Prediction-driven computational auditory scene analysis

by

Daniel P. W. Ellis

B.A.(hons) Engineering (1987) Cambridge University

S.M. Electrical Engineering (1992) Massachusetts Institute of Technology

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Electrical Engineering

at the
Massachusetts Institute of Technology

June 1996

Signature of author .....
Department of Electrical Engineering and Computer Science
April 17, 1996

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

Accepted by .....
Frederic R. Morgenthaler
Chair, Department Committee on Graduate Students

Prediction-driven computational auditory scene analysis

by Daniel P. W. Ellis

Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of Doctor of Philosophy of Electrical Engineering.

Abstract

The sound of a busy environment, such as a city street, gives rise to a perception of numerous distinct events in a human listener – the 'auditory scene analysis' of the acoustic information. Recent advances in the understanding of this process from experimental psychoacoustics have led to several efforts to build a computer model capable of the same function. This work is known as 'computational auditory scene analysis'.

The dominant approach to this problem has been as a sequence of modules, the output of one forming the input to the next. Sound is converted to its spectrum, cues are picked out, and representations of the cues are grouped into an abstract description of the initial input. This 'data-driven' approach has some specific weaknesses in comparison to the auditory system: it will interpret a given sound in the same way regardless of its context, and it cannot 'infer' the presence of a sound for which direct evidence is hidden by other components.

The 'prediction-driven' approach is presented as an alternative, in which analysis is a process of reconciliation between the observed acoustic features and the predictions of an internal model of the sound-producing entities in the environment. In this way, predicted sound events will form part of the scene interpretation as long as they are consistent with the input sound, regardless of whether direct evidence is found. A blackboard-based implementation of this approach is described which analyzes dense, ambient sound examples into a vocabulary of noise clouds, transient clicks, and a correlogram-based representation of wide-band periodic energy called the weft.

The system is assessed through experiments that firstly investigate subjects' perception of distinct events in ambient sound examples, and secondly collect quality judgments for sound events resynthesized by the system. Although rated as far from perfect, there was good agreement between the events detected by the model and by the listeners. In addition, the experimental procedure does not depend on special aspects of the algorithm (other than the generation of resyntheses), and is applicable to the assessment and comparison of other models of human auditory organization.

Thesis supervisor: Barry L. Vercoe

Title: Professor of Media Arts and Sciences

Acknowledgments

First and foremost,. my thanks are due to Barry Vercoe for creating the group in which I was able to do this work, for inviting me, sight unseen, to join his little ensemble back in 1989, and for supporting me since then. Barry opened the door for me to a whole new world of knowledge and practice, and his unwavering commitment to providing a creative and protected environment for his students stands to his considerable credit.

Next, my thanks go to my thesis readers, Louis Braida and Ben Gold. Lou's insightful reading of my original proposal spurred me into conducting subjective tests, immeasurably improving the final work. Ben's breadth of knowledge combined with a tremendous patience and positivity provided invaluable support in some of the more difficult moments of the process.

At this point, the last stop on a very long path through the education system, I am inspired to reflect on all the teachers who have brought me here. Ultimate credit goes to my parents; I have memories, thickly fogged by childhood, of being helped with analyzing soil from the garden in jars of water as part of a primary school project. I treasure this image as the archeological evidence of my early channeling into the scientific order. I hope that my eventual direction, such as it is, has provided some vindication of my parents' continual support of my education in advance of their other concerns

I have gained benefit from many other teachers between then and now, and I feel (in some cases belatedly) grateful to them all. At the risk of unfair omissions resulting from poor memory, I would like to mark my gratitude to Sophie Tatchell, Carol Allen, John Armstrong, Paul Brna, Alan Gent, Dennis Archer, Keith Fuller, Harry Pearson, Don Spivey, Pam Parsons, David Butcher, Tim Williams, Ken Snuggs, Ruth Whiting, John Rogers, David Moore, John Denton, Peter Rayner and Frank Fallside.

For providing the means for me to travel to the United States in the first place, and for renewing my visa far beyond the terms of the original agreement, I am indebted to the Harkness Fellowships of the Commonwealth Fund of New York and their most agreeable administration by Robert Kostrzewa.

I'd like to thank all members, past and present, of the old Music and Cognition group and its heir, the Machine Listening group, for their part in making a wonderful and supportive working community. Particular honors go to Bill Gardner, Mike Casey, Judy Brown, Keith Martin, Eric Scheirer, Adam Lindsay and Paris Smaragdis, not forgetting Nicolas Saint-Arnaud, Jeff Bilmes, Alan Ruttenberg, Mike Travers, Shahrokh Yadegari, Mary-Ann Norris, Kevin Peterson, Andy Hong, Joe Chung and Robert Rowe.

Other people who I would like to thank for having helped make my time at the MIT Media Lab such an ideal experience include (and I apologize for the inevitable omissions): Stan Sclaroff, Irfan Essa, Lee Campbell, Baback Moghaddam, Fang Liu, Sourabh Niyogi, Lisa Stifelman, Barry Arons, Betty-Lou McClanahan, Laureen Chapman, Marty Friedmann, Mike Hawley, Sandy Pentland, Ken Haase, Pushpinder Singh, Judy Bornstein, Molly

Bancroft, Ben Lowengard, Greg Tucker, Doug Alan, Viet Anh, David Blank-Edelman, Jane Wojcik and Marc Bejarano.

Special thanks for continued emotional support to Bill Gardner, Janet Cahn and Dave Rosenthal.

For interest in these ideas and advice both technical and professional, I am grateful to: Malcolm Slaney, Dick Duda, Al Bregman, Hideki Kawahara, Martin Cooke, Guy Brown, Nelson Morgan, Steve Greenberg, Alon Fishbach, Lonce Wyse, Kunio Kashino, Ray Meddis, Bill Woods, Tom Ngo, Michele Covell, Ted Adelson, Whitman Richards, Pat Zurek, Bill Peake, and Hiroshi Okuno.

I was given specific practical help in completing this thesis by Bill Gardner, Keith Martin, Eric Scheirer, Mike Casey, Adam Lindsay, Paris Smaragdis, Janet Cahn, Barry Arons, Sarah Coleman, Lorin Wilde, Malcolm Slaney and Betty-Lou McClanahan. All their contributions are acknowledged with many thanks.

I am grateful for the indispensable software tools made freely available to me by their authors. These include Malcolm Slaney's filter code, Joe Winograd's blackboard implementation, John Ousterhout's Tcl language with the object-oriented extensions of Dean Sheenan and Mike McLennan, and Pete Keleher's Alpha editor.

I am specifically indebted to Guy Brown for permission to use the sound examples from his thesis. Guy, Dave Mellinger and Martin Cooke also were kind enough to give me permission to adapt figures from their work.

Finally, to Sarah for your sometimes inexplicable but always unwavering love, support and understanding: thank you. I hope that my reaching this milestone provides some recompense for the privations it has entailed, and I promise not to do it again.

Contents

1 Introduction 9

1.1 Auditory Scene Analysis for real scenes 9
1.2 Modeling auditory organization - motivation and approach 10
1.3 The prediction-driven model 11
1.4 Applications 13
1.5 Ideas to be investigated 14
1.6 Specific goals 15
1.7 Outline of this document 16

2 An overview of work in Computational Auditory Scene Analysis 17

2.1 Introduction 17
2.1.1 Scope 17
2.2 Foundation: Auditory Scene Analysis 18
2.3 Related work 20
2.3.1 Sound models 20
2.3.2 Music analysis 21
2.3.3 Models of the cochlea and auditory periphery 22
2.3.4 Speech processing and pre-processing 23
2.3.5 Machine vision scene analysis systems 25
2.4 The data-driven computational auditory scene analysis system 26
2.5 A critique of data-driven systems 35
2.6 Advances over the data-driven approach 37
2.6.1 Weintraub's state-dependent model 37
2.6.2 Blackboard systems 39
2.6.3 The IPUS blackboard architecture 40
2.6.4 Other innovations in control architectures 43
2.6.5 Other 'bottom-up' systems 44
2.6.6 Alternate approaches to auditory information processing 45
2.6.7 Neural network models 46
2.7 Conclusions and challenges for the future 47

3 The prediction-driven approach 48

3.1 Psychophysical motivation 48
3.2 Central principles of the prediction-driven approach 53
3.3 The prediction-driven architecture 56
3.4 Discussion 59
3.5 Conclusions 62

4 The implementation 65

4.1 Implementation overview 65
4.1.1 Main modules 65
4.1.2 Overview of operation : prediction and reconciliation 66
4.2 The front end 67

4.2.1 Cochlear filterbank 67
4.2.2 Time-frequency intensity envelope 70
4.2.3 Correlogram 71
4.2.4 Summary autocorrelation (periodogram) 76
4.2.5 Other front-end processing 78
4.3 Representational elements 80
4.3.1 Noise clouds 81
4.3.2 Transient (click) elements 84
4.3.3 Weft (wideband periodic) elements 87
4.4 The reconciliation engine 92
4.4.1 The blackboard system 92
4.4.2 Basic operation 98
4.4.3 Differences from a traditional blackboard system 104
4.5 Higher-level abstractions 105

5 Results and assessment 107

5.1 Example analyses 107
5.1.1 Bregman's alternating noise example 107
5.1.2 A speech example 114
5.1.3 Mixtures of voices 117
5.1.4 Complex sound scenes: the "city-street ambience" 121
5.2 Testing sound organization systems 122
5.2.1 General considerations for assessment methods 123
5.2.2 Design of the subjective listening tests 124
5.3 Results of the listening tests 131
5.3.1 The training trial 131
5.3.2 The city-sound 133
5.3.3 "Construction" sound example 137
5.3.4 "Station" sound example 142
5.3.5 The double-voice example 143
5.3.6 Experiment part B: Rating of resyntheses 144
5.3.7 Experiment part C: Ranking of resynthesis versions 148
5.4 Summary of results 151

6 Summary and conclusions 153

6.1 Summary 153
6.1.1 What has been presented 153
6.1.2 Future developments of the model 154
6.2 Conclusions 155
6.2.1 Reviewing the initial design choices 156
6.2.2 Insights gained during the project 157
6.2.3 A final comparison to real audition 158
6.3 The future of Computational Auditory Scene Analysis 161

Appendix A: Derivation of the weft update equation 163

Appendix B: Sound examples 169

Appendix C: Computational environment 171

References 173

1.1 Auditory Scene Analysis for real scenes

I have in front of me a ten-second fragment of sound. It is from a sound-effects collection, and is described as "city street ambience". When played, I hear a general background noise over which there are a couple of car horns, a loud crash of a metal door being slammed, some squealing brakes, and the rumble of an engine accelerating away. This description may be considered an analysis of the acoustic scene embodied in the sound; the goal of my research is to build a computer system that can make this kind of analysis of these kinds of sounds – not so much in terms of the causal accounts (car horn, door slam) but in terms of the number and general properties of the distinct events.

