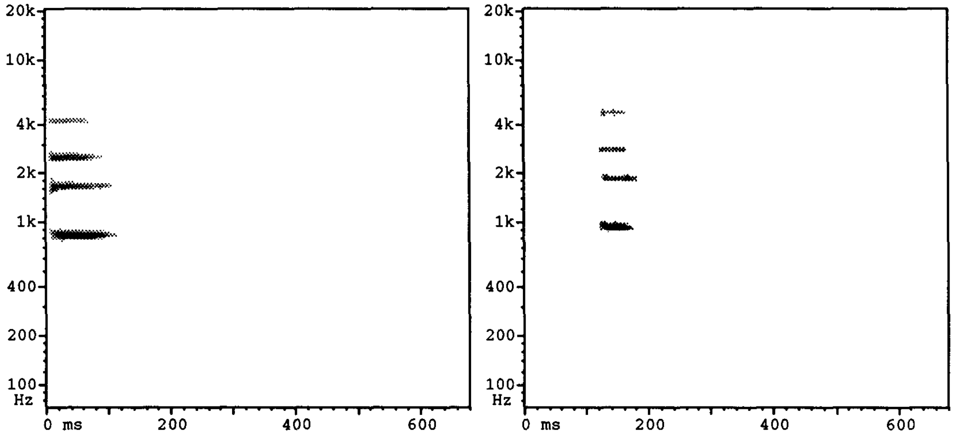


Figure 5.4: Events 1 and 2 from the Frescobaldi toccata.

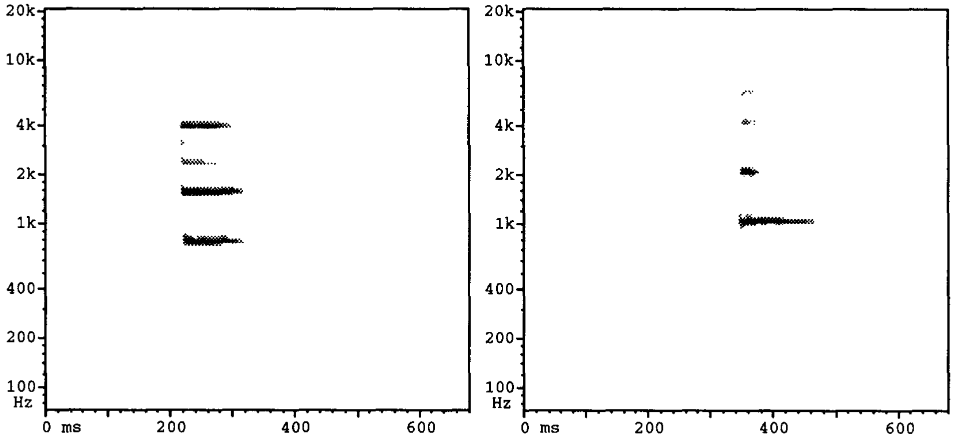


Figure 5.5: Events 3 and 4 from the Frescobaldi toccata.

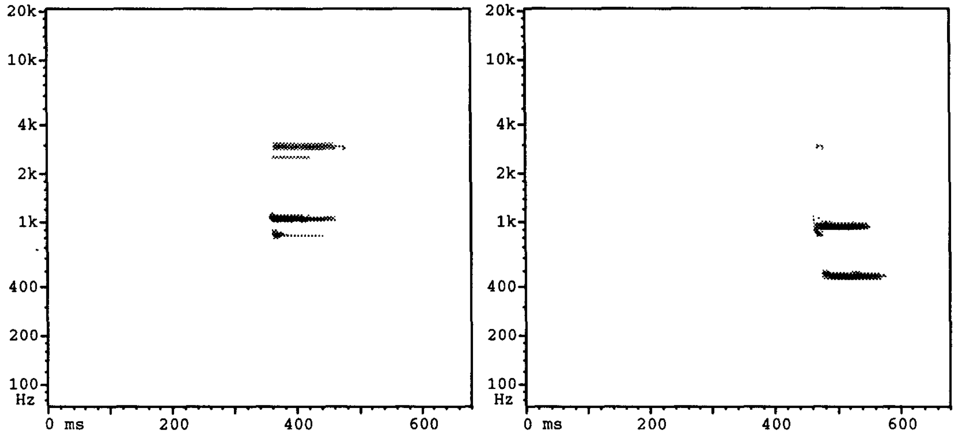


Figure 5.6: Events 5 and 6 from the Frescobaldi toccata.

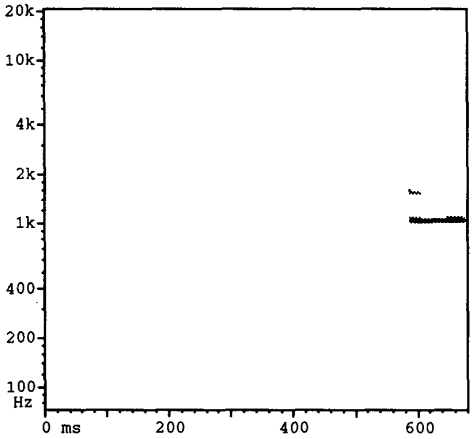


Figure 5.7: Event 7 from the Frescobaldi toccata.

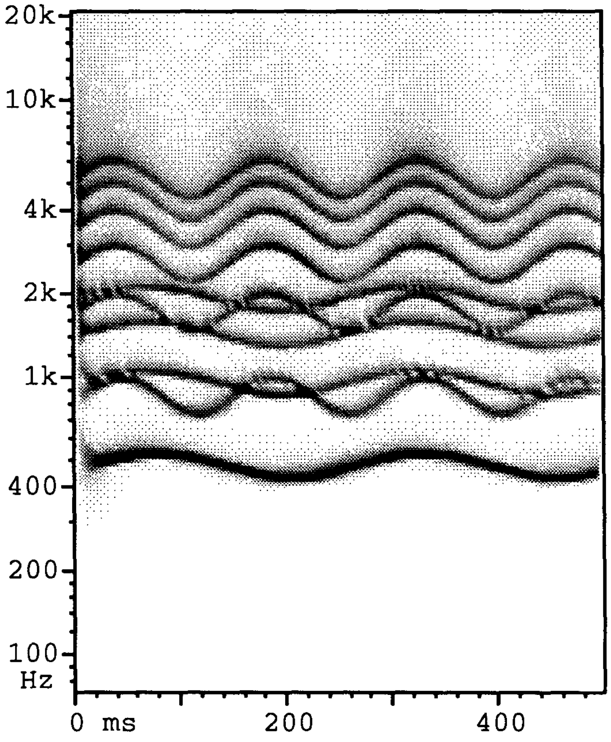


Figure 5.8: Mixture of two notes with separate vibrato.

A second sound is the mixture of two notes which have frequencies of 500 and 900 Hz. Each note has sinusoidal frequency modulation at a different rate and depth. In this sound, which is reproduced in fig. 5.8, the two events heard when you listen to the sound are the two notes. These events are difficult for the program to separate correctly for several reasons. They both start at the same time, leaving no common onset cue to distinguish them. At onset, the partials are muddled, not becoming clear for 60 or 70 ms. The partials cross one another repeatedly and sometimes touch without crossing, making partial tracking difficult. There are numerous small features, such as the destructive interference of partials pointed out in fig. 5.1, that also make tracking difficult.

The feature maps resulting from running the algorithm on this sound are shown

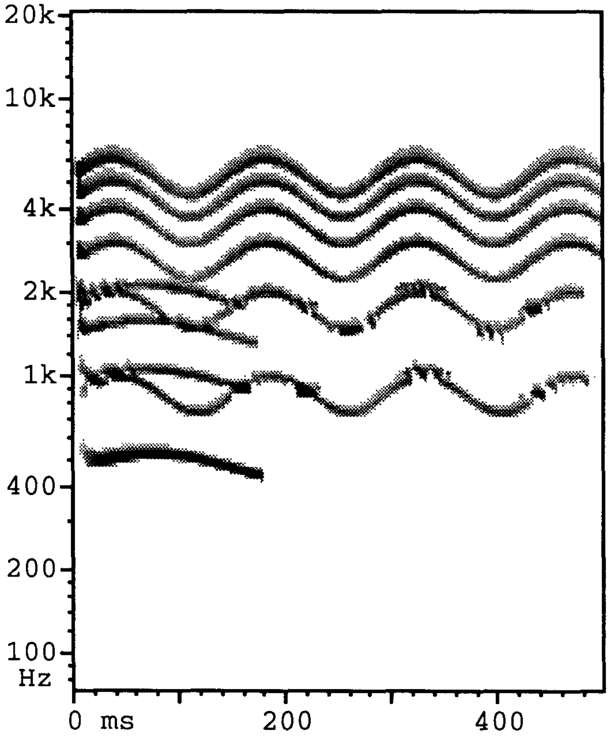


Figure 5.9: Event 1 of the two-note mixture.

in figs. 5.9 and 5.10. At the beginning of the sound, all of the partials are grouped together because they have a common onset. After about 180 ms, enough FV information has been collected — enough updating of the *aff* values based on FV differences — to pull apart the two notes, assigning them to separate events. Both events share the same history up to the break point, as explained in section 5.3.3, because there is no reason to distinguish one event from another at the time of the split.

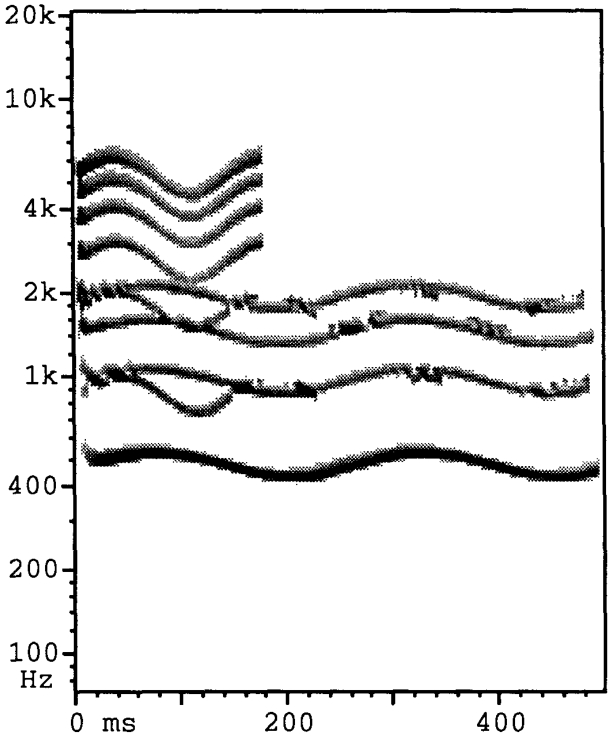


Figure 5.10: Event 2 of the two-note mixture.

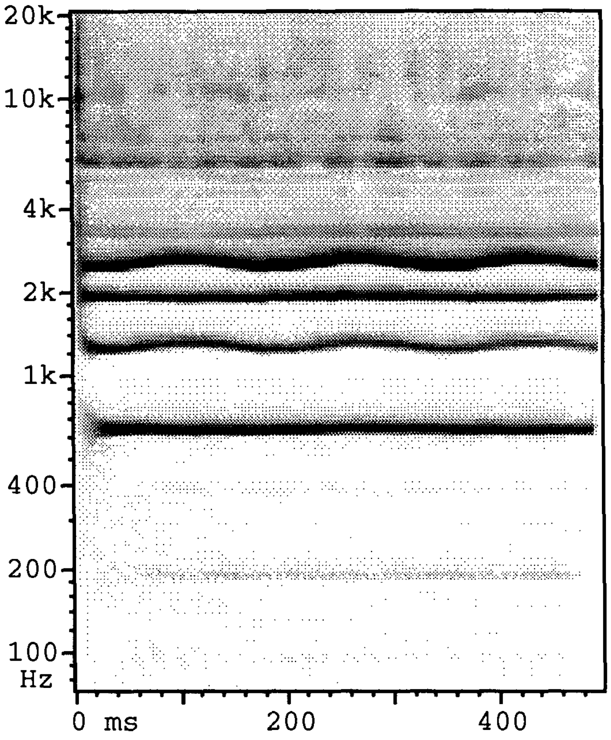


Figure 5.11: McAdams's oboe sound.

A third sound is the McAdams oboe sound of fig. 5.11. The goal here is to separate out the even and odd harmonics because of their independent FV. This sound is difficult because of the subtlety of FV in it, providing little for the event formation algorithm to work with. The output maps produced by the algorithm are shown in figs. 5.12 and 5.13. Again, at the onset of the sound, the even and odd harmonics are grouped together because of their common onset. After a little while, again, the two sets of harmonics split apart into two events, exactly as happens perceptually when listening to the sound.