This ability in humans has been considered in psychoacoustics under the titles of auditory perceptual organization or auditory scene analysis [Breg90]. These studies construct experimental stimuli consisting of a few simple sounds such as sine tones or noise bursts, and then record subjects' interpretation of the combination. The work has been very revealing of the mechanisms by which structure is derived from sound, but typically it fails to address the question of scaling these results to more complex sounds: In a real-world environment there may be any number of contributors to the total sound scene; how can we even define the basic elements, the analogs of the simple sine tones and bursts of white noise? When elements are distinct in a straightforward time-frequency representation, the segmentation is obvious (or at least the obvious segmentation turns out to be perceptually acceptable). But the spectrogram of the "city street ambience," which is an almost featureless mess, highlights the limitations of these results.

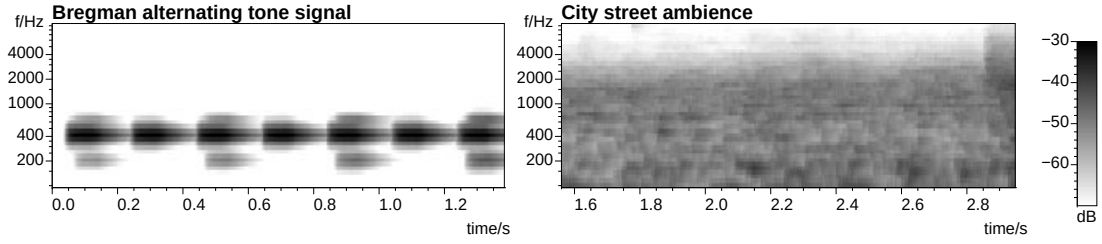

Figure 1.1: Cochlea-model spectrogram of a typical psychoacoustic stimuli (left) compared with a real environmental sound (the "city street ambience," right).

An ability to structure a continuous sound mixture into a set of independent sources is absolutely basic to an auditory system that can confer an evolutionary advantage. Without it, we would only be able to interpret sounds when they occurred against a silent background. The combinatoric problem of recognizing mixtures of sounds as single, composite entities is intractable given the range of sounds we must handle. Rather, there must be a process of organization or *segmentation* of the auditory signal that is

applied prior to (although most likely in conjunction with) the function of recognizing and describing the individual sound-objects. Since the nature of the sound world is that mixtures are the rule and isolated events the exception (or so it seems from my noisy apartment overlooking Harvard Square), it is to be expected that the auditory system has evolved to perform this separation transparently and effectively, despite the fact that the cues to separation in the "city street ambience" are basically invisible under conventional signal analysis. This aspect of auditory processing is quite unconscious, reinforcing the realization that the ability to organize sound mixtures is not unique to humans, but must be possessed by all higher mammals, indeed by any creature that wants to adapt its behavior based on the mishmash of available acoustic information.

1.2 Modeling auditory organization - motivation and approach

This thesis originates in the view that the human auditory system can be regarded as a complex signal- and information-processing machine, and as such can be understood and explained in the same way that we can explain complex systems built by engineers, for instance the tracking systems used by air-traffic controllers. Like those human-engineered systems, our hearing mechanism relies on some underlying physics to glean information-carrying signals from the environment, which are then processed to reveal specific details to be integrated into a simplified representation of the world. Although this is not an entirely uncontroversial starting point (there are researchers who object strongly to the idea of world models as the output of perception [ChurRS94]), it is certainly conventional. It does however, have certain implications for the nature of the work to be presented.

The overriding goal of the work of which this thesis is a part is to approach a functional model of this total information processing system. The primary motivation behind this project is a desire to understand the human auditory system. There are several approaches to unraveling this mystery, the more obvious being to practice experimental psychology and physiology to probe directly the behavior and structure of the mechanism we are trying to understand. The building of computational models is a rather different but singularly valuable complement to these approaches: Taking the observations and theories of experimental science, the modeler attempts to duplicate the function of the system under study. In this way, the theories can be tested and refined. Ideally, we may build a model that suitably reproduces the observed phenomena, in which case there is strong circumstantial support to the contention that the theory upon which the model is based is in fact the basis of the biological prototype. More often we find that the models fail to accommodate the full range of behaviors we had considered, but this gives us the equally valuable information that our theories are mistaken or at least incomplete, pointing to the need for further research. In the course of creating these computer systems, we may also hope for spin-off benefits, for instance valuable sound-processing techniques for practical applications like speech recognition or sound enhancement.

The modelers' methodology, if that is the best term, is to take phenomena of human audition (from both psychology and physiology) and to consider their implications for a computational model – in the case of psychology, behaviors that the model should mimic, and for physiology, hints about how the model might work internally. The difficulty in addressing this goal comes from its breadth; if the auditory system was founded on a single principle, or served

one overriding function, it might be possible to make a clean and elegant model of it, one that perhaps approached some measure of optimality in the exactly-specified task. In some creatures we see more specialized functions that might perhaps be regarded this way – the fly-catching visual routines of the frog made famous in [Lettv59], or the echo-detection systems in bats [Suga90]. However, the entirety of the human auditory system is too sophisticated for a clean mathematical description founded on a couple of axioms of signal detection theory; it has evolved as a compromise general-purpose solution to the range of auditory tasks that people perform, from detecting very faint but steady tones, to making precise interaural timing judgments for localization of transients. At the same time, it has specialized to exploit many of the peculiar characteristics of the world of sounds in which we exist, such as the prevalence of common-period modulation, the kinds of reflections introduced by acoustic environments, and the basic constraints of existential continuity in the world.

As a result, a model that comes anywhere near emulating this not-purpose-specific but highly domain-adapted system is going to be a hodge-podge of different techniques working in combination. As in Minsky's famous quote (reproduced in [Mell91]), "the brain is a kluge," and any computer system capable of reproducing some of its more interesting behaviors is likely to be a similarly unattractive and seemingly arbitrary mixture of components. Would that it were not so; elegance is surely the ultimate aesthetic good in science as in art, but my chosen goal dictates otherwise.

This thesis, then, presents a collection of components assembled into a partial model that mimics some of the aspects of auditory information processing. The particular focus is on areas that I consider important to address at this stage in the development of such models – namely, the processing of sounds of all types rather than a limited subclass, and robustness in the face of obscured cues in different acoustic contexts. As such it is a mix of techniques from signal processing and artificial intelligence. I hope, however, that the processes presented will serve as useful solutions to the phenomena they address; a main goal of the thesis is to offer an example framework and a range of specific techniques that can serve as raw material for the future projects of the community of computational auditory modelers.

Another aspect of this thesis is the presentation a particular view of the problem: Different researchers have very different intentions when they say that they are building models of the auditory system; my research starts from a certain set of beliefs about what is important, interesting, valuable and feasible in this domain, and I want to present and justify these choices. It is also my hope to articulate a perspective on the history and background of work in this area that imposes some kind of structure and relation between the wide range of existing literature.

1.3 The prediction-driven model

This thesis describes a project to develop a computer system capable of analyzing a complex sound mixture into a set of discrete components. The intention is that these components correspond to individual sound-producing events that a listener would identify in the sound, and, moreover, that the system should arrive at this analysis by using the same features and information processing techniques that are at work in the human listener, at some suitably stylized level. Specifically, the characteristics of known structures in the auditory physiology are *not* the concern of this model; the

focus is at the higher, functional level of the psychoacoustic experience of distinct sound events.

I will present a framework for such a model which seeks to address some of the weaknesses of previous auditory scene analysis models (such as [Brown92]) in dealing with sounds that are not harmonically structured, and in dense mixtures where any given feature may have been arbitrarily corrupted. I will argue that all previous work in this area is based on a data-driven architecture, that is, within each time-slice a set of operations is applied to convert uniquely from the concrete input data to the more abstract output data, without any significant influence of the analysis thus far. In contrast with the data-driven approach, the current system performs an analysis that is strongly dependent on the predictions of an internal, abstracted model of the current state of the external acoustic environment. The system converts the acoustic signal of a sound mixture into a collection of generic sound-object abstractions. At each instant, the system has a prediction of the sound in the future based on this internal representation; assuming the model is correct and the external world behaves predictably, this prediction will match the actual sound observations, and the analysis proceeds with only minor parameter adjustment. However, when some unexpected sound-event occurs, it is detected as an irreconcilable deviation between prediction and observation, which triggers a major modification to the internal model, perhaps by the addition of a new element.

The principal advantages of this architecture over its precedents are:

(a) The **internal representation can be rich and over-general**, allowing the system to analyze and represent a full range of sounds. A data-driven system is unable to resolve the ambiguity of converting an observed sound into a space with many alternative representations for any given observation; the prediction-driven model is impervious to the fact that there may be other possible explanations of the sound it is observing; it arrives at a particular internal state by analyzing a past context, and simply confirms that the observations continue to be consistent.

(b) In situations where there is significant ambiguity concerning the representation of a sound, **alternative explanations can be developed in parallel** on the hypothesis blackboard at the core of the system, until such time as a reasonable choice can be made between them.

(c) By characterizing the sound objects in a **probabilistic domain** (i.e. in terms of expected properties and their predicted variance), the model can incorporate a range of sensitivities from the relative indifference to the fine structure of signals perceived as noise to the delicate sensitivity to isolated sinusoids.

(d) The **architecture is intrinsically extensible**, unlike the pre-ordained processing sequences of previous models. The blackboard system progresses through a sequence of 'problem-solving states' (for instance, excess energy has not been explained) then chooses from a repertoire of actions to resolve outstanding uncertainties [CarvL92a]. New rules, perhaps involving new cues, can simply be added to this repertoire and will then be employed as appropriate.

(e) A second dimension of extensibility is in the **explanation hierarchy**. The generic objects forming one level of this hierarchy can themselves be explained as the support for a more abstract hypothesis of some larger

structured event, which may itself support a still-higher hypothesis such as “my roommate is playing her Billy Joel record again.” While this aspect of highly abstract analysis has not been developed in the current work, it is an open door for future development within the same framework.

(f) Abstract representations are the ingredients crucial for **context-sensitive inference** for noisy or corrupted sound mixtures. When a particular attribute, such as the energy of a component in a certain frequency band, cannot be extracted locally owing to an obscuring sound, the object of which it is a part will still be able to make a reasonable guess as to its likely value, and the prediction-driven reconciliation will confirm that the guess is consistent with the overall observations. A prediction may have propagated down from a high-level abstraction, representing a more specific hypothesis about the cause of the sound, and thus capable of a more detailed prediction. In this way, the primary goal of the system, to analyze sounds correctly even when they are dense and corrupted, is intrinsically promoted by the architecture.

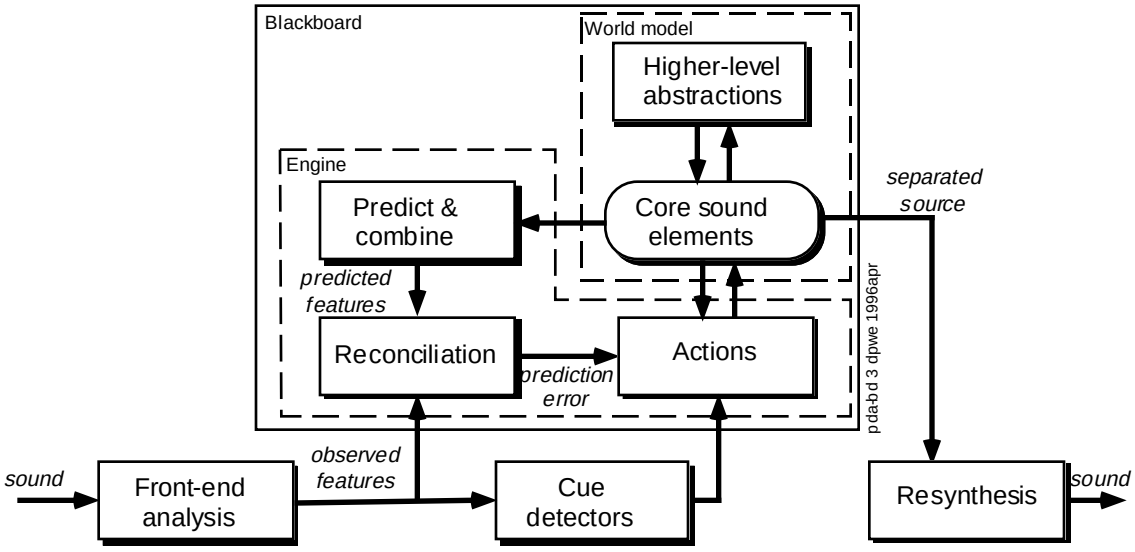

Figure 1.2: Overview of the goal, a computational auditory scene analysis system (described in detail in chapters 3 and 4).

1.4 Applications

What is the point of conducting research into this problem? The broadest motivation is intellectual curiosity, born of an increasing sense of awe as the full subtlety and sophistication of the auditory system is revealed. Although it is possible to imagine approaches to finding structure in sound that are obviously different from that employed in the brain, the task is hardly well enough defined to know if a given approach is ‘auditory’ or not until we have a much deeper understanding; my emphasis is specifically on models of the auditory system.

Apart from the basic quest of understanding the auditory system through functional modeling, there would be several practical applications for a usable sound organization module:-

- **Speech recognition systems**, and more generally sonic-aware robots, need to be able to make sense of their acoustic environments to separate the interference from the target or to recognize the arrival of new sources. It is ironic that there has been such success in the performance of speech recognition systems – an ability otherwise limited to humans – but that today's systems have great difficulty distinguishing speech from irrelevant intrusions – something that is probably quite easy even for cats and mice. In my opinions, this is the most serious indictment of the compact feature-space / hidden Markov model approach to speech processing.
- **Hearing prostheses**: One cause of interest and research in the area of auditory scene analysis is the frustrating inadequacy of conventional hearing aids. In spite of the restoration of lost sensitivity through amplification and dynamic range compression, hearing-impaired subjects still have difficulty separating mixtures of voices (e.g. [KollPH93]). An understanding of the basis by which this separation occurs might inform the construction of a processing system to enhance the relevant cues; or a module that can successfully separate out the different voices could be built into the hearing aid itself.
- **Multimedia indexes:** Ever increasing storage, processing and communication capacities are creating a world in which vast realms of information are available – but can we find the one piece that we want? Automatic indexes of the text-based portions of, for instance, the World-Wide Web are increasingly sophisticated and useful, but other media can only be searched through an associated manually-created textual description. An automatic sound-processing system capable of segmenting and classifying the elements of a movie soundtrack is required for useful automatic content-based indexing of these kinds of data [Hawley93] [Keis96].
- **Enhancement applications**: There are many examples of existing recordings which include unwanted interference - coughs in concerts, or recordings that have been degraded through age or equipment shortcomings. Listening to these recordings, we perceive separately the desired and unwanted components, and a sound organization system (that included a resynthesis capability) could remove the latter. Of course, a computer model of the auditory system would not be expected to reveal any *new* information that was previously inaudible to human listeners (at least until we understood the separation process well enough to improve upon it), but the cosmetic advantages of such cleaning-up would certainly be popular.

1.5 Ideas to be investigated

In its purest form, a dissertation is supposed to present a 'thesis' – an intellectual claim – and then investigate and support that claim. What, then, is the thesis of this work, a description of a partially-functional model of a incompletely-understood natural system?

Beyond the general idea that this is a useful collection of techniques for building such models, there are in fact a couple of fairly strong and perhaps slightly unusual positions behind this work. The first is the idea that perception proceeds via the indirect reconciliation of the internal representation with the perceived signal from the external world – the

motivation behind 'prediction-driven' analysis. This is in contrast to the direct, data-driven approach more usually adopted, where the information added to the internal model must be directly derived from the input signal, without provision for inference or guessing in the absence of more direct cues. Any number of illusory phenomena, as well as reasoning about the nature of perception in messy real worlds, support indirect inference as a preferable foundation for models of perception.

The second contention is a little more unusual: the idea that the full spectrum of 'real' sounds are adequately spanned by the combinations of a few simple parameterized sound primitives, and moreover that it is by decomposing real sounds into these primitives that the auditory system is able to analyze complex mixtures. This contention is implicit in the architecture of the system, which approaches sound analysis as a search for an adequate representation of the observed signal in terms of some generic primitives such as periodic tones and noise bursts. An alternative to this position would be to assume that there is no intermediate general-purpose representation of sound, merely the fully-detailed sound mixture and perhaps a collection of templates representing the exact, ideal waveform of 'known' sounds – a model-based analysis system where *every* distinct sound has a separate model. This detailed-model hypothesis is problematic, both because of the amount of storage it implies, but more seriously owing to the logical problems in determining if two sounds – perhaps different notes on the same instrument – should be represented as one or two models. Of course, these issues of classification are not entirely avoided by an intermediate level of generic representation, but such an arrangement would simplify the storage involved in each model, and it might also be amenable to hierarchical structure, with each model successively specialized by the addition of features to correspond to smaller and smaller categories within a particular class of sounds.

These are the two contentions of which I am most explicitly aware, and whose validity will hopefully be promoted by the results of this work. In building a model of a complex system there are myriad assumptions and implicit theories regarding every detail of the system to be modeled, most of which are unconsidered or forgotten by the modeler. One problem with pursuing the net result of so many hypotheses is the difficulty in assigning credit and blame among the assumptions and ideas; since my model cannot claim to be a fully functional model of human audition, do we reject the main hypotheses? Or do we give credit for the aspects of auditory organization that have been successfully reproduced to the nominated hypotheses, arguing that other simplifications in the model inevitably limited its ultimate scope? My inclination is to the latter, although I wish there were a more principled way to make this allocation.

1.6 Specific goals

A project in computational auditory scene analysis can go in many different directions. In this work the particular goals that were pursued, and to a greater or lesser extent achieved, are as follow:

- **Computational auditory scene analysis:** The broadest goal was to produce a computer system capable of processing real-world sound scenes of moderate complexity into an abstract representation of the sources in the sound as perceived by a human listener.

- **Dense ambient sound scenes:** The particular driving focus was on dense, noisy sounds like city-street ambience, since previous sound-analysis systems have not considered the information in such sounds, and since the particular issue of inharmonic sounds presented important new challenges.
- **Perceptual event output:** The highest level output should be a small number of relatively large-scale objects that correspond directly to distinct perceptual events in the sound. (Organization at certain highly abstracted levels, such as the formation of successive events into streams from a single source, was beyond the scope of this model).
- **Adequate sound representation and reconstruction:** To confirm the sufficiency of the representation, a resynthesis scheme was devised to generate perceptually acceptable reproductions of the represented sounds.
- **Assessment of scene-analysis systems:** The system is framed as a model of real audition; subjective listening tests were performed to test this assertion. The experiments were designed with a view to general applicability, and can be used for comparisons between different scene-analysis models. None of the assessment metrics used in previous models of auditory organization meet this need.

1.7 Outline of this document

The dissertation has six chapters. After this introduction, chapter 2 presents an overview of the field of computational auditory scene analysis and its underpinnings, including an interpretation of several major previous systems within a 'data-driven' framework. Chapter 3 presents the alternative approach of the current work, the prediction-driven architecture, which seeks to permit a far more context-sensitive analysis by basing analysis on predictions generated by an internal model. In chapter 4, the implementation of a system based on this approach is described in detail, ranging from the front-end signal processing through the internal representational elements through to the blackboard-based analysis engine. Chapter 5 presents the results of the system, firstly by looking at the behavior of the system in analyzing some sound examples, and secondly by describing the structure and outcome of the subjective listening tests. Finally, the conclusion in chapter 6 summarizes the project and considers how well it has achieved its goal of being a model of human audition.

Chapter 2 An overview of work in Computational Auditory Scene Analysis

2.1 Introduction

There is an emerging body of work in which researchers attempt to build computer models of high-level functions of the auditory system. The past couple of years has seen a surge of interest in this work, with workshops and sessions on this topic at numerous meetings, including the 12th International Pattern Recognition Conference (Jerusalem, October 1994), an Institute of Acoustics workshop on Speech Technology and Hearing Science (Sheffield, January 1995), the International Joint Conference on Artificial Intelligence (Montreal, August 1995) and the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (Mohonk, New York, October 1995). Despite all this activity, the essence of the field is sometimes hard to discern. This chapter is an attempt to define a scope for this work, trace its origins, and assemble within a single framework the various projects that have been completed. I will also discuss some major problems awaiting resolution.

2.1.1 Scope

What falls under the heading of Computational Auditory Scene Analysis? Arguably, the answer is anything that involves computer processing of acoustic data for a purpose similar to one performed by listeners. This ranges from models of the firing patterns of cochlear neurons to inference of metrical time-signature from streams of musical events. In the interests of brevity, however, I will adopt a more restricted interpretation limited to models of the process by which mammals use the sound energy reaching their ears to infer characteristics of external physical sources. Thus an empirical model of nerve firings that matches observations but does not consider why such phenomena exist falls outside this scope, since it does not address the *functional* role of the auditory system in organizing acoustic stimuli. At the other extreme, an abstract process such as time-signature inference relies on the distinct musical notes produced by the sound-organization function for its input, and can thus be seen as external to that function. However, the scope so defined is hardly narrow, covering everything from simple onset-detection algorithms (undoubtedly a very important part of auditory scene analysis) to complete sound-description systems that aim to 'explain' unrestricted real-world sound ambiences into symbolic representations reflecting the perceived source structure.

The remainder of this chapter is organized as follows: The next section considers the psychoacoustics of auditory scene analysis, the principal experimental basis for these computational models. Section 2.3 rounds out the background to this topic by considering related work in music and speech processing, and in models of the auditory physiology. Then, in section 2.4, I will present the core of the chapter, a synthesis of the work of Cooke, Brown, Mellinger and others into a 'standard' model of auditory organization, whose

qualities and limitations will be investigated in section 2.5. Following that, section 2.6 examines some of the more recent work and the particular problems researchers are currently seeking to solve. Finally, in section 2.7, I summarize the achievements of the systems described and speculate about the likely development of the field.