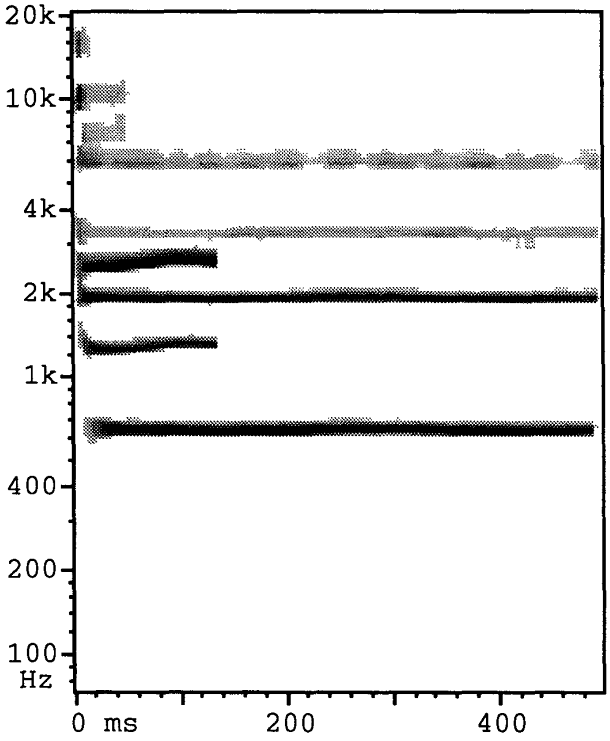


Figure 5.12: Event 1 of the McAdams oboe sound.

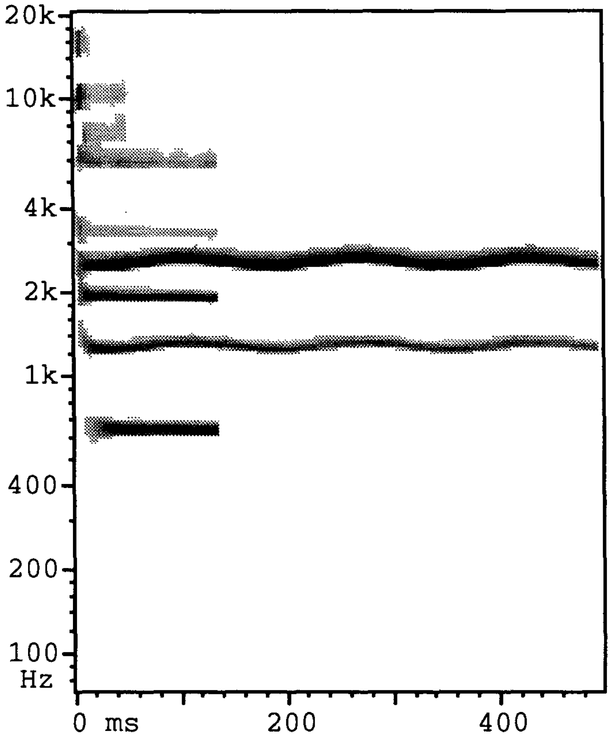


Figure 5.13: Event 2 of the McAdams oboe sound.

The event formation algorithm given here gets mixed results when more instruments are playing. Fig. 5.14 is a short excerpt out of Beethoven's octet in E-flat major, opus 103; the score is given below. This sound is interesting because it involves several instruments playing chords: first with different onsets for the different instruments, then (in the second measure, at about 1820 ms in the sound) with a common onset.

The algorithm produced the output maps shown in figs. 5.15 and 5.16. The event algorithm misses the first tone entirely because at no time is there a strong onset. The instrument here is a bassoon, whose tones have an attack in which the onsets of the different harmonics are spread out in time compared to the other instruments present, the oboe and clarinet [Grey75, p. 67]. It is probably this smearing of onsets that causes the bassoon not to trigger any strong response in the onset filter, making the instrument undetectable by the event-formation algorithm.

It is possible to tune the algorithm to be more sensitive to onsets, so that it picks out the initial bassoon tone. Unfortunately, it is then overly sensitive, enough to find many onsets in the mix of partials between 950 and 1800 ms. In this case the chording of the notes is probably a factor: harmonics of different instruments fall into the same cochleagram frequency channel, leading to beats in the channel's amplitude. These beats cause a moderately strong response from the onset filter; it is this suprathreshold response that makes the event formation algorithm incorrectly detect new partials between 950 and 1800 ms.

The other notes are detected by the algorithm, though it fails to separate the two tones, clarinet and oboe, that begin nearly simultaneously at about 950 ms, and finds only two of the four tones that begin at 1820 ms — and only a few partials of one of the latter. In both of these cases, there does not seem to be enough onset asynchrony in the tones to separate them by onset alone; the small amount frequency jitter present did not enable separation.

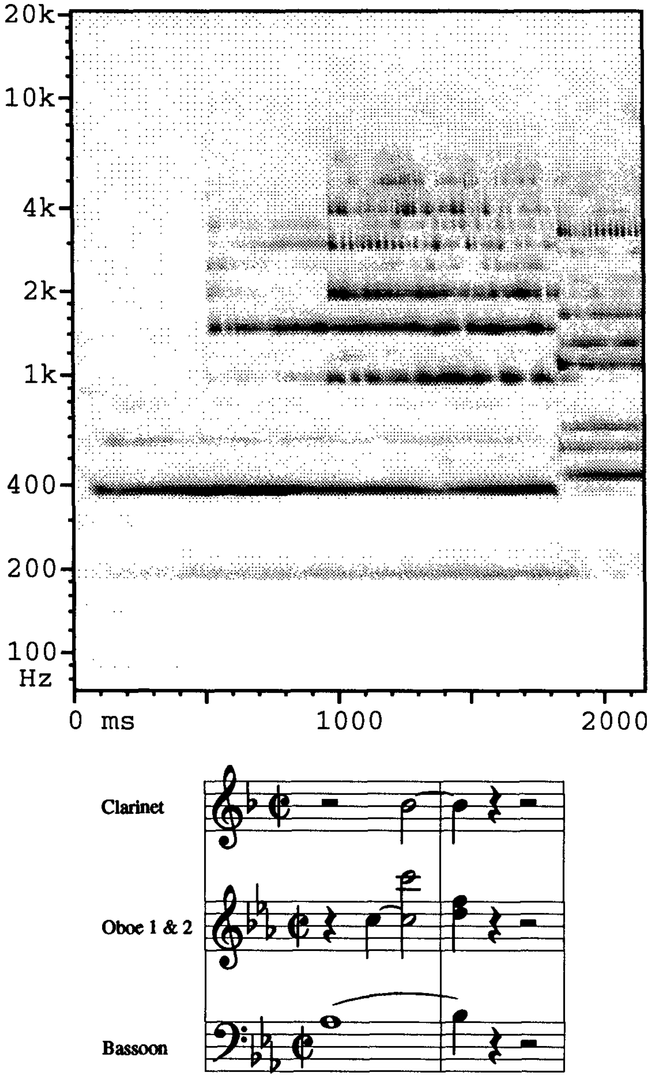


Figure 5.14: An excerpt (the first) from the Beethoven octet.

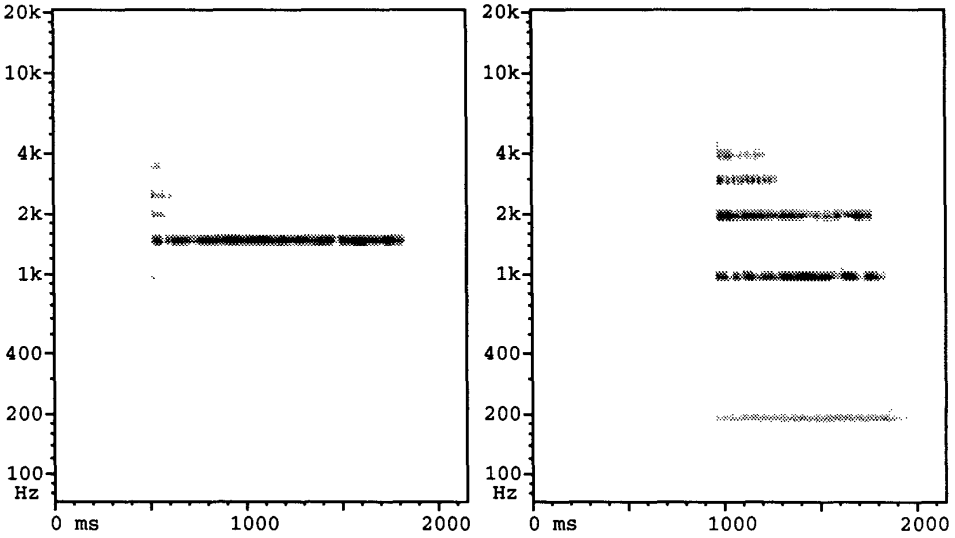


Figure 5.15: Events 1 and 2 from the Beethoven octet.

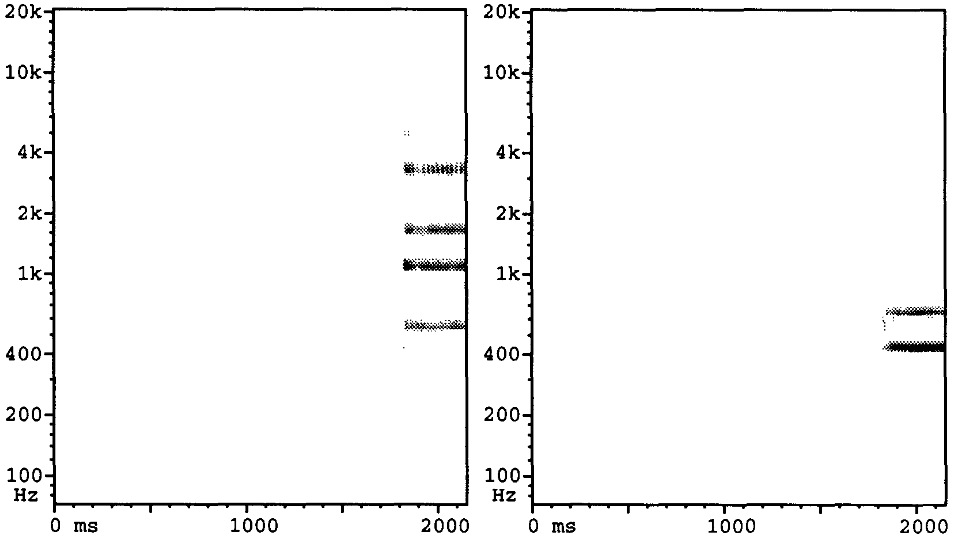


Figure 5.16: Events 3 and 4 from the Beethoven octet.

An excerpt from Beethoven's violin concerto in D-major, Opus 61, is shown in fig. 5.17, with the score below. The event formation algorithm produced the maps shown in figs. 5.18 and 5.19. The last three maps (in fig. 5.19, all but the top-left image) are composite images made by adding together several event maps. The groupings of partials in these three maps that are distinctly separate in time are different events, and are drawn together here only to save space.

The algorithm was not able to place into one event the partials that belong to the violin, despite the fairly strong vibrato being played. The event formation algorithm, despite having some effort applied to tune it to be sensitive to this vibrato, was not able to correctly match the partials. Part of the reason for this is the amplitude modulation that happens simultaneously with the vibrato. The partials drop in amplitude at some points to less than one tenth of their intensity at surrounding points; bear in mind that this signal has been compressed by the 4-stage AGC in the cochlear model, so the dynamic range in the original sound signal is much greater than 10 dB. Part of this amplitude variation may be due to partials moving closer and farther from the frequency of a resonance of the violin body, but it is sufficiently extreme that it seems some other process must be operating as well. The algorithm is able to track partials through these gaps by virtue of continuity constraints, but apparently the on-and-off nature of the partials prevents the algorithm from using FV feature maps to merge events that contain different partials from the same instrument.

One might think that a common AM filter would be useful here, but close inspection reveals that many of the partials are not modulating in the same phase: When one partial has an amplitude dip, another partial will be strong, and conversely. The partials at about 2600 and 3900 Hz, between 2200 and 2700 ms are a case in point.

The algorithm tracked only one partial from all the sound made by the orchestra. The orchestra is much quieter than the violin in this passage, but still at least one other frequency component (at 440 Hz), albeit intermittent, is apparent to the human eye. One would expect the partial tracker to detect it, but it does not; perhaps because of the very slow onsets of this partial, it is never detected at all.

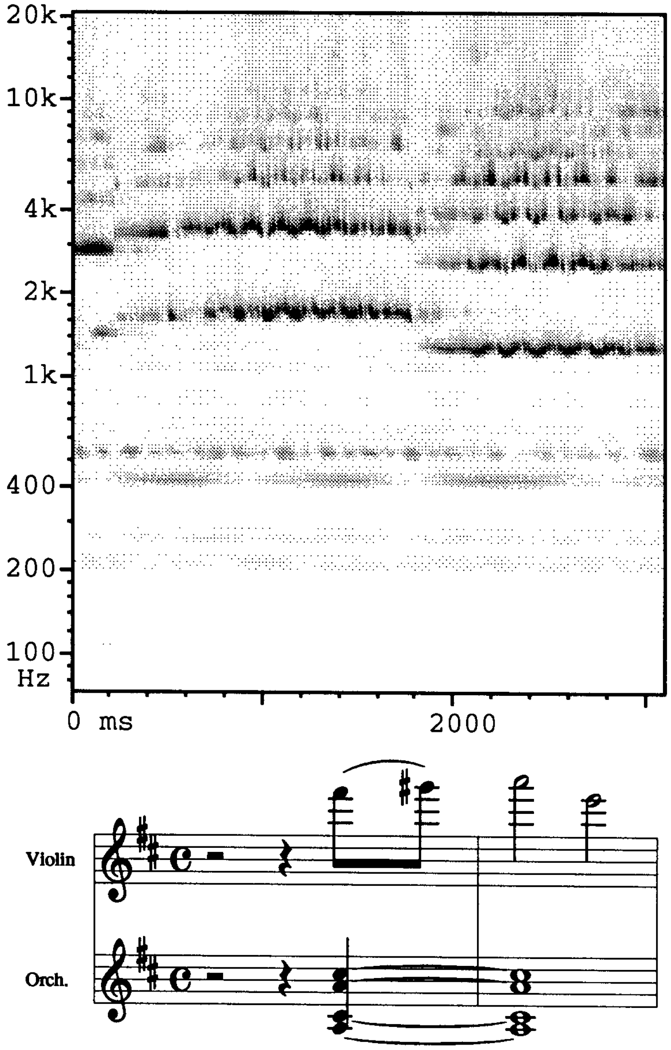


Figure 5.17: Excerpt from Beethoven's D-major violin concerto, with score.