2.2 Foundation: Auditory Scene Analysis

Although people have been using computers to process sound ever since the emergence of digital signal processing in the 1960s, it was not until the mid-1980s that the idea of modeling some of the more abstract aspects of auditory processing emerged. This lag may largely be explained by our relative ignorance of auditory processes, reflecting a common problem in perception research: Much of what our perceptual apparatus actually does is implicitly invisible to us, and we have the greatest difficulty separating the concepts of external reality and our perception of it. Thus it was not until 1990 when Bregman published his book "Auditory Scene Analysis" [Breg90] that we had anything approaching a theory of the crucial step in auditory organization where continuous distributions of acoustic energy are converted into sharply delineated perceptions of distinct events. As an experimental psychologist, Bregman had been studying phenomena of perceptual organization in hearing for several decades, having realized in the late 1960s that, despite great successes in characterizing low-level detection and discrimination abilities, the sound domain had no work equivalent to the *ecological* (in the sense of [Gibson79], i.e. environmentally-relevant) problems that were attracting attention in vision, such as size and color constancy. However, it took two decades of research for a coherent picture of the kinds of processes and principles in operation to appear.

Bregman's book is the explicit foundation of all the modeling work described in this paper, as reflected in the most common title for the field, which simply prepends "Computational" to his phrase, "Auditory Scene Analysis" (itself an appropriation of the machine vision concept of "Scene Analysis", or an environmental explanation of a visual image). The book has had such an impact because, despite being a prudent account of empirically-observed psychoacoustic phenomena, it carries a strong sense of being close to a formal specification of the rules by which acoustic stimuli are converted into separate percepts. Bregman investigates different kinds of cues with clever experiments to set them competing against one another, permitting accurate modeling of such phenomena. However, the impression that there are rules simply waiting to be translated into computer code is mistaken; the biggest problem seems to arise in translating simple, isolated principles inferred from highly constrained psychoacoustic stimuli such as sine-tones and gated white noise, to the much, much messier domain of real sounds and sound mixtures. The human auditory system is known to be able to deal with complex real-world ambiences, and thus it is sensible and valid to investigate human hearing using simplified stimuli. However, if we were to build a computer system that did the 'right' thing in typical psychoacoustic tests, there is no reason to suppose it could then handle the real world. The terrible complexity of sounds in our physical environment presents a whole series of additional problems which aren't really addressed in the book.

In order to understand the computational systems to be discussed in this paper, let us briefly review the essence of Bregman's theory of auditory scene analysis. Early experiments with tapes of edited speech sounds forced him to

conclude that a high degree of organization is imposed on the sound we experience; without the 'correct' rate and sequence of speech cues, the same segments of sound changed in perception from a single, continuous speech stream into a collection of perceptually distinct pieces. In particular, the high-frequency noise energy of sibilants was no longer integrated with the lower-frequency vowel energy. This evidence implied the existence of a process at work in the auditory responsible for collecting speech sounds from a single speaker into the same stream, a function that could be defeated by the editing modifications.

Bregman's subsequent work sought further to investigate these organizational processes. His theory describes two broad classes of organization: The first is simultaneous organization, which is responsible for the *fusion* of acoustic energy that occurs concurrently in different frequency regions into a single percept. The most important example of this kind is the fusion of harmonically-related sinusoid (Fourier) components into a single 'rich' tone whose pitch and tonal color are related to the frequencies and strengths of the various component harmonics. This simultaneous fusion can be investigated by detuning one of the harmonics (which eventually causes it to be perceived as a separate sound [Hart88]) or by temporal asynchrony, that is by starting or stopping one of the tones at a slightly different time from the others. Fusion is strongly promoted by simultaneous onset; an onset asynchrony of a few tens of milliseconds can cause a tone to be perceived as separate from a tone-complex, even if the harmonic frequency relations would otherwise lead to fusion [DarwC92].

These fused auditory events governed by the cues of harmonicity and onset synchrony (and also common spatial location and common modulation) are then subject to the second class of *sequential* organization, where a series of sound-events is built up into one or more *streams*. Each stream is a sequence of events treated as coming from a single external source; events will be segregated into separate streams according to Gestalt-like principles of dissimilarity of pitch, loudness and timbre, implying that their chance relations at the ear of the listener are of no significance. Bregman suggests that this sequential organization might be accomplished by *schema* (remembered sequential patterns) which are learned through exposure to the environment, in contrast to the *primitive* simultaneous organization that might be pre-wired into the auditory apparatus.

It is intrinsically difficult to measure something as abstract and unobservable as the internal experience of sounds 'belonging' together. The success of experimental psychoacoustics stems from careful experimental design, where the properties of the resulting sound organization (the pitch of fused tones, the perceived order of streamed sequences) can be measured to infer the behavior of the organization process itself. The skill lies in devising experiments to distinguish between competing proposed mechanisms, such as the question of whether modulated harmonics are grouped by their common modulation pattern or their harmonicity [Carly91].

Bregman's explicit enumeration of the cues (related to physical attributes of the sound energy) and principles (rules governing the construction of perceived sources based on the available cues) furnishes a very attractive theoretical basis for researchers concerned with the processing and understanding of sound by computers. The implication is that the manner by which human listeners assemble useful, high-level information about the real world from the acoustic signals at their ears is more-or-less understood, and

the way is open to make computer implementations of these processes for a new generation of 'smart' sound understanding systems.

2.3 Related work

Before examining the kinds of computer models of the auditory system that Bregman's theory has inspired, it will help us to look at some other work in computer processing of real-world signals that formed the context of auditory scene analysis models. We will focus in particular on systems that seek to identify the more abstract information in sound signals (from both the musical and speech domains), as well as the more physiologically-motivated computer models of the auditory periphery. We will also consider briefly some of the analogous work in machine analysis of visual scenes, from which hearing researchers have borrowed many concepts related to abstractions of the real world.

2.3.1 Sound models

Before looking at computer systems that analyze sound mixtures, it is worth mentioning a few of the approaches to analyzing isolated, single sounds that have influenced this work.

Fourier analysis has long been an extremely powerful basis for dealing with sound, starting with Helmholtz's investigation of the harmonic structure of pitched sounds [Helm77]. However, the advent of digital computers and the fast Fourier transform (FFT) algorithm made possible a whole new approach to sound processing conducted in the narrowband frequency domain. Although applications of the phase vocoder (i.e. short-time Fourier transform representation) were first envisaged in the 1960s [FlanG66], it was not until the late 1970s that such algorithms became cheap enough for exotic algorithms such as timescale modification and pitch shifting [Port81]. A particularly flexible instance of this kind of processing is the Sinusoid Transform Coder of [McAuQ86], where pseudoperiodic sounds such as speech are represented as a collection of distinct sinusoid tracks following the individual harmonics extracted by narrowband FFT analysis. Sinusoidal representation provided not only for unusual transformations but also considerable coding efficiency in separating important and irrelevant information in the original signal.

Sinusoidal analysis was very applicable to the pitched sounds used in computer music, but it was less convenient for the non-periodic aspects of such sounds such as noise transients at the beginnings of noise or background 'breathiness'. This limitation was addressed in [Serra89] by modeling the short-time Fourier transform of musical notes as sinusoid tracks where such tracks were pronounced, and an autoregressive (all-pole) noise model for the remainder of the spectrum. This 'deterministic-plus-stochastic' decomposition was successful in representing these different classes of sound – pitch and harmonic balance for near-periodic sound, broad spectral envelope for 'noisy' sound. As will be seen in chapters 3 and 4, this assumed perceptual distinction between noisy and pitched sounds is deeply influential on the current work.

The concept of classifying regions of sound energy according to categories that reflect the perceptual distinction between pitched and noisy sounds was taken further in [Goldh92] which segmented the time-frequency plane according to the first three cepstral coefficients in the analysis of both time

and frequency profiles. This research was framed in terms of designing an aid for the deaf that would automatically classify or describe arbitrary environmental sounds in perceptually relevant terms. Cepstral coefficients (Fourier decomposition of the log of the magnitude envelope) nicely separate level, slope and finer detail (the interpretation of the three coefficients). Looking at the cepstrum both along frequency (the usual application) and along time defines a rich yet rather compact space for discriminating classes of sounds such as tones, bangs, rough and smooth noises.

2.3.2 Music analysis

Given the early successes of digital signal processing in revealing the structure of sound, and with the rise of computer musical instruments and performance systems, it seemed natural to imagine a computer system that could listen to a piece of music and produce from it a description of the instruments involved and the notes they played. But like the modeling of other perceptual tasks (e.g. speech recognition, visual object recognition), this goal of 'polyphonic pitch tracking' and automatic transcription turned out to be far harder than had been anticipated. An early system that refused to be discouraged by these difficulties is described in [Moorer75]; it does a reasonable job of deriving a score from recordings of certain kinds of musical performance; however, there are serious constraints on the instruments involved, and the structure of the music they play.

The domain of automatic transcription continued fascinate researchers. The extreme difficulties arising from the collision between Fourier components in common polyphonic music (an inevitable consequence of the conventions of western harmony) led to interest in the transcription of non-harmonic music such as drum or other percussion performances. Early work by Schloss [Schlo85] (who detected drum sounds via the slope of a sound's energy envelope) and Stautner [Staut83] (who recognized the particular value of a variable-bandwidth filterbank in detecting the rapid transients of these sounds) was followed by a recent project by Bilmes [Bilmes93] who built an array of special-purpose detectors for the different instruments in an Afro-Caribbean percussion ensemble, as a precursor to an automatic music interpretation system.

The thornier problems of separating harmonically-structured instrument sounds also attracted a series of projects. The Kansei system of [KataI89] sought to translate recordings of music all the way to textual descriptions of their emotional impact, solving numerous difficult issues along the way. [Maher89] addressed the problem of harmonic collision directly by analyzing recordings in the domain of sinusoidal models (where each extractable harmonic is a distinct entity) and devising algorithms to make a best guess of the note or notes that would result in that pattern. [Baum92] employed a similar approach, using a constant-Q sinusoidal analysis (the Frequency-Time Transform of [Hein88]) to convert musical recordings to MIDI note-event streams. The full auditory-scene-analysis model of [Mell91], described in section 2.4, has this approach at its core.

However, the continued elusiveness of this problem demands a more resourceful approach. Starting with the piano-transcription problem formulation of [Hawley93], [Scheir95] exploits the additional prior knowledge of the original score to guide the extraction of the precise parameters of a given performance. He makes the point that this is an instance of a more general class of systems that use musical knowledge to guide the performance

extraction process, which is exactly the philosophy of the 'signal and knowledge-processing' system of [Kash95], which is discussed in section 2.6.

Despite all the attention attracted by this problem, even the best systems only work moderately well in limited domains. The idea of a machine that can convert a recording of a symphony into the printed parts for an orchestra, or a MIDI encoding for storage and resynthesis, remains something of a fantasy.