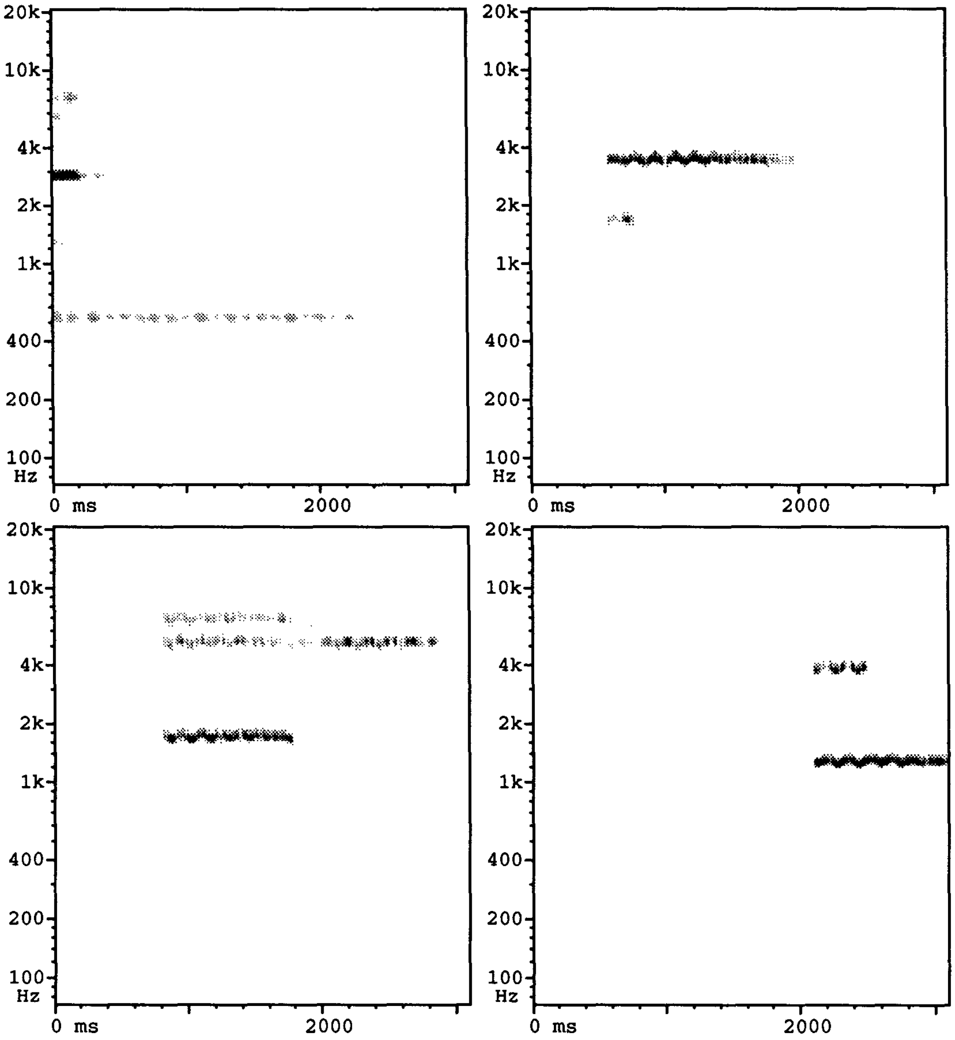


Figure 5.18: Four events from the Beethoven violin concerto.

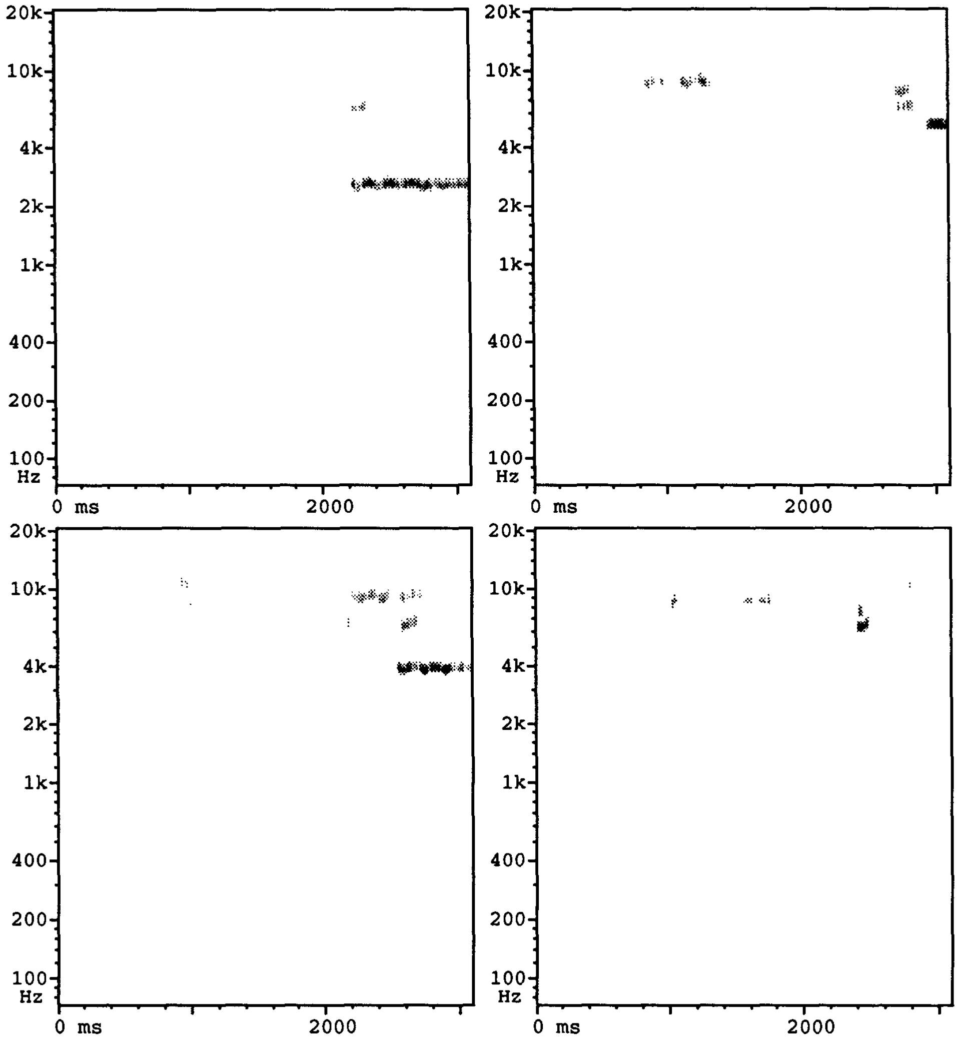


Figure 5.19: The remaining events from the Beethoven violin concerto.

A final sound is shown in fig. 5.20, along with the score. This is a second excerpt from the Beethoven octet in E-flat major. The result maps were made without special fine-tuning of the event-formation algorithm; the threshold values and other tuning parameters were the left the same as for the other octet example of fig. 5.14. The output maps are in figs. 5.21 and 5.22.

As you can see, the algorithm did not do very well in this case. It found a few of the tones present in the piece, but also placed into the same event several tones that should have been separated and missed some partials entirely. As in the previous example, there are several events that consist of nothing more than a brief onset burst. As a fully automatic event separator mimicking the human auditory system, this algorithm leaves quite a bit to be desired.

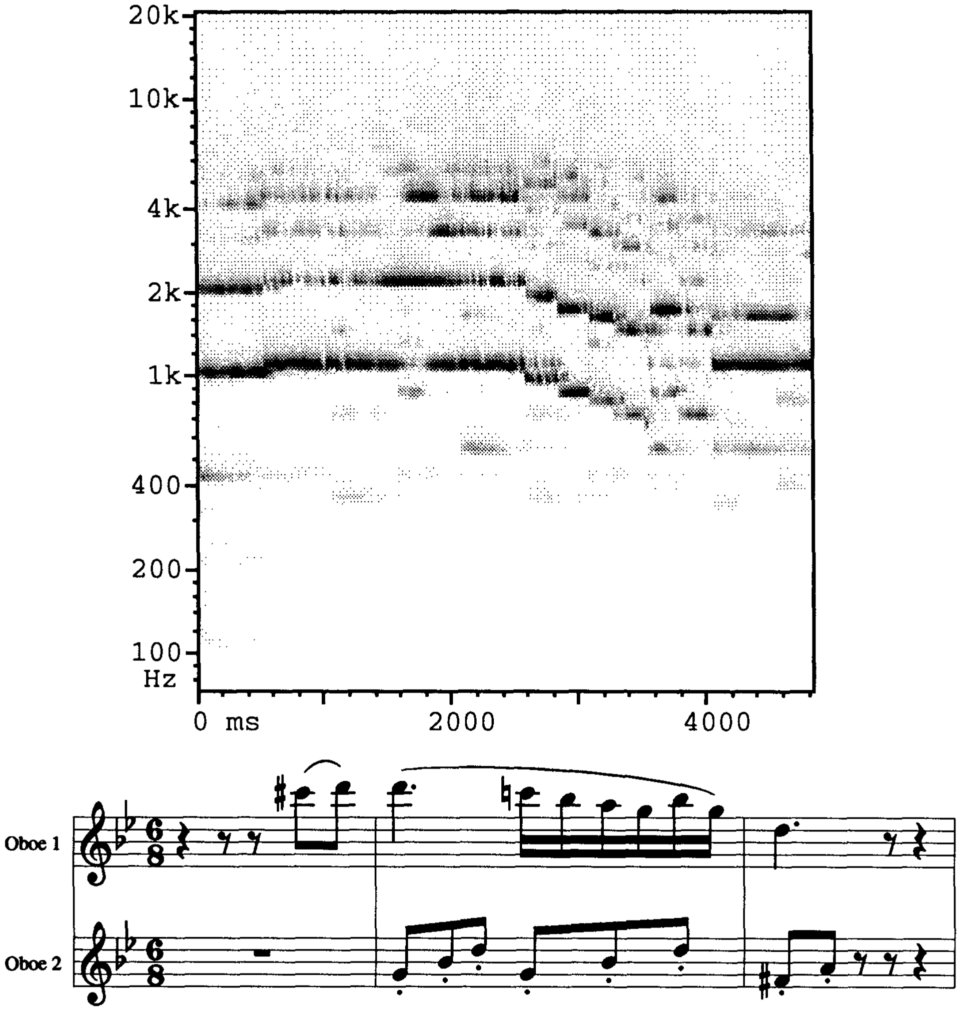


Figure 5.20: A second excerpt from the Beethoven octet, with score.

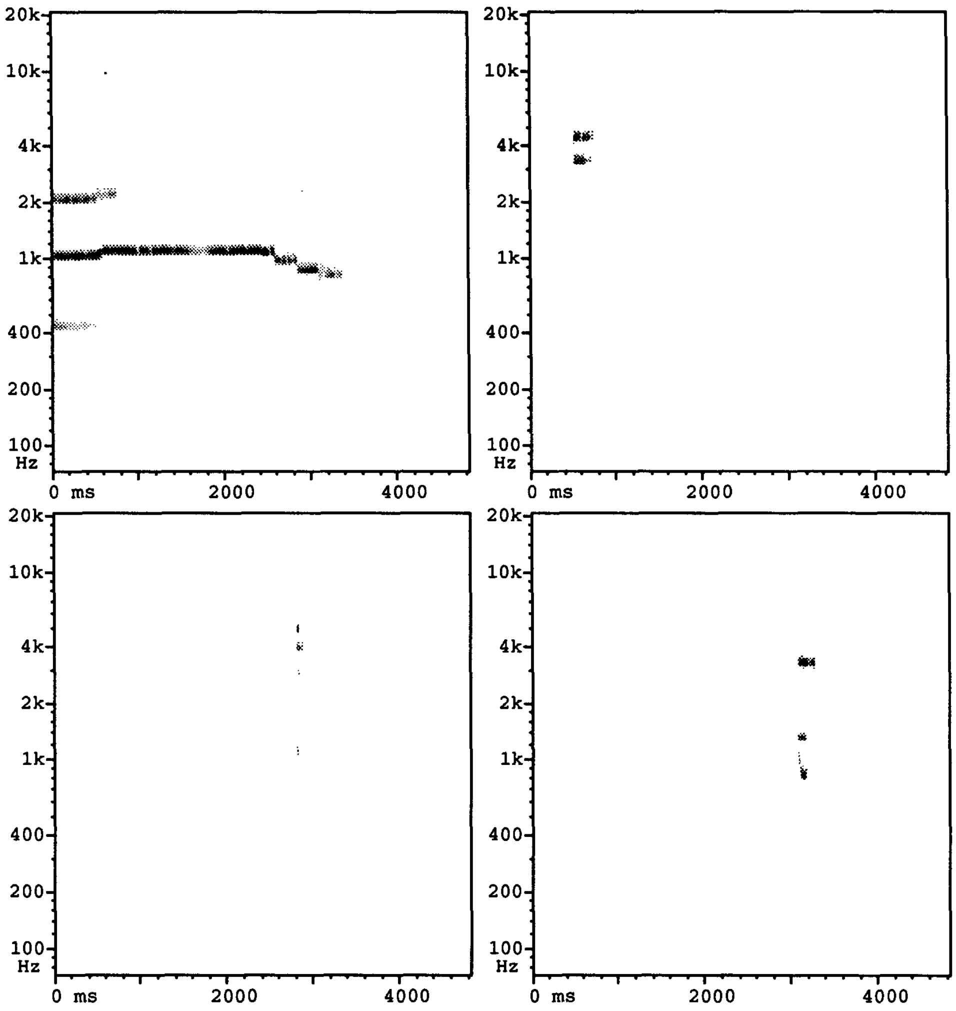


Figure 5.21: Four events from the Beethoven octet, excerpt 2.

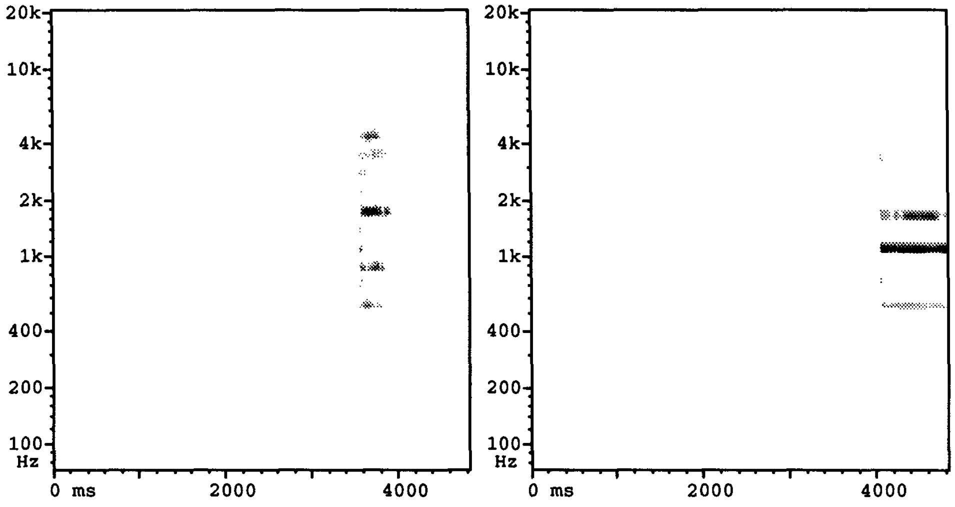


Figure 5.22: The remaining events from the Beethoven octet, excerpt 2.

5.6 Limitations of the Model

The model presented here fails to work well in some musically significant situations — namely those in which the onset and FV cues used by the algorithm are weak or misleading, as in the examples in the previous section.

One example of this is when different instruments, or different tones of a multi-voiced instrument like the piano, are played simultaneously, or nearly so. This can happen in playing a chord. There may be an onset cue at the start of the chord as the voices commence at slightly different times, but because of the absence of continuing cues, the model will soon group all of the partials together. (This did not happen in the Frescobaldi excerpt because of the rapidity of the playing.) This problem could be cured, at least partially, with a harmonicity cue detector.

Another instance of the model's failure occurs when too few harmonics are resolved by the cochlear filter. This happened partially in figs. 5.4 and 5.6, where some harmonics have vanished because other notes' harmonics have interfered. When such interference happens, the amplitude of a cochlear channel drops and the partial tracker usually does not detect a partial. This problem gets especially severe when there are many sources, as in orchestral or choral music.

Large intensities of noise can also make the model perform poorly. String instrument tones, as seen in fig. 3.19 (p. 69), are relatively noisy; a rapid succession of such tones, as in a string quartet (or larger group), is very difficult for the model to separate, or even to filter out reasonable onsets.

5.7 Neural Speculation

The algorithm presented above made no attempt to model in detail the action of the auditory system, though it did incorporate some of the principles found by psychoacousticians which are somehow used in real-world hearing processes. This section speculates on how the auditory system may represent the knowledge that different parts of a sound signal belong to different sources.

5.7.1 Labelling and Filtering

Two different ways that sources may be separated in the auditory system spring to mind. These may be called the *filtering* and *labelling* hypotheses. The filtering hypothesis says that different sources are separated into different neural channels, so that each source gets it own set of neurons for propagating information to the brain. This approach is the one used in the result maps of the event-formation algorithm, shown in section 5.5. The labelling hypothesis says that the same set of neurons transmits information about different sources to the brain, but that this information is somehow labelled to identify which parts of it belong to which sources. A way of displaying labelled events for visual inspection suggested by Mont-Reynaud [Mont-Reynaud90] is to use a different color for each event or source, so that they can all be displayed on a common feature map.

Both filtering and labelling have obvious strengths and weaknesses. It is easy to see how filtered events could be transmitted neurally, while a method by which a neural signal could labelling a set of firings as being distinct from one another seems difficult to imagine. On the other hand, the auditory system extracts many different kinds of patterns from a sound — including at least the properties covered in chapters 2 and 4 — and replicating the data-transmission circuits in the auditory system once for each possible source would seem to be an immense neural burden.

5.7.2 Cortical Oscillation

A potential answer to this dilemma has appeared within the last five years. If this theory is applicable, it would imply that the labelling hypothesis is correct.

The basic idea is that *synchronized neural firing* is the label by which different features of a sound signal are identified as belonging to the same source. This representation, somewhat like time-division multiplexing, has different sources sharing the same neurons to compute and transmit information, but using these neurons at different times to do so. Since neurons typically fire repeatedly at a rate of up to several hundred spikes per second, different sources could be represented by having each one claim a distinct phase and/or frequency of repetitious firing as its mode of identification.

Though questions abound — the most prominent asking how different neural processes which perform parts of the source formation process, like feature detection, agree on the phase and/or frequency to use — evidence for such a mechanism in other senses continues to grow, and it is possible to imagine the mechanism applying to audition as well.

Theoretical Work

Theoretical underpinnings of the cortical oscillation theory were advanced by von der Malsburg and Schneider [Malsburg86b], who suggested that synchronous oscillation (firing) of cortical neurons could bind together sense impressions in vision and perhaps other senses as well. They refer back to the idea of Hebb [Hebb49], who suggested that *neural assemblies* were how features were organized and processed [Malsburg86a]. Freeman had long recorded synchronicity of firings in olfactory neurons [Freeman75] and suggested that it might be binding together separate features in that sensory pathway.

Experimental Evidence

Singer and others, while probing neurons in the visual system, noticed that units fairly far apart from each other were firing in synchrony at a rate of roughly 40 Hz [Barinaga90]. Gray and Singer [Gray89b] then recorded synchronous firing in different columns of the visual cortex in cats in response to certain stimuli. Gray *et al.* [Gray89a] found the same result for properties of the entire visual field, leading them to conclude,

> [T]he synchronization of oscillatory responses of spatially distributed, feature-selective cells might be a way to establish relations between features in different parts of the visual field.

One factor complicating the understanding of auditory synchronization that is not present in vision is the presence of phase locking of neural firing to the waveform. It may be difficult to tell whether a given example of synchrony is due to feature binding or merely to phase locking. Indeed it may be that phase locking acts as a trigger for feature binding — perhaps this is why harmonic sounds hang together so well. Certainly common onset of energy at different parts of the spectrum, triggering as it does a cascade of neural firing in a short time, could initiate synchronicity in the firing of feature filters. Jenison *et al.* point out the extent of synchrony of auditory nerve fibers. In the absence of competing periodicities, a speech formant at one frequency triggers synchrony across fibers of a wide range of frequencies [Jenison91].

As far as I can determine, the only auditory work that has been done in this area is on the grassfrog. Johannesma *et al.* [Johannesma86] recorded up to four neurons simultaneously, finding that their firing became synchronous with certain stimuli (but not at a rate that implied mere phase-locking to the waveform). They suggest a search for neural assemblies in the auditory system, ones that would "contribute to the generation of figure out of ground, percept out of stimulus, [or] sense out of fact."

These questions and more will have to be answered by neurophysiologists.

Chapter 6

Summary and Conclusions

6.1 Summary of the Model

Sound enters the ear and is transduced, after some gain control processes, into neural firing representing sound energy at each frequency. This representation, in time and height, is the first of several feature maps, arrays in some number of dimensions that represent the presence or intensity of some feature. The map of cochlear neuron firings is the input to a number of elementary processes that filter features in the sound, including onset, frequency variation, harmonicity, amplitude variation, and spatial location. These features, in turn, drive an event formation process that integrates information from the feature filter output maps and the cochlear nerve firings to decide over time which parts of the spectrum are associated with which events. Event formation influences and is influenced by another process, source formation, that places events into separate sources such that all events that are in an auditory stream originate from the same source. Information about what events are in what sources is passed to higher levels of the auditory system, including the brain.

6.2 Questions Revisited

Chapter 1 presented a number of questions which this thesis aimed to address. These questions can now be answered.

- How well does such an implementation work?

  It works reasonably well on the sounds tested there, identifying the events and capturing most of the partials that belong to them. It fails most often on sounds with coincident vibratoless harmonics and sounds where onsets are not prominent.

- What are some of the principles that influence the grouping of events into sources?

  In addition to those that affect event formation, the factors for source formation include pitch separation, timbre differences, repetition rate, repetition number, overall loudness, and loudness differences.

6.3 Contributions

The contributions of this thesis to the field of auditory scene analysis are these:

- The structure of the auditory model outlined above; this provides an overall framework in which computational experiments may be done.
- A survey of the psychoacoustic and neurophysiological literature for information about how humans and other animals may respond to features and use them for source separation.
- A representation for features that may work for several levels of the auditory model and for other tasks like pattern recognition.
- A method for onset filtering and two methods for frequency variation filtering, along with knowledge about how to tune them for musical sounds.
- An algorithm that uses some of the principles of scene analysis presumably present in the auditory system to integrate information from cochlear, onset, and frequency-variation data to make event decisions.

6.4 Comparison with Other Models

6.4.1 Weintraub

Weintraub [Weintraub85] sought to do source separation by modelling the auditory system. He identified many of the principles covered here, including the importance of using a variety of cues — harmonicity, common FV, common onset and offset, common AM, lateralization, and higher-level context cues. He also talked about some higher-level grouping principles, such as continuity over time of spectral components. The system was aimed at separating two simultaneous talkers.

His implementation shares some characteristics with mine. The analysis began with Lyon's cochlear model and used the per-channel autocorrelation function. He used the autocorrelation function to estimate the pitch period of the signal, effectively employing a harmonicity cue for spectral grouping. This estimation was done in a process similar to summing vertical strips in a correlogram, then the using peaks in the resulting function as periodicity values. These pitch estimates were used as input to a dynamic programming algorithm to form pitch tracks over time. This algorithm incorporated continuity constraints for more accurate tracking of pitches, just as mine does for tracking of partials. Weintraub also used a Markov state machine to model the onset, continuation, and offset of pitched sounds. This process has no analog in my model, nor did his next steps, an iterative probabilistic process to assign estimated amplitudes of spectral components to each source, and resynthesis of a sound signal. He tested his system with speech sounds; musical sounds were used in my system. He mixed two voices uttering digits, separated the voices with his system, and fed the output to a speech recognizer; the results were that his process helped fairly well to improve recognition accuracy of separated male speakers but not of female ones.

Weintraub's model has an advantage over the current one in using harmonicity for source separation. This cue is vitally important in musical sound, since most music consists of pitched notes. However, his work does not attempt to model the auditory system, and it is not clear how additional cues such as common FV and common AM could be used in his system.

6.4.2 Cooke

Cooke [Cooke91] presented a model of source separation sufficiently similar to mine that it is worthwhile to delimit the differences. Cooke's model, like Weintraub's, aimed at separation of speech from interfering sounds, in contrast with the musically-oriented model presented here.

Cochlear Model

Cooke's cochlear model starts out with a bank of gammatone filters [Boer78] which operate somewhat similarly to the filterbank in Lyon's cochlear model. As in Lyon's model, this is followed by a non-linearity, though in Cooke's case it is a sigmoid instead of the halfwave-rectification used by Lyon. Cooke also models the functioning of the hair cell with a leaky integrator/reservoir and a firing threshold; this is not done in the Lyon model. The latter does, however, contain multi-stage coupled automatic gain control not in Cooke's model. The AGC helps bring out weaker harmonics which may be missed by models that lack AGC; in addition, its compression greatly reduces the dynamic gain of the signal, leading to an output that is more neurally reasonable.