2.3.3 Models of the cochlea and auditory periphery

An obvious starting point for modeling the auditory system is to look closely at the biological original to see what may be learned. Our understanding of the auditory periphery to the level of the hair-cell synapses and beyond has been steadily increasing over the past four decades. Improvements in experimental techniques, progressing from studies on cochleae removed from cadavers [vBek60], to auditory-nerve recordings from anaesthetized cats [YoungS79], to direct observations of basilar membrane motion in live gerbils [Zwis80], have yielded a wealth of information for modelers to attempt to explain.

The cochlea is probably the single most critical component in the auditory chain. After coupling to the acoustic free-field via the outer and middle ears, the one-dimensional sound pressure fluctuation is applied to the oval window, which starts a traveling wave down the spiraled transmission line of the cochlea. The variation of the mechanical structure of the basilar membrane – the central division of the cochlea – effectively forms a continuous array of band-pass filters; Fourier components in the pressure variation will travel some distance down the cochlea (further for lower frequencies) before reaching a point where the membrane is in resonance, causing a maximum in basilar membrane motion and the dissipation of the traveling wave. Thus, the cochlea performs a spectral analysis, converting the incident sound-pressure variation into motion at different places down the basilar membrane, with the place of motion encoding spectral location and amplitude of the motion indicating intensity.

However, the precise behavior of the cochlea is rather subtle and subject to debate [AllenN92]. Consequently, no single mathematical model has been adequate, but rather a number of models have been proposed, each with its own particular strengths. On the assumption that the most important quality of the cochlea is its behavior as an array of band-pass filters, the Patterson-Holdsworth model [PattH90] presents simple four-pole resonator approximations to this filtering as a good compromise between computational simplicity and accuracy of fit to observed tuning curves. The Lyon model [SlanL92] also reproduces this filtering behavior, but as a result of a more accurate transmission-line model of the physics of the cochlea. This is followed by a laterally-coupled stage of automatic-gain-control, to account for the very wide range of intensities over which our hearing can operate. Still more sophisticated models of the cochlea acknowledge that the resonant tuning of the cochlea appears to become progressively more damped as the intensity increases, something a fixed resonant structure cannot exhibit. The model of [GigW94] incorporates a damping-feedback component, intended to model the function of the outer hair-cell bundles of the basilar membrane (which are enervated by nerves carrying information *out* from higher neural centers) to vary the damping of the peripheral model in response to the signal. These ideas have recently been incorporated into Patterson's original model [PattAG95].

The motion of the basilar membrane is converted to nerve firings via synapses at the base of the inner hair cell bundles. This transduction forms a crucial component of any model of the auditory periphery, since it is the firing patterns on the approximately 30,000 fibers of the auditory nerve that comprise the raw description of sound used by higher levels of the brain. Ranging from simple integrated nonlinearities followed by threshold-discharge models [Ross82] to detailed models of neurotransmitter depletion, a detailed comparison of a range of nerve models with known firing behaviors appears in [HewM91]. Analysis of these systems is difficult, because the information-carrying-capacity of a single nerve is rather limited; the brain accommodates this by using a very large number of them and combining their outputs. However, running tens of thousands of nerve models in parallel is prohibitively expensive computationally, favoring approximations to ensemble behavior instead (such as firing *probability* for a certain class of fiber, as opposed to actual firing patterns for a number of instances of that class). Our ignorance of how the nerve information is combined in the brain makes this an uncertain business.

Looking further along neural pathways, there have been several studies of cells at higher brain centers including the Cochlear Nucleus, focusing particularly on their response to modulated tones. An interesting behavior of certain of these cells is their selective transmission of certain modulation rates (in the range 5 to 500 Hz), i.e. they can act as bandpass filters for intensity modulation in different frequency bands [Lang92]. Models of this behavior, based on simplifications of the biophysics of neurons, have been produced by [HewM93] and [BerthL95]. Such behavior provides evidence for the existence of a two-dimensional array or 'map' of neurons, where each element responds most vigorously to a particular combination of peripheral frequency channel (one dimension) and amplitude modulation rate (the second dimension); this is exactly the kind of map required for the pitch-based organization scheme of [Brown92], discussed in section 2.4.

Physiological and neural models of this kind can stand purely as predictors of experimental observations. Our goal, however, is to understand how these mechanisms facilitate hearing. A few of the models have been taken further to address functional questions of how the ear performs useful tasks, for example, the separation of a mixture of pitched sounds, and we will return to them in the next subsection.

2.3.4 Speech processing and pre-processing

As observed in the introduction, by far the most thoroughly-researched area of sound processing is the recognition of speech signals. This is because of the enormous practical attractions of controlling machines by spoken commands. Despite early systems based on explicit theories of human information processing [LessE77], more recently the dominant approach has tended towards a 'brute-force' statistical approach, asking only that the representation distribute speech fragments fairly uniformly around a low dimensional space [SchaeR75], and employing the powerful technique of hidden-Markov-models to find the most-probable word or phrase to account for a given speech recording [Jeli76]. Despite its success, such research does not belong in this survey because the only aspect of human audition reproduced is the abstract and narrow task of converting speech sounds into words. There is no real effort to duplicate the auditory function at a deeper level, and, crucially, such systems tend simply to ignore nonspeech sounds, making the working assumption that the input is clean, closed-mic'ed voice.

Of more relevance to the development of models of auditory scene analysis is research into the separation of target speech from unwanted interference. This work has been motivated both by the desire to clean-up transmitted or recorded speech prior to presentation to a listener, and as a possible pre-processing stage for speech-recognition machines. Indeed, this concept of an automatic-speech-recognition aid has been the justification for much of the work in computational auditory scene analysis [Wein85] [CookCG94], which reflects the potential value of such a device, as well as the benefits of working in a problem domain where achievement can be quantified (as a reduction in word-error rate).

The problem of separating simultaneous, pitched voices has attracted attention as a well-formed, seemingly tractable problem (similar to the polyphonic music problem, but perhaps easier since harmonic collisions are short-lived and less common). Early projects by [Pars76] and [HansW84] sought to completely identify the Fourier spectrum of the target voice, to permit uncorrupted reconstruction, an approach which has been considerably refined in the work of [DenbZ92]. By using an explicit sinusoidal track model, [QuatD90] were able to reconstruct voicing through pitch-crossings (which would otherwise play havoc with extracting the spectrum) by identifying ill-conditioned time-frames, and interpolating the frequencies and magnitudes of each component across the gap.

The objective of the above models was processing and reproducing a speech signal to remove interference. In contrast, the problem of mixed, pitched speech, framed as 'double-vowel perception', has also been addressed by researchers concerned with faithful models of the human auditory periphery. A series of papers from the Institute for Hearing Research in Nottingham and associated researchers [SummA91] [AssmS89] [StubS91] [CullD94] has considered various physiologically-motivated representations and how they might fare in identifying combinations of pitched vowels; these models can then be compared to the results of psychoacoustic experiments where real listeners attempt the same tasks. The recent models of [MeddH92] and [AssmS94] have been particularly successful at duplicating the recognition rate for vowel-pairs as a function of fundamental-frequency difference, providing some validation of their particular theories of sound separation by pitch difference; the [MeddH92] model relies on autocorrelation peaks within each frequency channel to gather energy related to a given vowel, and the [AssmS94] system extends the model to consider the particular effects of time-separation of individual pitch pulses from the two voices. A class of similar approaches to separation of concurrent periodic sounds based on neurally-plausible delay-and-combine structures was thoroughly investigated in [deChev93].

Ghitza has also examined the possible benefits of models of the auditory periphery for speech recognition systems. [Ghitza88] describes how an automatic-gain-control, modeled on known physiology, can benefit the performance of a speech recognizer when confronted with noisy speech. In [Ghitza93], the performance of a standard hidden-Markov-model recognizer is assessed in a highly-constrained task which should be dominated by the adequacy of the representation; an auditory model, based on the number of hair-cells firing at a particular interval, performs better than the more common cepstral front-end, but neither approaches the performance of real listeners in the same task.

Along similar lines, Lazzaro has recently investigated the use of physiologically-motivated front-end processing for conventional speech

recognizers. The use of "silicon cochleae" to calculate spectral decompositions, autocorrelations, and onset features is motivated towards finding small, low-power, real-time alternatives to the costly and complex digital signal processing of typical auditory models [LazzW95]. Lazzaro makes a clear analysis of the mismatch between unevenly-sampled, highly-redundant auditory model features and the regular, uncorrelated coefficients best suited to current pattern-recognition technologies [LazzW96]. He notes that alternative recognizer structures (such as neural-net classifiers) might ultimately realize the latent processing advantage of such plausible front-ends, which to date have shown only modest improvements at best.

2.3.5 Machine vision scene analysis systems

As observed in the introduction, many of the basic concepts of the psychology of auditory scene analysis were inspired by or borrowed from theories of visual perception. Similarly, computational models of auditory organization cannot help but be influenced by the considerable body of work concerned with modeling the analysis of visual scenes. Minsky relates how it was his casual assignment of image-object-extraction as a summer project for an undergraduate in the early 1960s that first exposed the nascent artificial intelligence community to the difficulty of perceptual processing problems, and the insight that "easy things are hard" [Minsky86].

Visual scene analysis continues to be a major preoccupation of artificial intelligence research, both because of its obvious practical benefits (e.g. for robots useful in manufacturing) and because of the undiminished intellectual mystery of how it may be achieved. One highly influential contribution to this work was Marr's book *Vision* [Marr82], which presented one of the first viable theoretical foundations for computational models of perception, seen as the process of converting raw stimuli into 'useful' information about the real world. Marr stressed the importance of distinguishing between different levels of analysis of perceptual systems; in particular, he identified three distinct levels for any information-processing task: At the lowest level is the *implementation*, which is a description of the actual physical elements employed to perform the computation; work on the biology of vision tends to fall in this area. Above implementation is the *algorithmic* layer, an abstraction of the calculations performed by the implementation, which could be equivalently performed by different hardware. Theoretical analysis of vision, such as modeling the retina as an array of spatial filters, falls into this level. Marr adds a third, higher level, the *computational theory*, which concerns the fundamental question of what the system is really trying to do; in the case of vision, what interesting aspects of the external world are available in visual information and would therefore be worth computing. Marr argued that this level of analysis was almost completely absent from work in vision at that time, leading to inevitable confusion; in order correctly to abstract the computational behavior from an implementational instance it is necessary to have a good idea of the overall purpose or goal of the system; else, there will be no basis upon which to distinguish between the essential purpose of the system and irrelevant artifactual aspects of the implementation. Marr's ideal form for a computational theory is some mathematical expression of the physical facts of the situation, such as the dependence of image intensity on reflectance, illumination, geometry and viewpoint. Starting from such a formulation, algorithms may be proposed to extract the interesting information (the separation of the four contributing factors), and then implementations can be devised for those algorithms. Without a computational theory, the entire analysis is rootless. Modelers of

computational auditory scene analysis, starting with Cooke [Cooke91] have acknowledged their considerable debt to Marr.