Partial and Formant Tracking

The next major part of Cooke's model computes synchrony strands, which correspond to both partials (of low-numbered harmonics) and speech formants (at higher harmonic numbers where several harmonics fall into the same frequency channel). This contrasts with my model, in which features are filtered out before any information about partials is known.

Cooke's model first estimates the dominant frequency in each channel. These estimates are then aggregated into *place groups* that model areas of the basilar membrane vibrating at nearly equal frequencies. This recruitment of groups of adjacent channels can also be seen in correlogram images, where each spot corresponding to a single strong frequency extends vertically over many frequency channels. Place groups are extended over time to form *synchrony strands* that correspond to partials and formants. This technique is similar to my technique for tracking partials except that (1) it also captures formants, which are useful to detect in the speech domain Cooke is working in, and (2) it happens much earlier in the processing than my algorithm for tracking partials.

Scene Analysis Theory

Cooke next discusses some of the principles and heuristics used in auditory scene analysis, such as exclusive allocation and competition between different groupings. Like me, he sees that the problem at its most basic level is identifying which partials or, more generally, which synchrony strands belong in which group over time.

Scene Analysis Implementation

It is only at this level, after synchrony strands have been computed, that Cooke's model begins to filter out what I call features. In contrast, my system filters information about these features first, then uses the information to identify and track partials. Accordingly, Cooks looks for features only at those frequencies through which the synchrony strands pass. Cooke's first cue is harmonicity, which he calculates by a harmonic sieve that allows, but places a slight penalty on, deviations from perfect harmonic alignment. The other main cue is common amplitude modulation, calculated via measuring the difference in instantaneous AM rates between pairs of synchrony strands. These two cues are used to place synchrony strands into groups — either harmonicity groups or AM groups. These groups can potentially be merged by a higher-level process that considers all of the partials in all of the groups and searches for the best grouping(s), using also pitch contour similarity as a cue. That is, the strands in each group are compared for similarity of pitch track. If the similarity score is high enough, the groups are merged.

This summary reveals several other important differences between Cooke's work and mine. Cooke's model uses harmonicity and AM as the main separation cues, while mine uses common onset and common FV. Cooke's pitch contour similarity is similar to common FV, but it plays a very limited rôle in his model, operating only after harmonicity and AM groups have been identified. Cooke, like me, uses principles discovered by psychoacoustic experiments to justify the choice of cues and grouping algorithm. His partial grouping (strand grouping) process is also motivated by psychoacoustically-discovered principles, though perhaps not to as great an extent as mine. He does not consider neurophysiological data. Like my model, his does no source formation — thus his speech examples are restricted to continually-voiced utterances such as "I'll willingly marry Marilyn."

Cooke's model has certain advantages and disadvantages relative to mine in separating concurrent musical tones from one another. The cues used in his model, especially harmonicity, are important musically; so are mine, and any complete system would need filters for all of these cues. My model probably has an advantage in handling those sounds, such as drum beats, which have little periodicity, while Cooke's would suffer because it looks for source-formation cues only after finding synchrony strands. Cooke's model seems to be better able to track weak harmonics; this gives it an advantage for harmonically rich musical tones – *i.e.*, most of them. Because of its use of psychoacoustic constraints, my model is probably slightly better than Cooke's at tracking partials through interfering noise, such as would happen with loud drum beats.

To sum up, Cooke's model is similar in overall form to the present one but differs in many of the details. His model tracks partials and formants first, then looks for features in the strand tracks; in contrast, mine filters the features first and uses them to track partials. His main cues are harmonicity and common AM, while mine are common onset and common FV. My partial grouping algorithm is perhaps more sophisticated and better motivated by psychoacoustics than his, but his method of speech-based method of evaluation is more extensive than my musical one.

6.5 Future Work

We are still quite a long way from the level of performance of human auditory perception. We can understand voices even in a room crowded with other talkers; no current source separation system comes even close to achieving this level of performance at the separation task. We don't yet know even what all of the cues are that might be used for source separation, much less how these cues might be found in a sound signal and filtered out. What is the rôle of pattern recognition — for example, of phoneme or word recognition, of musical instrument timbre recognition, or of recognition of footsteps and telephone bells? How does the auditory system deconvolve source characteristics and the effects of room acoustics? How do we filter out noise? How do short- and long-term acoustic memory affect source separation? How do linguistic, semantic, and cross-modal clues come into play?

Short of these longer-term research issues, there are quite a few relatively straightforward possibilities for work to follow this effort.

One possibility is to improve the feature filters. The onset filter could be expanded to capture slower amplitude variation as well as onsets. Most likely the simple decaying-exponential spike used here would need some parameter of variation. Some kind of non-linearity would likely help with this and possibly with onset filtering as well. The frequency variation filter could perhaps be improved by non-linearities — possibly lateral inhibition — and expanded to cover wider-band partials than the musical ones concentrated on here.

Other feature filters need to be added to the implementation. The issues of pitch extraction, its relationship to harmonicity and/or periodicity, and its utility for source separation are still unclear; psychoacoustic, neurophysiological, and computer modelling experiments could shed some light. Harmonicity is probably one of the most important cues for source separation; Cooke's amplitude modulation techniques may be important here, as well as his method of filtering and representing harmonic structure. Location has been neglected in the field of auditory modelling for scene analysis: no one has yet tried to incorporate localization cues into a sophisticated multi-cue separation system.

Another open question for computer modellers is how to improve the partial tracking and event formation algorithm of chapter 5. As noted there, that algorithm incorporates only a few of the principles and factors that affect event formation — enough to show the utility of the implemented feature filters. A more complete algorithm would use all of them and probably other as-yet-undiscovered cues as well. In addition, that algorithm does no source formation. Thus it cannot, for instance, decide which notes in ensemble music belong to which instruments; it can at best notice that single notes exist and determine their spectral content over time. A complete source-formation system would use the cues and principles outlined in chapter 4 and in Bregman's book to track sources over time.

Another direction of work would be to investigate types of sounds other than music. Speech and environmental sounds each have their constraints and idiosyncrasies that require attention before applying a source-separation system.

For practical applications of source separation, a re-synthesis system could be built that takes the output of the event-formation algorithm and regenerates sound. Such a system would need to use a modified front end, either an ear model that records the amounts of gain applied so it could be reversed on output, or a linear front-end like the Constant-Q Fourier Transform [Schwede83]. Since the output of the event formation algorithm is essentially binary — what parts of the spectrum are and are not included in each event — the linearity of stages past the front end need not matter; the initial time-frequency representation is sufficient to regenerate the sound. Most likely a sinusoidal partial tracker [Serra88] would improve reconstruction quality.

Another area of work is to examine other auditory processes than source separation in the framework described in the auditory model presented here. For instance, the features filtered here could be used as input to a pattern recognizer, one that identifies time-frequency patterns on the basis of what features are present in what frequency and time relationships.

A final interesting line of work would be to follow up on the cortical oscillation theory of section 5.7.2. Could one develop a physiologically compatible algorithm for event formation and source formation incorporating synchronous neural firing as an object representation? How would it work?

6.6 Conclusion

This thesis presents some auditory-system evidence, a model, and some implementational techniques for scene analysis, and suggests several directions for further work. These lines of development increase, and will continue increasing, our knowledge of the auditory system and its source-separation methods. In addition, we may soon have a fairly complete source-separation system that can be used for practical applications, including speech separation from interfering a mixture, detection of sounds in some kinds of noise, and music transcription. Such a system will be able to listen to that jazz band and pick out the strokes of the drums, write down the bass lines, and separate out that solo that Miles Davis is playing.

Appendix A

Growable Triangular Matrices

This appendix describes a simple technique for storing and acessing a growable symmetric table-lookup function, such as the *aff* function of chapter 5, in computer memory.

Requirements of the Storage Method

Given two partials p_1 and p_2, the *aff* function returns their affinity value, a real number. *aff* is a table-lookup functions, in that a value is first stored for $\mathit{aff}(p_1, p_2)$, then later retrieved or possibly changed. Since each partial has an integer index, what is needed is just a function

$$\mathit{aff}\colon Z^+ \times Z^+ \rightarrow \Re$$

aff is symmetric in its arguments, meaning that $\mathit{aff}(p_i, p_j) = \mathit{aff}(p_j, p_i)$ for all pairs i, j. Also, *aff* must be *growable*, in this sense: there can be arbitrarily many partials, so the table of *aff* values must be able to grow arbitrarily in size to accomodate new partials.

Implementation

The requirement of a symmetric table-lookup function suggests that a triangular matrix be used. Such an array stores the *aff* value for two partials p_i and p_j only once. It can be implemented by storing the elements of the matrix in the computer's memory in the order they appear in the matrix. Elements are stored in this pattern:

$$\begin{array}{llll}
(0,0) & & & \\
(1,0) & (1,1) & & \\
(2,0) & (2,1) & (2,2) & \\
(3,0) & (3,1) & (3,2) & (3,3) \\
\vdots & & & \quad \ddots
\end{array}$$

Here, the two numbers of each ordered pair are the partial index numbers, and the position of the pair in the triangular matrix shows where the *aff* value for that pair is stored.

If the matrix is stored in row-major order in computer memory, then the use of a *lower*-triangular matrix implies that the matrix need not be reshuffled when room for a new partial is needed. The matrix can just have a new row added to the end, so the new partial will have a position for an *aff* value with each existing partial. (If the storage system had been, say, a square matrix with duplicated values for (i,j) and (j,i), then the matrix would have to be reshuffled in memory whenever a new partial was added.)

To access the aff value for a partial pair (p_i, p_j), just find the values $m = \min(i,j)$ and $M = \max(i,j)$. Then the position of the *aff*(p_i, p_j) is given by

$$m + \frac{M(M+1)}{2}.$$

This formula just adds m to the M'th triangular number.

If there were to be a large number of possible partials, then it might be worth making each row of the matrix be separately allocated, with a one-dimensional array of pointers to each row. The access method would then be simply to find where in memory the M'th row is stored and extract the m'th element from it.

Appendix B

Parameters

This appendix lists the parameters that control the operation of the event-formation mechanism discussed in chapter 5.

`checkThresh` : Minimum cochleagram value needed, in a channel adjacent to a partial peak, to copy the value to the output map showing an event. See also `loSide` and `hiSide`.

`earRecent` : Maximum amount of time a partial's neural firing rate can remain below `earThresh` before it is terminated.

`earThresh` : Minimum neural firing rate necessary to sustain a partial. See also `earRecent`.

`eDivThresh` : Minimum average affinity value between a partial and the other partials in its affinity group necessary to keep the partial in the group. Below this value, the partial splits off to form another event.

`fvDecr` : Value that the *aff* function between two partials is decremented by when two partials have the different FV. See also `fvIncr` and `fvDiffThresh`.

`fvDiffThresh` : Maximum difference between FV values of two partials for the partials to be considered to have the same FV. Only applies if the `fvSumThresh` threshold is exceeded. See also `fvIncr` and `fvDecr`.

`fvIncr` : Value that the *aff* function between two partials is incremented by when two partials have the same FV. See also `fvDecr` and `fvDiffThresh`.

`fvSumThresh` : Minimum sum of values in a FV feature map at one time slice needed to trigger FV processing.

`hiSide` : Number of frequency channels above a partial peak that are copied to the output map showing an event. See also `loSide` and `checkThresh`.

`loSide` : Number of frequency channels below a partial peak that are copied to the output map showing an event. See also `hiSide` and `checkThresh`.

`maxGroupAge` : Maximum time that a floating partial in a neighbor group can exist before being forced into an event or terminated.

`maxNewAge` : Maximum time that a newly-created floating partial can exist before being forced into an event.

`minMeanAff` : Minimum average affinity between a partial and the partials in an affinity group necessary for the partial to join the group.

`onConsec` : Amount of time the sum of values of the onset feature maps must stay above onThresh for it to be noticed. See also `onThresh`.

`onFvHiSize` : Number of frequency channels above a given channel that are inhibited from having new partials created when FV is detected in the given channel. See also `onFvStop`.

`onFvLoSize` : Number of frequency channels below a given channel that are inhibited from having new partials created when there is an onset in the given channel. See also `onFvStop`.

`onFvLook` : Amount of time in which new partials are inhibited from being created after FV is detected. See also `onFvStop`.

`onFvStop` : Minimum threshold of FV to trigger inhibition of creation of new partials. See also `onFvLook`, `onFvLoSize`, and `onFvHiSize`.

`onLowThresh` : Minimum value at one frequency in an onset feature map needed for that frequency to start a new partial. This threshold is used once the requirements set by `onThresh` and `onConsec` have been satisfied.

`onParRecent` : Time within which two partials must be created for them to be placed in the same affinity group.

`onStartThresh` : Threshold below which values in an onset map must fall to determine the end of an onset. (An onset has a "duration" because values in an onset feature map stay above this threshold for some period of time.) See also `onTime`.

`onThresh` : Minimum sum of values of the onset feature-maps needed for an onset to be noticed. See also `onConsec`.

`onTime` : Minimum time allowed between the end of onset of one partial and the start of another onset at the same frequency. See also `onStartThresh`.

`splitNear` : Maximum distance, in frequency channels, that a new spectral peak must be from an existing partial to be considered to have split off from the partial (and consequently join a neighbor group with the partial). See also `valleyDepth`.

`valleyDepth` : Minimum value in the cochleagram, as a fraction of the partial height, that must exist between a partial and another spectral peak for the other peak to be noticed. See also `splitNear`.

Appendix C

SoundExplorer

This appendix describes the interactive system that allows one to run the filtering and event-formation processes described in chapters 3 and 5, and to view their output feature maps, change filtering parameters, and so on. It was presented at the 1991 International Computer Music Conference in Montreal [Mellinger91].

SoundExplorer: A Workbench for Investigating Source Separation

David K. Mellinger and Bernard M. Mont-Reynaud
Center for Computer Research in Music and Acoustics
Stanford University
Stanford, CA 94305-8180
davem@ccrma.Stanford.EDU, bmr@ccrma.Stanford.EDU

Abstract

SoundExplorer is a system for computing and displaying information for audio source separation. It provides an interactive workbench with which you can listen to a sound or view any of several representations of it, process the sound with filters for various source-separation tasks, and view the outputs of this filtering. SoundExplorer displays the variables that control the operation of these filters, allowing you to read and change them interactively to see what the effects are. By using a standard for storing computed values and parameters, parts of SoundExplorer are portable to different types of computers.

Introduction

Source separation in human hearing [Bregman90] is the process by which a mixture of sounds arriving at the ear is separated into its constituent components. Simultaneous, as contrasted with sequential, separation is the part of this process that operates at very short time intervals. Simultaneous separation works by extracting from the sound mixture several different types of sound cues that provide evidence about which parts of the sound spectrum come from one source and which from another.

SoundExplorer, working from input of a sampled, recorded sound, computes (using computational models of auditory processes) and displays several of the cues used for simultaneous source separation: harmonicity, common frequency change, common onset and offset, common amplitude modulation, and so on. It shows visual images of the computed results; it does not currently have a method for re-synthesizing sound from the separation images.

Feature Maps

Data in SoundExplorer is organized in *feature maps.* A feature map is a rectangular array of up to (currently) three dimensions. The elements of this array are real values (floating-point numbers) that represent intensity of one sort or another. These values are produced by the various computational filters in SoundExplorer. For example,

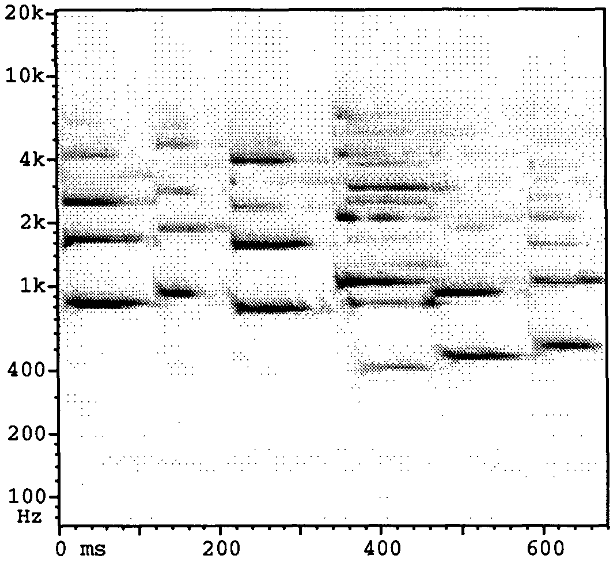


Figure C.1: A feature map for a segment of piano music.

one feature map for a snippet of piano music is displayed in fig. C.1. It is a two-dimensional intensity map with logarithm of frequency on the vertical axis and time on the horizontal. SoundExplorer is a system for computing, displaying, and storing feature maps.

The Computation Engine

The backbone of SoundExplorer is a set of programs for computing various types of feature maps for studying source separation. To study a sound in SoundExplorer, you must first obtain a sampled, recorded version of the sound in Next Soundfile format. SoundExplorer uses the recorded sound as input to a computational model of the cochlea [Lyon82, Slaney88], which produces the first two feature maps: a two-dimensional image (time × frequency) of cochlear neuron firing rates, and a three-dimensional image (lag × frequency × time) of the per-frequency-channel autocorrelation of the neural firings.

From these images, you can have SoundExplorer do any of a variety of things to process the above images to useful extract information. Each further processing step takes as input one or more of the existing feature maps and produces as output one or more feature maps. You run each filter by typing its name, any necessary parameters, and name of the sound. The program reads the sound file and any feature maps it needs, computes its output feature map(s), and writes the results to a new files.

Each type of filter in SoundExplorer extracts a different type of information useful for sound source separation. Details of these filters is beyond the scope of this paper. It is fairly easy to add new filters to SoundExplorer to extend its capabilities.

Running all of the filters in SoundExplorer takes up to 10 hours for one second of sound in the worst case. You can speed up this process considerably if other Unix computers are available on a local network. The slower filters are capable of spreading out their workload over locally networked machines. Since SoundExplorer's computation engine is portable to other kinds of computers, the others need only be connected by a network.

With all of the above options available, you easily forget which filters need which input parameters, which filters must be run before which others, and even the names of the filters. SoundExplorer has several solutions to this problem. First, each filter program knows the standard values for its parameters, so you need only provide the name of the sound to make the filter do the standard filtering. Second, a program named `crunch` runs any or all of the filters in order. For example, you can say "`crunch -on -ear miles`" to run the autocorrelation and onset filters in that order, as necessary, or "`crunch -all miles`" to run every existing filter. Finally, you can avoid all of these unmnemonic commands entirely by using SoundExplorer's interactive browser.

The Interactive Browser

Sitting down to use SoundExplorer, you see the interactive browser. This part of the system displays feature maps and allows you to examine and change them in various ways.

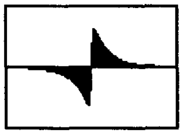


Figure C.2: One-dimensional filled-in map display.

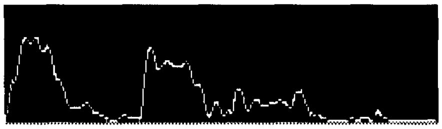
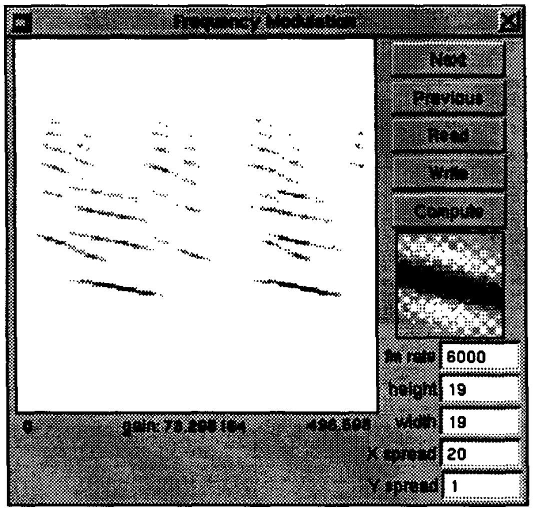


Figure C.3: One-dimensional line map display.

Feature maps come in one, two, and three dimensions. The browser displays one-dimensional maps in the ways shown in figs. C.2 and C.3, where the height of the lines intensity. Two-dimensional maps are shown as grey-level images, as seen in fig. C.1 above. In these images, the grey level shows intensity, with darker areas for higher (more intense) values. You can change the transfer function of intensity values to grey levels, enhancing low-contrast areas of the image.

Three-dimensional maps are displayed as two-dimensional animated images. Since all types of three-dimensional feature maps currently used have a time axis, it works quite naturally to display the time dimension of the feature map as the time dimension in animation. The remaining two dimensions map to the screen.

Each map displayed has a set of associated parameters. The set varies depending on the type of filter used to compute the map. The SoundExplorer browser, when displaying a map, allows you to see the type of map and its associated parameter values. These values are editable text, so you can simply point the mouse at one of them, type in a new value. Some types of maps also have additional graphical information that SoundExplorer displays with the map. For example, fig. C.4 shows the result of a convolution operator applied the cochlear model map for a sound, together with the kernel of this two-dimensional operator.

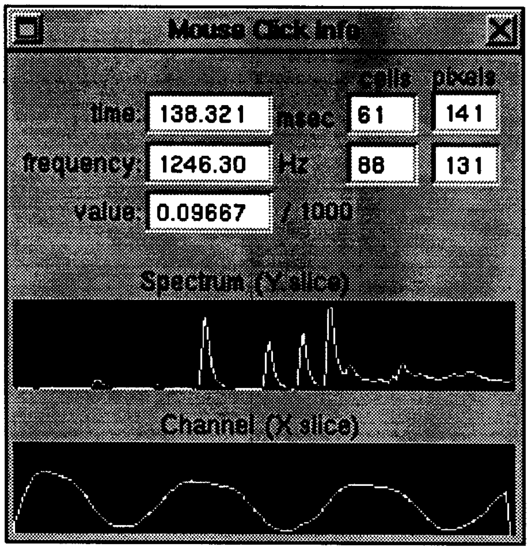


fm rate 6000
height 19
width 19
X spread 20
Y spread 1

Figure C.4: FM map with convolution kernel.

Each feature map has its own window on the Next machine screen. This lets you hide the maps you no longer need or re-display the ones you need again, and makes it easy to move them around independently. The latter feature has proven especially useful, since you can line up different images next to each other on the screen for comparison.

Another useful feature is the mouse tracker. Any time you click the mouse button on a feature map image, the intensity of the map at that point is recovered and displayed, along with the mouse's position with the map. Fig. C.5 shows the mouse position in frequency and time and the map value at that position. It also shows the horizontal and vertical slices through this time/frequency map — a frequency slice and spectrum, respectively. You can also perform filtering with SoundExplorer's computational engine via the interactive browser. After changing parameter values for a filter, you can press a "Compute" button, which makes the browser take all the relevant parameter values and start up a new process that performs the filtering you asked for. When the computation is finished, the resulting feature map is displayed for further perusal.



			cells	pixels
time:	138.321	msec	61	141
frequency:	1246.30	Hz	88	131
value:	0.09667	/ 1000		

Figure C.5: Mouse position information.

Map Storage Standard

One goal of SoundExplorer is to have the computational engine be portable to different types of computers, so that it's not restricted to just Next machines. For this to be possible, the files that SoundExplorer stores its feature maps in must be standardized. SoundExplorer's standard has three parts: a standard way to store feature maps, a standard way to store the parameters associated with a map, and a standard way of placing map files in a directory structure to keep track of them all. Details are available on request.

This work was supported in part by National Science Foundation grant IRI-8613574.

Bibliography

[Adelson86] Edward H. Adelson and James R. Bergen. The extraction of spatio-temporal energy in human and machine vision. In *Proceedings, Workshop on Motion: Representation and Analysis*, pages 151–155. IEEE Computer Society, Computer Society Press, May 1986.

[Anstis85] Stuart Anstis and Shinya Saida. Adaptation to auditory streaming of frequency-modulated tones. *Journal of Experimental Psychology: Human Perception and Performance*, **11**(3):257–271, 1985.

[Assman89a] P. F. Assman and Q. Summerfield. Modelling the perception of concurrent vowels: Vowels with the same fundamental frequency. *Journal of the Acoustical Society of America*, **85**:327–338, 1989.

[Assman89b] P. F. Assman and Q. Summerfield. Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies. *Journal of the Acoustical Society of America*, **88**:680–697, 1989.

[Barinaga90] Marcia Barinaga. The mind revealed? *Science*, **249**:856–858, August 1990.

[Békésy63] G. von Békésy. Three experiments concerned with pitch perception. *Journal of the Acoustical Society of America*, **35**(4):602–606, April 1963.

[Boer78] E. de Boer and H. R. de Jongh. On cochlear encoding: Potentialities and limitations of the reverse-correlation techinque. *Journal of the Acoustical Society of America*, **63**:115–135, 1978.

[Borden84] Gloria J. Borden and Katherine S. Harris. *Speech Science Primer: Physiology, Acoustics, and Perception of Speech.* Williams & Wilkins, Baltimore, second edition, 1984.

[Bregman73] Albert S. Bregman and Gary Dannenbring. The effect of continuity on auditory stream segregation. *Perception & Psychophysics*, **13**(2):308–312, 1973.

[Bregman75] Albert S. Bregman and Alexander Rudnicky. Auditory segregation: Stream or streams? *J. Experimental Psychology: Human Perception and Performance*, **1**(3):263–267, 1975.

[Bregman78a] Albert S. Bregman. Auditory streaming is cumulative. *J. Experimental Psychology: Human Perception and Performance*, **4**(3):380–387, 1978.

[Bregman78b] Albert S. Bregman and Steven Pinker. Auditory streaming and the building of timbre. *Canadian Journal of Psychology*, **32**(1):19–31, 1978.

[Bregman85] Albert S. Bregman, Jack Abramson, Peter Doehring, and Christopher J. Darwin. Spectral integration based on common amplitude modulation. *Perception & Psychophysics*, **37**:483–493, 1985.

[Bregman90] Albert S. Bregman. *Auditory Scene Analysis.* The MIT Press, Cambridge, Massachusetts, 1990.