More than a decade after Marr's seminal work, the vision problem remains unsolved, and naturally his approach has been criticized. A recent theoretical development [ChurRS94] argues that it is a mistake to view vision as some abstract task of creating a neat internal symbolic model of the world independent of the immediate goals and tasks faced by the organism. Logically, evolution would have optimized visual processing to obtain only the information needed at any particular moment. Thus, the correct way to approach machine vision might be to focus on the completion of a particular task, such as catching prey, or avoiding predators or collisions [Woodf92]. [ChurRS94] make this case very powerfully for vision, and [Slaney95] points out the direct analogies to work in audition. [Brooks91] argues in some detail that the only way to solve problems of perception (as well as locomotion, planning etc.) is to build real robots that exist in the real world, and whose information-processing, rather than being neatly abstract and symbolic, is largely implicit in the emergent behavior of goal-specific sensor systems.

Despite these counter-arguments, there is still currency to the idea of a 'pure perception' system whose function is to produce a general-purpose symbolic description of the external world. There are no insurmountable problems in formulating a research project in these terms, and it is the goal, implicit or explicit, of a number of current efforts [Ellis95a].

2.4 The data-driven computational auditory scene analysis system

The central part of this paper is a unified description of several similar systems, representing the most direct, complete, and ambitious projects to implement computational models of the mechanisms presented in Bregman's book. These systems are described in the theses of Cooke [Cooke91], Mellinger [Mell91] and Brown [Brown92], and in the paper [Ellis94]. Each one of these systems sets out to duplicate the organizational function of the auditory system with a view to separating mixtures of sounds – speech plus interference for Cooke and Brown, ensemble music for Mellinger.

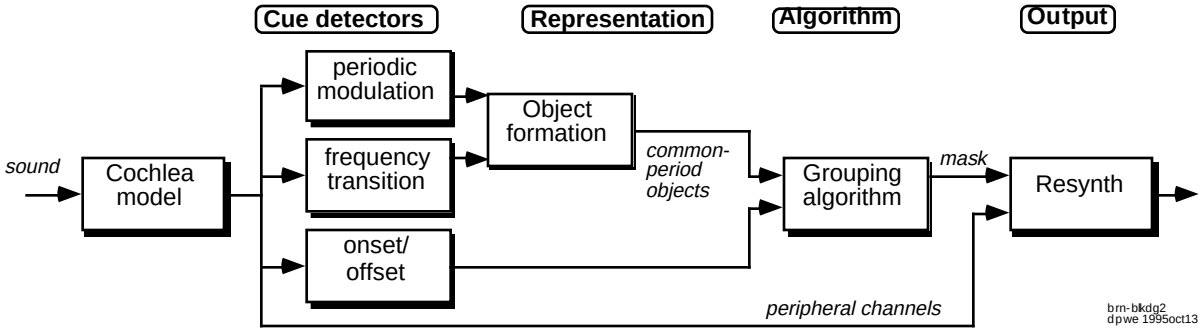

Figure 2.1: Block diagram of a typical data-driven sound analysis system (based on [Brown92]). The system consists of four stages: cue detectors, representation, grouping algorithm and output, and information flows exclusively from left to right i.e. from more concrete to more abstract.

It makes sense to consider these systems all at once because they share remarkable architectural similarities. I characterize these systems as 'data-driven', since the information flow is exclusively unidirectional, from concrete

to abstract, as schematized in figure 2.1. According to the diagram, the overall system structure can be broken into four main pieces:

(1) **Front-end**: A model of the auditory periphery, most importantly its effect of breaking the acoustic signal into different frequency bands. This may also include special-purpose processing to reveal particular acoustic features or 'cues'.

(2) **Basic representation**: The unstructured data coming from the auditory filterbank is organized into discrete elemental units, segmented on the basis of local coherence of low-level properties. Taken together, these elements account for the entire input sound; their discrete nature permits the calculation of useful attributes, and forms the fundamental granularity of the scene analysis.

(3) **Grouping algorithm**: The core of the system is an implementation of Bregman's grouping principles, to collect together the appropriate subsets of the elements that appear to come from single sources, based on the information of their properties and the other cues extracted in stage (1). The output of this stage is one or more groups of the elements generated in stage (2).

(4) **Output assessment / resynthesis**: The internal representation of the groups from stage (3) must be converted to some output representation appropriate to the goal or assessment method of the system. In some cases, this involves resynthesis of an acoustic equivalent to the detected source.

Each of these stages is considered in more detail below:

Front end: The input signal is processed by some equivalent of the auditory periphery up to the level of auditory nerves. In the most stylized system, this is simply a linear, constant-Q filterbank followed by intensity envelope extraction [Ellis94]. In the more detailed models, it may include filtering for the outer ear transfer function, a cochlear filterbank, and some kind of auditory nerve simulation. Cooke and Brown both use the gammatone filterbank as proposed in [PattH90], whereas Mellinger uses the transmission-line model of [SlanL92], whose output is nerve firing probability for each frequency channel. Brown uses the [HewM91] model of nerve-cell neurotransmitter concentration to derive firing probabilities.

In addition to indicating the time-frequency location of acoustic energy, the front-end may also incorporate specific feature detectors. While Cooke and Ellis derive their equivalents of the psychoacoustic cues of onset and harmonicity from their intermediate discrete representations, both Mellinger and Brown have cue detectors operating directly on the unstructured filterbank output to generate feature 'maps', an array, typically in the time-frequency plane, indicating the likelihood of a particular cue at each coordinate. This approach is partly motivated by the known use of parameterized maps in neural processing systems [Suga90]. Mellinger's maps detect energy onset and frequency variation; Brown has a very similar onset detector, uses a somewhat comparable frequency-transition map, and also includes auto- and cross-correlation to produce maps used in his periodicity (harmonicity) representation, discussed below.

Representation: The data coming from the front-end model is still somewhat 'raw'. In particular, it does not have any structure associated with it; ultimately the task of these models will be to assemble regions of time-frequency that contain energy judged as belonging to a single source, so some

division of the sound into a collection of distinct elements is required. These representational elements will then be formed into groups in subsequent processing to assemble all the energy believed to belong to a single external source. The desirable properties of these elements are that (a) each element should consist as far as possible of energy from only a single source, otherwise the grouping will be unable to allocate it unambiguously, yet (b) that each element should be as large as possible, and the total number of elements be as small as possible. Having 'large' elements (i.e. elements including information from relatively large portions of time-frequency) permits the calculation of informative attributes such as frequency variation or amplitude stability which cannot be calculated for a single sample of time frequency; these properties can then be used as the basis for grouping. Having relatively few elements is the logical consequence of making each element larger (assuming no overlap), and has the benefit of reducing the total computation required to allocate each element into a group.

The criterion for the formation of these fundamental representational elements is local coherence of low-level properties. In Ellis's system, the elements are *sinusoidal tracks*, fit to the filterbank output by picking spectral peaks, and grown or terminated based on simple smoothness conditions [EllisVQ91] [Ellis92] (a representation borrowed from the fixed-bandwidth speech-analysis system of [McAuQ86]). Similarly, in Mellinger's model, the first level of discrete representation is the *partial*, intended to correspond to sinusoidal Fourier components for the musical sounds he considers. Partials are formed when sufficient activity in the onset map coincides with a local maximum in an energy spectrum, which is tracked along time until it disappears.

Cooke's system uses *synchrony strands*, which are notionally almost identical (each being characterized by a frequency and magnitude contour defined over the time support), but are calculated rather differently. Ranges of cochlea-model channels whose outputs have high correlation to their neighbors, indicating response to a common dominant signal component, are grown into synchrony groups, which are then tracked through time. Brown's 'auditory objects' are intended to provide the same level of granularity (one representational element per distinct Fourier component in the lower spectrum, or formant track in the upper spectrum) and are derived in a related but enhanced manner where a range of peripheral frequency channels whose running-autocorrelations are highly correlated at a given time instant are recruited into a single element. The evolution of this element along time is then guided by the frequency-transition map which indicates the direction of broad spectral shifts in the acoustic energy.

Grouping: Once the sound has been converted to a collection of discrete, atomic elements, each with its own properties, which together form a nonredundant account of the entire sound, the grouping algorithm can be applied. In these models, this algorithm has the goal of organizing the elements into one or more distinct groups, where the elements in each group correspond to all the acoustic evidence from one 'perceived' source. Following auditory scene analysis, this grouping is made on the basis of cues derived from the elements (such as onset time), which lead to group formation according to particular principles (such as fusion by common onset).

In keeping with the data-driven philosophy of these models, the algorithms are devised to operate in a single pass (i.e. without revision or iteration), employing heuristic strategies to accomplish near-optimal grouping without backtracking according to some within-class similarity metric. Cooke's

system is presented as two stages. The first forms groups of his strands based on the cues of harmonicity (i.e. frequency contours that are integer multiples of a fundamental) and correlated amplitude modulation – needed for the grouping of formant tracks, whose frequency contour follows the formant rather than the voice pitch, but whose amplitude contours exhibit pitch-rate modulation. It uses a 'greedy' strategy of working through all the strands in descending order of total energy, forming a group for each one by (a) adding groups that overlap in time with the current strand if they have sufficient similarity according to the cue in operation, (b) choosing a new 'current strand' from among the newly-added strands, such that its time-support extends beyond the time range considered so far, and (c) repeating until the time-range is not extended. Groups are formed in this way for *every* strand; inclusion in a group does not remove a strand from subsequent processing.

His second grouping stage takes all the groups formed by the first stage, then *subsumes* all the groups with high degrees of overlap into a single, large group of which they are all subsets, drastically reducing the total number of groups and removing redundancy among those groups. He also applies a stage of pitched-based grouping, where pitch contours are calculated for the large groups, and then any groups whose pitch contours match over a significant time interval are merged. (Pitch contours are derived from the fundamental frequency for the harmonic groups, and from the amplitude-modulation period for the common-modulation groups.) This is the only way in which the low-frequency resolved harmonics for voiced speech (grouped by harmonicity) become connected to the higher-frequency formant tracks (grouped by amplitude modulation). Any speech detected in the original signal will have formed at least one of each of these groups, which are eventually joined only at this late stage.

Ellis's grouping algorithm similarly has two stages, where the first generates a large number of possible groups of his 'tracks', and the second prunes and corroborates to improve the confidence and robustness of the ultimate output groupings. His first stage applies four grouping principles in parallel, generating groups based on harmonicity, common onset, continuity (reconnecting tracks with short breaks) and proximity (intended to group together the many short tracks arising from non-harmonic noise energy). His second stage combines these in specific ways, looking for sets of tracks that were similarly grouped by both the harmonicity and onset groupers (indicating well-articulated pitched sounds) and merging together continuity and proximity groups that overlap with the same common-onset group. Like Cooke's, this system suffers from the separate treatment of resolved and unresolved harmonics of the same voice.