[Carlyon89] Robert P. Carlyon and Richard J. Stubbs. Detecting single-cycle frequency modulation imposed on sinusoidal, harmonic, and inharmonic carriers. *Journal of the Acoustical Society of America*, **85**(6):2563–2574, June 1989.

[Carlyon91] Robert P. Carlyon. Discriminating between coherent and incoherent frequency modulation of complex tones. *Journal of the Acoustical Society of America*, **89**(1):329–340, January 1991.

[Chafe85] Chris Chafe, David A. Jaffe, et al. Techniques for note identification in polyphonic music. In *Proc. International Computer Music Conference*, pages 399–405, 1985.

[Chafe86] Chris Chafe and David A. Jaffe. Source separation and note identification in polyphonic music. *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, **2**:25.6.1–25.6.4, April 86.

[Chowning70] John M. Chowning. The simulation of moving sound sources. In *Proc. Audio Engineering Society Convention*, May 1970.

[Chowning74] John M. Chowning et al. Computer simulation of music instrument tones in reverberant environments. Technical Report STAN-M-1, Stanford University Department of Music, Stanford, CA 94305, June 1974. Available from the Stanford Center for Computer Research in Music and Acoustics.

[Chowning80] John M. Chowning. Computer synthesis of the singing voice. In *Sound Generation in Winds, Strings, Computers*, pages 4–13. Royal Swedish Academy of Music, Stockholm, 1980. Publ. No. 29.

[Chowning84] John M. Chowning, Loren Rush, et al. Intelligent systems for the analysis of digitized acoustic signals. Technical Report STAN-M-16, Stanford Department of Music, January 1984.

[Ciocca89] Walter Ciocca and Albert S. Bregman. The effects of auditory streaming on duplex perception. *Perception & Psychophysics*, **46**(1):39–48, 1989.

[Cohen84] Elizabeth A. Cohen. Some effects of inharmonic partials on interval perception. *Music Perception*, **1**(3):323–349, spring 1984.

[Cohen87] Marion F. Cohen and Earl D. Schubert. Influence of place synchrony on detection of a sinusoid. *Journal of the Acoustical Society of America*, **81**(2):452–458, February 1987.

[Cooke91] Martin Peter Cooke. *Modelling Auditory Processing and Organisation.* PhD thesis, University of Sheffield, May 1991.

[Darwin84] Christopher J. Darwin. Perceiving vowels in the presence of another sound: Constraints on formant perception. *Journal of the Acoustical Society of America*, **76**(6):1636–1647, 1984.

[Deutsch75] Diana Deutsch. Two-channel listening to musical scales. *Journal of the Acoustical Society of America*, **57**(5):1156–1160, May 1975.

[Dirks70] Donald D. Dirks and Deborah Bower. Effect of forward and backward masking on speech intelligibility. *Journal of the Acoustical Society of America*, **47**(4):1003–1008, 1970.

[Dowling78] W. J. Dowling. Scale and contour: Two components of a theory of memory for melodies. *Psychological Review*, **85**(4):341–354, 1978.

[Duifhuis82] H. Duifhuis, L. F. Willems, and R. J. Sluyter. Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception. *Journal of the Acoustical Society of America*, **71**(6):1568–1580, June 1982.

[Durlach63] Nathaniel I. Durlach. Equalization and cancellation theory of binaural masking-level differences. *Journal of the Acoustical Society of America*, **35**(8):1206–1218, August 1963.

[Durlach64] Nathaniel I. Durlach. Note on binaural masking-level differences at high frequencies. *Journal of the Acoustical Society of America*, **36**(3):576–581, March 1964.

[Elliott79] Lois L. Elliott. Backward and forward masking of probe tones of different frequencies. In Earl D. Schubert, editor, *Psychological Acoustics*, chapter 28, pages 297–298. Dowden, Hutchinson & Ross, Stroudsburg, Pa., 1979. Reprinted from J. Acoust. Soc. Am. 34:1116-1117 (1962).

[Erickson82] Robert Erickson. New music and psychology. In Diana Deutsch, editor, *The Psychology of Music*, chapter 18, pages 517–536. Academic Press Limited, London, 1982.

[Fay36] R. D. Fay. A method for obtaining natural directional effects in a public-address system. *Journal of the Acoustical Society of America*, **7**:239 (A), 1936.

[Fox86] Peter T. Fox et al. Mapping human visual cortex with positron emission tomography. *Nature*, **323**(6091):806–809, October 1986.

[Freeman75] Walter J. Freeman. *Mass Action in the Nervous System.* Academic Press Limited, London, 1975.

[Fucks62] Wilhelm Fucks. Mathematical analysis of formal structure of music. *IRE Transactions on Information Theory, IT*, **8**:225–228, 1962.

[Gardner68] Mark B. Gardner. Historical background of the Haas and/or precedence effect. *Journal of the Acoustical Society of America*, **43**(6):1243–1248, 1968.

[Gardner79] R. B. Gardner and J. P. Wilson. Evidence for direction-specific channels in the processing of frequency modulation. *Journal of the Acoustical Society of America*, **66**(3):704–709, September 1979.

[Gardner86] R. B. Gardner and Christopher J. Darwin. Grouping of vowel harmonics by frequency modulation: Absence of effects on phonemic categorization. *Perception & Psychophysics*, **40**(3):183–187, 1986.

[Glasberg90] Brian R. Glasberg and Brian C. J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, **47**:103–138, 1990.

[Goldstein73] Julius L. Goldstein. An optimum processor theory for the central formation of the pitch of complex tones. *Journal of the Acoustical Society of America*, **54**(6):1496–1516, 1973.

[Gordon84] John William Gordon. *Perception of Attack Transients in Musical Tones.* PhD thesis, Stanford University, May 1984. Published as Department of Music Tech. Rept. STAN-M-17.

[Gray89a] Charles M. Gray, Peter König, Andreas K. Engel, and Wolf Singer. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature*, **338**:334–337, March 1989.

[Gray89b] Charles M. Gray and Wolf Singer. Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proc. Natl. Acad. Sci. USA, Neurobiology*, **86**:1698–1702, March 1989.

[Grey75] John M. Grey. *An Exploration of Musical Timbre.* PhD thesis, Stanford University, Stanford, CA 94305, 1975. Available as Stanford Department of Music Technical Report STAN-M-2.

[Hafter71] Ervin R. Hafter. Quantitive evaluation of a lateralization model of masking-level differences. *Journal of the Acoustical Society of America*, **50**(4):1116–1122, February 1971.

[Hall36] W. M. Hall. A method for maintaining in a public address system the illusion that the sound comes from the speaker's mouth. *Journal of the Acoustical Society of America*, **7**:239 (A), 1936.

[Hall84] Joseph W. Hall, Mark P. Haggard, and Mariano A. Fernandes. Detection in noise by spectro-temporal pattern analysis. *Journal of the Acoustical Society of America*, **76**:50–56, 1984.

[Hartmann88] William Morris Hartmann. Pitch perception and the segregation and integration of auditory entities. In Gerald M. Edelman, W. Einar Gall, and W. Maxwell Cowan, editors, *Auditory Function: Neurobiological Bases of Hearing*, chapter 21, pages 623–645. John Wiley and Sons, New York, 1988.

[Hebb49] Donald O. Hebb. *The Organization of Behavior.* John Wiley and Sons, New York, 1949.

[Heeger88] David J. Heeger. Optical flow using spatiotemporal filters. *International Journal of Computer Vision*, :279–302, 1988.

[Heeger91] David J. Heeger. Nonlinear model of neural responses in cat visual cortex. In Michael S. Landy and J. Anthony Movshon, editors, *Computational Models of Visual Processing.* The MIT Press, Cambridge, Massachusetts, 1991.

[Helmholtz54] H. L. F. Helmholtz. *On the Sensations of Tone.* Dover Press, New York, 1954. Second English Edition. Translated by A. J. Ellis.

[Henry51] Joseph Henry. *Scientific Writings of Joseph Henry, Part II, 1847-1878*, pages 295–296. Smithsonian Institution, Washington, D.C., 1851. Referenced in Gardner (1968).

[Houtsma79] A. J. M. Houtsma. Musical pitch of two-tone complexes and predictions by modern pitch theories. *Journal of the Acoustical Society of America*, **66**(1):87–99, 1979.

[Houtsma87] A. J. M. Houtsma, T. D. Rossing, and W. M. Wagenaars. *Auditory Demonstrations (Compact Disk).* Philips, 1987. Philips recording no. 1126-061. Available from the Acoustical Society of America, Woodbury, NY.

[Ingvar89] David H. Ingvar. On music and its cerebral correlates. In Søren Nielzen and Olle Olsson, editors, *Structure and Perception of*

Electroacoustic Sound and Music, pages 181–184. Excerpta Medica, Amsterdam, 1989.

[Jeffress56] Lloyd A. Jeffress, Hugh C. Blodgett, Thomas T. Sandel, and Charles L. Wood, III. Masking of tonal signals. *Journal of the Acoustical Society of America*, **28**:416–426, 1956.

[Jeffress72] Lloyd A. Jeffress. Binaural signal detection: Vector theory. In Jerry V. Tobias, editor, *Foundations of Modern Auditory Theory, Vol. II*, chapter 9, pages 351–368. Academic Press Limited, London, 1972.

[Jenison91] Rick L. Jenison, Steven Greenberg, Keith R. Kluender, and William S. Rhode. A composite model of the auditory periphery for the processing of speech based on the filter response functions of single auditory-nerve fibers. *Journal of the Acoustical Society of America*, **90**:773–786, 1991.

[Johannesma86] P. Johannesma et al. From synchrony to harmony: Ideas on the function of neural assemblies and on the interpretation of neural synchrony. In G. Palm and A. Aertsen, editors, *Brain Theory*, pages 25–47. Springer-Verlag, Berlin, 1986.

[Julesz81] Bela Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, **290**:91–97, March 1981.

[Kanizsa79] Gaetano Kanizsa. *Organization in Vision: Essays on Gestalt Perception.* Praeger Scientific Press, New York, 1979.

[Kashima85] Kyle L. Kashima and Bernard M. Mont-Reynaud. The bounded-Q frequency transform. Technical Report STAN-M-28, Stanford University Department of Music, 1985.

[Knudsen78] Eric I. Knudsen and Masakazu Konishi. Space and frequency are represented separately in auditory midbrain of the owl. *Journal of Neurophysiology*, **41**(4):870–884, 1978.

[Knudsen81] Eric I. Knudsen. The hearing of the barn owl. *Scientific American*, :113–125, December 1981.

[Lakatos91] Stephen Lakatos. Personal communication. June, 1991.

[Langner88] Gerald Langner and Christoph E. Schreiner. Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms. *Journal of Neurophysiology*, **60**(6):1799–1822, December 1988.

[Lee90] Kai fu Lee, Hsiao-Wuen Hon, and Raj Reddy. An overview of the Sphinx speech-recognition system. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **38**:35–45, 1990.

[Licklider51] J. C. R. Licklider. A duplex theory of pitch perception. *Experientia*, **7**:128–133, 1951.

[Lindemann86] W. Lindemann. Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals. *Journal of the Acoustical Society of America*, **80**:1608–1622, December 1986.

[Lyon82] Richard F. Lyon. A computational model of filtering, detection, and compression in the cochlea. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 1982.

[Lyon84] Richard F. Lyon. Computational models of neural auditory processing. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 36.1.1–36.1.4, 1984.

[Lyon86] Richard F. Lyon. Experiments with a computational model of the cochlea. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1975–1978, 1986.

[Lyon88] Richard F. Lyon and Carver A. Mead. Cochlear hydrodynamics demystified. Technical report, California Institute of Technology, February 1988.

[Malik89] Jitendra Malik and Pietro Perona. A computational model of texture segmentation. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 1989.

[Malsburg86a] Christoph von der Malsburg. Am I thinking assemblies? In G. Palm and A. Aertsen, editors, *Brain Theory*, pages 161–176. Springer-Verlag, Berlin, 1986.

[Malsburg86b] Christoph von der Malsburg and Werner Schneider. A neural cocktail-party processor. *Biological Cybernetics*, **54**:29–40, 1986.

[Marr82] David Marr. *Vision.* W. H. Freeman and Company, New York, 1982.

[Massaro87] Dominic W. Massaro. *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1987.

[Mathews80] Max V. Mathews and John R. Pierce. Harmony and nonharmonic partials. *Journal of the Acoustical Society of America*, **68**(5):1252–1257, November 1980.

[McAdams84] Stephen McAdams. *Spectral Fusion, Spectral Parsing, and the Formation of Auditory Images.* PhD thesis, Stanford University, May 1984.

[McAdams89] Stephen McAdams. Segregation of concurrent sounds I: Effects of frequency modulation coherence. *Journal of the Acoustical Society of America*, **86**(6):2148–2159, December 1989.

[McAulay86] Robert J. McAulay and Thomas F. Quatieri. Speech analysis/resynthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **ASSP-34**(4):744–754, August 1986.

[Meddis89] Ray Meddis and Michael Hewitt. Virtual pitch and phase sensitivity studied using a computer model of the auditory periphery. Submitted to Journal of the Acoustical Society of America, September 1989.

[Mellinger91] David K. Mellinger and Bernard M. Mont-Reynaud. SoundExplorer: A workbench for investigating source separation. In *Proceedings of the International Computer Music Conference*, pages 90–93, October 1991.