This problem is nicely addressed in Brown's system. He includes an autocorrelation-based periodicity map that can reveal the common period of both resolved and unresolved harmonics. His grouping scheme first calculates a *pitch contour* attribute for each of his elemental objects (resolved harmonics or single formant tracks). This is accomplished by combining the autocorrelations for the range of frequency channels included in the object to form a summary short-time autocorrelation as a function of time. Typically, this will have its most dominant peak at a period reflecting the center-frequency of the channel from which it is derived. However, he then weights this per-object autocorrelation with the global summary autocorrelation for the unsegmented sound. Since averaging across all frequency channels will favor the common fundamental period of pitched signals, this weighting effectively picks out just that autocorrelation peak that matches a plausible

fundamental period for that time frame, and thus the appropriate subharmonic is selected as the pitch of high harmonics. This is clever because it successfully attaches a context-sensitive property (the underlying pitch-period) to a locally-defined element (the harmonic or formant track) without explicitly deciding *a priori* what pitches are present in the sound.

With each object labeled by its pitch, the process of grouping is as follows: Starting with the longest as-yet unexplained object, a group is built for it 'greedily' by adding any objects remaining unexplained that pass a similarity test with every object in the expanding group. Similarity is scored by closeness of pitch contour during time-overlap, and also includes a term to favor the grouping of objects with common onset, as judged by the presence of simultaneous energy in the peripheral onset-map at the start of both objects. Since objects are removed from further consideration once they have been added to a group, Brown's system implicitly applies the psychoacoustic principle of *exclusive allocation* (the tendency for acoustic energy to be associated with only one perceived source), and does not require a subsequent pruning or subsumption stage.

All these systems process an entire segment of sound in *batch mode* (i.e. after it is completely known). In contrast, Mellinger frames his algorithm as *incremental*, recognizing the very real requirement of the auditory system to provide a best-estimate of the correct organization at every instant in time. He updates the hypothesized groupings of the current harmonic elements at each time step based on their accumulated resemblance; an *affinity* score is maintained for every pair of harmonics which is initialized according to their onset synchrony, then updated subsequently to reflect the coherence of their frequency variation. Perceived sources are then simply sets of harmonics whose mutual affinity scores exceed some threshold. This 'real-time' grouping algorithm has the interesting property of 'changing its mind' as evidence accumulates: Mellinger uses the particular example of the Reynolds-McAdams oboe [McAd84], where progressive frequency modulation of just the even harmonics causes the percept to bifurcate from a single oboe note to the combination of a clarinet-like tone and a voice singing the same note an octave higher. This is exactly the analysis achieved by his system, initially grouping all the harmonics into a single group, then abruptly revising this to two groups as the frequency-variation mismatch pushes down the affinity between the odd and even harmonics.

Output: Once grouping has been performed, there is the question of how to use the analysis. All the systems under discussion are theoretical proof-of-concept investigations rather than solutions to specific problems, so the form of the output is not rigidly defined in advance.

Each system identified subsets of the input signal believed to originate in a single source; an obvious representation for these would be as an acoustic signal – a resynthesis – consisting solely of the identified energy. However, this is a difficult thing to produce, since the analysis is usually applied at a simplified or under-sampled level of representation. For instance, Mellinger's system operates at a basic 'tick rate' of 441 Hz (2.27 ms), which is generous when one's objective is the location of perceptually important events in onset maps, but quite inadequate for capturing the full details of modulation information in the kilohertz-wide upper frequency bands. Batch-mode computational models generally have the option of returning to the original source data if they find that more information is needed than was gathered in the first pass, but it is somewhat worrying if satisfactory resynthesis requires

information that was apparently ignored in the structural analysis of the sound.

Instead of trying to resynthesize sound, it may be preferable in many applications to generate a more abstract description of the identified sources. If an auditory organization system is intended to convert a musical performance into control information for a synthesizer, it would be much more valuable to have the abstract parameters from the grouping algorithm concerning the fundamental period used to group each source than to be given a carefully resynthesized acoustic signal. In the scenario of computational auditory scene analysis as a pre-processor for speech recognition, the 'confidence rating' for different portions of the detected target might be very useful to the speech recognition algorithm, but cannot easily be expressed in a resynthesis (Cooke *et al* have indeed developed modified speech recognition algorithms to exploit the extra information coming from their scene analysis systems [CookCG94]). Thus the most appropriate output entirely depends on the application.

That said, both Ellis and Brown have addressed the resynthesis of audio signals. Ellis goes to some lengths to hold enough information in his 'tracks' representation to permit good-quality resynthesis, including a variable sampling rate to accommodate the higher bandwidth of information in upper frequency channels. An output signal can be synthesized based on the information in the final grouped elements alone. Brown uses his grouped objects to generate a time-frequency 'mask' that indicates the regions where energy for the extracted source was detected. By re-filtering the original input mixture using this mask, energy that occurs in time-frequency cells not assigned to the source is removed. However, where energy from both target and interference occur in the same cell, they cannot be separated. He does not pursue resynthesis based on the grouping system's representation alone (i.e. without referring back to the original input signal).

Assessment: Each author had to face the question of assessing the performance of their systems, yet this brings up a peculiar problem of the auditory scene analysis domain: The models are trying to duplicate the operation of an internal perceptual process, but we cannot directly access that process' output to compare it with our systems. Rather, we investigate human auditory organization by psychoacoustic experiments – asking questions of real listeners. One approach to assessment would be to construct models that were effectively capable of participating in similar psychoacoustic experiments, although that might require the modeling of many additional aspects of the brain. Also, our goal is not really a computer model capable of organizing the simplified acoustic stimuli typically used in psychoacoustic experiments; such sounds are suitable reductionist tools for investigating a general-purpose auditory system such as a human listener, but their simplicity would be misrepresentative if used in the assessment of less capable models. Much of the difficulty of computational auditory scene analysis would be eliminated if the sonic world could efficiently be reduced to a few, stable sine tones; our interest lies in a system that can deal with noisy, complex, real-world sounds that are less amenable to neat psychoacoustic tests.

If we were to assume that the scene analysis system does an exact job of distinguishing sounds from independent real-world sources, we could assess the systems by constructing an acoustic mixture from two separately-recorded sounds (say speech and some background noise) and observing how closely the model reconstructs the originals from the mixture. Current

auditory organization systems tend to fare very poorly by this criteria, partly due to the problems with resynthesis mentioned above, but at a deeper level reflecting the weakness of the initial assumption: The internal perception of an individual source is not an *exact* representation down to the waveform level, but rather a highly selective perceptual impression. A fairer approach to assessment would compare the original and reconstruction on the basis of 'perceptual similarity' i.e. correspondence of perceptually important features; unfortunately, there are no reliable objective metrics to score this resemblance. The best available option is to conduct subjective listening tests, but this has been avoided by previous modelers, doubtless due to the logistical disadvantages of human subjects compared to computer-based objective measures; a firm psychoacoustic grounding has been sacrificed in the interests of a consistent, objective score that can be calculated repeatedly to compare and measure the evolution of a system.

The idea that a sound processing system might be doing a reasonable job even if its rms error figures looked terrible arose somewhat naturally in the speech recognition community, where the only important part of the sound was the sequence of words being conveyed, and all other attributes of the signal (such as whether it sounded 'clean' or 'noisy') were irrelevant so long as they did not influence the speech recognizer. Thus in Weintraub's speech separation system [Wein85], he assessed its performance by taking the reconstructed, separated voices, feeding them into a speech recognizer, and using the conventional metrics of recognition accuracy to rate the system. The weakness of this approach lies in the fact that a speech recognition system is a poor model of auditory perception, highly specialized to a particular domain. Thus, certain artifacts in the output of the source separation system may disturb the speech recognizer far beyond their perceptual significance (static spectral distortion, for instance), while at the same time the source separation system may introduce perceptually appalling distortions that the speech recognizer blithely ignores. Weintraub's results were inconclusive, with some mixtures having better recognition rates before the separation step.

Several of the smaller models of specific aspects of speech separation have devised their own reduced versions of 'salient feature recognition' to permit assessment. For example, the two-vowel separation system of [MeddH92] deals only with static vowel signals; they assess it via the error rate of a static-vowel identifier constructed specifically for that situation which classifies reconstructed sound as one of five vowels according to the first few bins of the cepstrum (i.e. the broad spectral shape). Thus a highly-reduced but task-adapted assessment metric highlights the successes of their approach, but is less useful outside their limited domain.

Cooke was interested in a system that could extract the full sound of a voice in a mixture, i.e. more than just its linguistic content [Cooke91]. His assessment was to compare the output of his separation scheme with the original isolated sounds used to form his mixture examples. However, lacking a good resynthesis scheme, he made the comparison in the domain of his 'synchrony strands', made possible by the solution of a thorny correspondence problem : He calculated representations of both original sounds and their mixture, then derived an 'optimal' correspondence between the mixture elements and each of the input sounds by allocating every strand from the mixture to whichever isolated source contained a strand most similar. His metrics then rated how well his grouping system had been able to reproduce the optimal segregation. This scheme gave useful results in terms of indicating which examples were most difficult, but it is intricately

tied up with the details of his system, precluding its use for comparison with other sound organization systems. It also ignores the problematic cases where his strand-representation itself is inadequate to permit separation (such as the common situation of spectral overlap).

Brown's time-frequency mask did provide for high-quality resynthesis. He assessed his system by defining metrics at least notionally corresponding to the target-to-interference ratio in his final output sounds. He faced a correspondence problem similar to Cooke's in terms of allocating the energy in his resynthesized signal to one of his input components, since the resynthesis is essentially a new signal with no intrinsic marking of the origin of each piece of energy. However, as his resynthesis scheme amounts to a linear (albeit time-variant) filtering of the input mixture to recover each of the original sources, he was able to apply this filter separately to each original signal to obtain 'target' and 'interference' components that sum exactly to his resynthesis. By comparing these two contributions, he could produce target-to-interference ratios both for the output as a whole and for each individual time-frequency cell. This clever scheme has certain weaknesses: The metric favors conservatism, since if a certain area of energy is ambiguous or contains both target and interference, it is best excluded from the output; as the signal-to-noise ratio is only calculated over regions that are passed, there is no direct penalty for deleting valid target energy such as the perceptually-important, low-energy, low signal-to-noise ratio channels in the higher end of the spectrum. Brown argues that his scheme can be used for comparison of different source separation systems, but in fact it only applies to systems that generate their results via time-frequency masks as his does; other approaches which synthesize an output signal based purely on abstracted parameters can't use the separate-filtering trick.

Brown had the courage to make his sound examples widely available. Listening to them provides insight into the assessment problem; despite marked improvements according to his metric, and although his system has evidently been very successful in removing interference energy, the quality of resynthesized targets leaves much to be desired. They are patchy (often sounding muffled owing to the missing upper spectrum) and subject to considerable artifactual frequency-dependent amplitude modulation where the system fails to identify the correct allocation of time-frequency cells. Informal comparison of the original mixtures and separated targets gives the impression that the target speech is more easily perceived before processing; the removal of interference energy by the separation process does not necessarily compensate for distraction caused by the artifacts. This general problem in sound-separation schemes has been more formally recorded by the work of Kollmeier *et al* on sound-segregating directional hearing-aid algorithms [KollPH93] [KollK94]. Despite having systems that are demonstrably capable of rejecting the majority of off-axis and reflected energy, they are obliged to limit the effects of their algorithms rather dramatically in order to minimize deleterious artifacts when tuning them for use by real hearing-impaired listeners. In one case, intelligibility tests showed an optimal balance gave a gain equivalent to only 2-3 dB improvement in plain signal-to-noise ratio.