[Mendelson85] J. R. Mendelson and M. S. Cynader. Sensitivity of cat auditory primary cortex (AI) neurons to the direction and rate of frequency modulation. *Brain Research*, **327**:331–335, 1985.

[Metz68] P. J. Metz, G. von Bismark, and N. I. Durlach. Further results on binaural unmasking and the EC model. II. Noise bandwidth and interaural phase. *Journal of the Acoustical Society of America*, **43**(5):1085–1091, 1968.

[Miller50] George A. Miller and J. C. R. Licklider. The intelligibility of interrupted speech. *Journal of the Acoustical Society of America*, **22**(2):167–173, 1950.

[Møller77] Aage R. Møller. Coding of time-varying sounds in the cochlear nucleus. *Audiology*, **17**:446–468, 1977.

[Mont-Reynaud85] Bernard M. Mont-Reynaud and Mark Goldstein. On finding rhythmic patterns in musical lines. In *Proc. International Computer Music Conference*, pages 391–397, 1985.

[Mont-Reynaud90] Bernard M. Mont-Reynaud, Richard O. Duda, Albert Bregman, and Jay M. Tenenbaum. A computational model of auditory perception. Research proposal submitted to DARPA, June 1990.

[Mont-Reynaud91] Bernard Mont-Reynaud. Personal communication. Feb., 1991.

[Moore83] Brian C. J. Moore and Brian R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, **74**:750–753, 1983.

[Moore85a] Brian C. J. Moore, Brian R. Glasberg, and Robert W. Peters. Relative dominance of individual partials in determining the pitch of complex tones. *Journal of the Acoustical Society of America*, **77**(5):1853–1860, May 1985.

[Moore85b] Brian C. J. Moore, Robert W. Peters, and Brian R. Glasberg. Thresholds for the detection of inharmonicity in complex tones. *Journal of the Acoustical Society of America*, **77**(5):1861–1867, May 1985.

[Moore89] Brian C. J. Moore. *An Introduction to the Psychology of Hearing.* Academic Press Limited, London, third edition, 1989.

[Moore90] Brian C. J. Moore. Co-modulation masking release: Spectro-temporal pattern analysis in hearing. *British Journal of Audiology*, **24**:131–137, 1990.

[Moorer74] James A. Moorer. The optimum comb method of pitch period analysis for continuous digitized speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **22**(5):330–338, October 1974.

[Moorer75] James A. Moorer. *On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer.* PhD thesis, Stanford University Department of Music, May 1975. Available as Stanford Department of Music Tech. Rept. No. STAN-M-3.

[Oppenheim75] Alan V. Oppenheim and Ronald W. Schafer. *Digital Signal Processing.* Prentice-Hall, Englewood Cliffs, New Jersey, 1975.

[Ortmann26] Otto Ortmann. On the melodic relativity of tones. *Psychological Monographs*, **35**(Whole No. 162), 1926.

[Parsons76] Thomas W. Parsons. Separation of speech from interfering noise by means of harmonic selection. *Journal of the Acoustical Society of America*, **60**(4):911–918, October 1976.

[Paul90] D. B. Paul. Speech recognition using hidden Markov models. *Lincoln Laboratory Journal*, **3**(1):41–62, 1990.

[Pickles88] James O. Pickles. *An Introduction to the Physiology of Hearing.* Academic Press Limited, London, second edition, 1988.

[Pierce83] John R. Pierce. *The Science of Musical Sound.* W. H. Freeman and Company, New York, first edition, 1983.

[Pierce90] John R. Pierce. Rate, place, and pitch with tonebursts. *Music Perception*, **7**(3):205–212, Spring 1990.

[Rabiner75] Lawrence R. Rabiner and Bernard Gold. *Theory and Application of Digital Signal Processing.* Prentice-Hall, Englewood Cliffs, New Jersey, 1975.

[Rand74] T. C. Rand. Dichotic release from masking for speech. *Journal of the Acoustical Society of America*, **55**(3):678–680, March 1974.

[Rasch78] R. A. Rasch. The perception of simultaneous notes such as in polyphonic music. *Acustica*, **40**:21–33, 1978.

[Rasch79] R. A. Rasch. Synchronization in performed ensemble music. *Acustica*, **43**:121–131, 1979.

[Reynolds83] Roger Reynolds. *Archipelago.* C. F. Peters, New York, 1983. For orchestra and computer-generated tape.

[Rhode86] William S. Rhode and Philip H. Smith. Encoding timing and intensity in the ventral cochlear nucleus of the cat. *Journal of Neurophysiology*, **56**(2):261–286, August 1986.

[Scharf70] Bertram Scharf. Critical bands. In Jerry V. Tobias, editor, *Foundations of Modern Auditory Theory*, chapter 5, pages 159–202. Academic Press Limited, London, 1970.

[Scheffers83] Michael T. M. Scheffers. Simulation of auditory analysis of pitch: An elaboration of the DWS pitch meter. *Journal of the Acoustical Society of America*, **74**(6):1716–1725, December 1983.

[Schloss85] W. Andrew Schloss. *On the Automatic Transcription of Percussive Music: From Acoustic Signal to High-Level Analysis.* PhD thesis, Stanford University, Stanford, CA 94305, May 1985.

[Schooneveldt87] Gregory P. Schooneveldt and Brian C. J. Moore. Comodulation masking release (CMR): Effects of signal frequency, flanking-band frequency, masker bandwidth, flanking-band level, and monotic versus dichotic presentation of the flanking band. *Journal of the Acoustical Society of America*, **82**(6):1944–1956, December 1987.

[Schooneveldt88] G. P. Schooneveldt and B. C. J. Moore. Failure to obtain comodulation masking release with frequency-modulated maskers. *Journal of the Acoustical Society of America*, **83**(6):2290–2292, June 1988.

[Schottstaedt91] William Schottstaedt. Personal communication. Aug., 1991.

[Schreiner86] Christoph E. Schreiner and J. V. Urbas. Representation of amplitude modulation in the auditory cortex of the cat. I. Anterior auditory field. *Hearing Research*, **21**:227–241, 1986.

[Schreiner88a] Christoph E. Schreiner and Gerald Langner. Coding of temporal patterns in the central auditory nervous system. In G. M. Edelman, W. E. Gall, and W. M. Cowan, editors, *Auditory Function*, chapter 11, pages 337–361. John Wiley and Sons, New York, 1988.

[Schreiner88b] Christoph E. Schreiner and John V. Urbas. Representation of amplitude modulation in the auditory cortex of the cat. II. Comparison between cortical fields. *Hearing Research*, **32**:49–64, 1988.

[Schreiner90] Christoph E. Schreiner and Julie R. Mendelson. Functional topography of cat primary auditory cortex: Distribution of integrated excitation. *Journal of Neurophysiology*, **64**(5), November 1990.

[Schroeder68] M. R. Schroeder. Period histogram and product spectrum: New methods for fundamental-frequency measurement. *Journal of the Acoustical Society of America*, **43**(4):829–834, 1968.

[Schwede83] Gary W. Schwede. An algorithm and architecture for constant-Q spectrum analysis. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1384–1387, April 1983.

[Seneff88] Stephanie Seneff. A joint-synchrony/mean-rate model of auditory speech processing. *J. Phonetics*, **16**:55–76, 1988.

[Serra88] Xavier Serra. *An Environment for the Analysis, Transformation, and Resynthesis of Music Sounds.* PhD thesis, Stanford University Department of Music, May 1988.

[Shamma85] Shihab A. Shamma. Speech processing in the auditory system I: The representation of speech sounds in the responses of the auditory nerve. *Journal of the Acoustical Society of America*, **78**(5):1612–1621, November 1985.

[Shamma91] Shihab A. Shamma, James W. Fleshman, and Philip R. Wiser. A functional model of primary auditory cortex: Spectral orientation columns. Submitted to *Hearing Research*, 1991.

[Sheeline82] Christopher W. Sheeline. *An Investigation of the Effects of Direct and Reverberant Signal Interactions on Auditory Distance Perception.* PhD thesis, Stanford University, 1982.

[Shepard82] Roger N. Shepard. Geometrical approximations to the structure of musical pitch. *Psychological Review*, **89**:305–333, 1982.

[Shepard89] Roger N. Shepard. Internal representation of universal regularities: A challenge for connectionism. In Lynn Nadel et al., editors, *Neural Connections, Mental Computation*, chapter 4, pages 104–134. The MIT Press, Cambridge, Massachusetts, 1989.

[Shepard91] Roger N. Shepard. Personal communication. Nov., 1991.

[Slaney88] Malcolm Slaney. Lyon's cochlear model. Technical Report 13, Apple Computer, 1988. Available from the Apple Corporate Library, Cupertino, CA 95014.

[Slaney90] Malcolm Slaney. Interactive signal processing documents. *IEEE ASSP Magazine*, **7**(2):8–20, April 1990.

[Smith87] Julius Orion Smith and Xavier Serra. PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation. In *Proc. International Computer Music Conference*, pages 290–297, 1987.

[Stevens37] S. S. Stevens, J. Volkmann, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, **8**:185–190, 1937.

[Stevens40] S. S. Stevens and J. Volkmann. The relation of pitch to frequency: A revised scale. *American Journal of Psychology*, **53**(3):329–353, 1940.

[Suga90] Nobuo Suga. Cortical computational maps for auditory imaging. *Neural Networks*, **3**:3–21, 1990.

[Takahashi84] T. Takahashi, A. Moiseff, and M. Konishi. Time and intensity cues are processed independently in the auditory system of the owl. *J. Neuroscience*, **4**(7):1781–1786, July 1984.

[Terhardt72] Ernst Terhardt. Zur Tonhöhenwahrnehmung von Klängen II: Ein Funktionsschema. *Acustica*, **26**:187–199, 1972.

[Tobias59] Jerry V. Tobias and Stanley Zerlin. Lateralization threshold as a function of stimulus location. *Journal of the Acoustical Society of America*, **31**(12):1591–1594, December 1959.

[van Noorden75] Leon P. A. S. van Noorden. *Temporal Coherence in the Perception of Time Sequences.* PhD thesis, Technische Hogeschool Eindhoven, 1975. Unpublished.

[van Noorden77] Leon P. A. S. van Noorden. Minimum differences of level and frequency for perceptual fission of tone sequences ABAB. *Journal of the Acoustical Society of America*, **61**(4):1041–1045, April 1977.

[Vercoe88] Barry Vercoe. Hearing polyphonic music with the Connection Machine. Technical report, MIT Media Lab, 1988.

[Waibel89] Alex Waibel. A time-delay neural network. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **37**(3):328–339, March 1989.

[Warren82] Richard M. Warren. *Auditory Perception: A New Synthesis.* Pergamon, New York, 1982.

[Warren84] William H Warren, Jr. and Robert R. Verbrugge. Auditory perception of breaking and bouncing events. *J. Experimental Psychology: Human Perception and Performance*, **10**(5):704–712, 1984.

[Watson85] Andrew B. Watson and Albert J. Ahumada, Jr. Model of human visual-motion sensing. *J. Optical Society of America*, **2**(2):322–342, February 1985.

[Weintraub85] Mitchel Weintraub. *A Theory and Computational Model of Auditory Monaural Sound Separation.* PhD thesis, Stanford University, Stanford, CA 94305, August 1985.

[Wessel79] David L. Wessel. Timbre space as a musical control structure. *Computer Music Journal*, **3**(2):45–52, Summer 1979.

[Whitfield65] I. C. Whitfield and E. F. Evans. Responses of auditory cortical neurons to stimuli of changing frequency. *Journal of Neurophysiology*, **28**:655–672, 1965.

[Wickesberg90] Robert E. Wickesberg and Donata Oertel. Delayed, frequency-specific inhibition in the cochlear nuclei of mice: A mechanism for monaural echo suppression. *J. Neuroscience*, **10**(6):1762–1768, June 1990.

[Wightman73] Frederic L. Wightman. The pattern-transformation model of pitch. *Journal of the Acoustical Society of America*, **54**(2):407–416, 1973.

[Witkin83] Andrew P. Witkin and Jay M. Tenenbaum. On the role of structure in vision. In *Human and Machine Vision*, pages 481–543. Academic Press, 1983.

[Yin88] Tom C. Yin and Joseph C. K. Chan. Neural mechanisms underlying interaural time sensitivity to tones and noise. In Gerald M. Edelman, W. Einar Gall, and W. Maxwell Cowan, editors, *Auditory Function: Neurobiological Bases of Hearing*, chapter 13, pages 385–430. John Wiley and Sons, New York, 1988.

[Young88] Eric D. Young, William P. Shofner, John A. White, Jeanne-Marie Robert, and Herbert F. Voigt. Response properties of cochlear nucleus neurons in relationship to physiological mechanisms. In Gerald M. Edelman, W. Einar Gall, and W. Maxwell Cowan, editors, *Auditory Function: Neurobiological Bases of Hearing*, chapter 9, pages 277–312. John Wiley and Sons, New York, 1988.

[Zurek80] P. M. Zurek. The precedence effect and its possible role in the avoidance of interaural ambiguities. *Journal of the Acoustical Society of America*, **67**(3):952–964, March 1980.