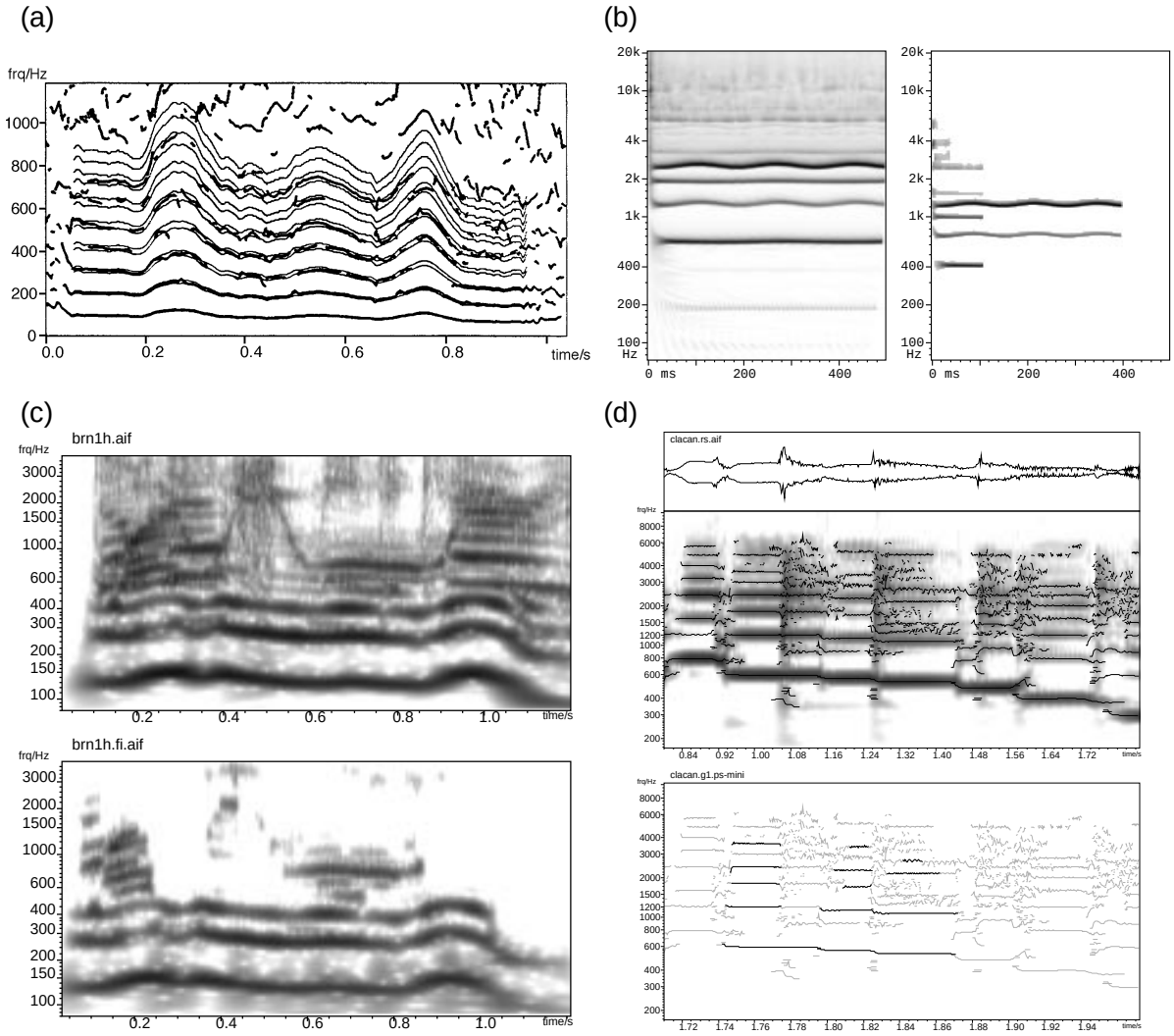

Figure 2.2: Illustrations of the four 'data-driven' systems discussed. Panel (a), drawn from [Cooke91], shows the synchrony strands extracted from a fragment of voiced speech, along with the 'harmonic sieve' used to group the resolved harmonics. Panel (b) is drawn from [Mell91], showing the spectrogram of the McAdams-Reynolds oboe-soprano sound, along with one of the sources extracted by his system. Note that up until 100 ms the system fuses all the harmonics, but then it segregates the even harmonics on the basis of their common modulation. Panel (c) shows the spectrograms of voice mixtures used in [Brown92] before and after processing to extract one voice; the effect of his time-frequency masking is clearly visible as the extensive 'white' regions where interference has been removed. Panel (d) is reproduced from [Ellis94], showing the sinusoidal tracks used to model a mixture of a harmonic sound (a clarinet) and a transient (a dropped tin can). The lower panel highlights the tracks corresponding to a clarinet phrase, grouped on the basis of harmonicity.

2.5 A critique of data-driven systems

The systems we have discussed are all highly original and significant pieces of work. However, when taken together, they reveal fundamental weaknesses of their common paradigm which must be acknowledged.

The most basic criticism is that, essentially, these systems don't work – at least not particularly well. Clearly, the motivation is to build a sound processing system that will isolate a target sound from interference in a manner similar to that achieved by a listener; this cannot be said to have been demonstrated. Cooke and Brown come closest with their impressive target-to-interference ratio figures, but the figures are far removed from perceptual reality. Listening to the resyntheses from Brown's system, it is obvious that the ratio of target energy to background noise has been greatly improved, but missing target energy and the artifacts resulting from discontinuities in the filter mask rather spoil the results.

In any case, the initial goal of a human-like sound understanding system has in each case been severely diminished in the interests of tractability. Thus the systems of Cooke and Brown are mainly aimed at continuously-voiced speech, and Mellinger only addresses the resolved harmonics of pitched musical sounds. Ellis's system appears to be limited to the specialized problem of removing rapid noise-like transients from steady, pitched tones.

Perhaps the most basic explanation for the limitations of these systems is that the authors were simply too ambitious and underestimated the difficulty of the tasks the initially addressed. However, it is also possible to identify some specific weaknesses in these systems that may be attributed to the data-driven approach common to them all. This category includes:

Inadequate cues: Each author acknowledges that their system functions less well than might be hoped, and makes the point that this situation could be improved if only they used more information from the input sound. According to the Bregman account of auditory organization as the integration of a variety of cues, each system would benefit from the addition of, say, spatial information from binaural cues. In most cases the authors suggest that such additional information could be incorporated rather easily, or at least outline the modifications necessary for such enhancements.

In addition to the acknowledged omissions, problems with cues include uncertainty about the detection even of the ones that are used: Despite a seemingly unambiguous list from psychoacoustics (common onset and fate, common location, harmonicity, continuity, ...), the precise formulation of these cues in signal-processing terms is not known. For instance, both Mellinger and Brown implement onset detector maps as rectified differentiators within each frequency channel, and both recognize the importance of having a family of maps based on different time-constants to be able to detect onsets at a variety of timescales. But there is no consensus over how to combine information from these different scales to generate the 'true' onset cue; Mellinger uses information from any map that indicates an onset, whereas Brown found that using only the very fastest map was adequate. Other authors have looked at combining information across scales to register an onset only if it occurs in multiple maps [Smith93]. Onset detectors also generate artifacts in the presence of frequency-modulated tones, since a tone gliding into a particular filter's band will cause an increase in energy in that band which may resemble the appearance of a wholly new

sound object. (Note, however, various techniques for suppressing such 'ghosts' such as suppressing the neighbors of excited channels [ShahD94]). Essentially, while the importance of an onset detector is universally accepted, the right way to build one is still an open question. A slightly more involved cue, such as the common amplitude modulation exploited in comodulation-masking release phenomena [HallG90], lacks even a speculative computational model at this point.

Inextensible algorithms: Despite their authors' protestations to the contrary, these systems are not particularly flexible in providing for the inclusion of new or modified cues that might become available. The systems of Cooke, Brown and Mellinger have an abstracted stage of similarity judgment, where basic elements are added to the same group if their 'score' is sufficient, and the score could be redefined to include contributions from additional cues. Ellis has parallel primary-grouping processes to which new modules based on new cues could be added. However, since each of these strategies employ the new cues only in comparisons between the same primary representational elements of the basic system, the scope of such extensions is limited. It is equally likely that cues need to be involved at an earlier stage in the formation of the elements themselves, and this is not addressed in these systems.

Rigid evidence integration: The fixed processing sequence embodied in the representation and grouping modules of these systems not only limits the role of additional cues, but also seems a poor model of the robustness of the human auditory system. In Ellis's system, for instance, a harmonic sound-event is formed by first searching through all the elements to form groups that appear to have harmonic relationships, then searching for groups of elements that have close onset times, then comparing both resulting sets of groups for corroborating pairs. This fixed, procedural sequence of operations is a definitive feature of these data-driven systems. By contrast, the human auditory system seems to be far more adaptable in the face of missing or obscured evidence: When a sound is presented with strong common-period cues but missing or ambiguous onset information, the common-period information will be used as the basis for the perceived object. If, on the other hand, the harmonic relationships are rather poor, but the onset is well-defined and synchronous, the common-onset cue will ensure the perception of a single event, at least for the first few hundreds of milliseconds of duration. If energy that otherwise seems independently structured shows a common spatial cue, that too can be the basis for a fused object [SteigB82]. Rather than following a fixed sequence of steps to form a percept, the auditory system appears to take stock of the reliability of whatever evidence is available, then choose the best-suited object-formation strategy. It is hard to imagine how the pre-programmed grouping algorithms of the systems considered so far could exhibit this kind of robust adaptation to the prevailing conditions.

Inability to handle obscured data: One particularly arresting example of this robust adaptation of the perceptual system is illustrated by the various 'auditory restoration' phenomena. In the classic example of [Warren70], listeners were played a speech recording in which an entire syllable had been deleted and replaced by a loud cough-like noise burst. Not only did the listeners succeed in inferring the linguistic information in the deleted syllable (on the basis of semantic context), but they were unable to identify precisely *where* in the speech the added cough had occurred; their preconscious perceptual processing had restored the obscured syllable so confidently that

their conscious experience was of 'hearing' the inferred, restored speech just as clearly as any other part.

Clearly such restoration is a useful attribute of perceptual processing. If some data is not directly observable as a result of chance signal collision or sensor limitations, but the information can be reliably guessed from the context, higher-level processing is best served by a peripheral system that can incorporate the 'guessed' information just as if it had been directly perceived. Data-driven systems, which operate by making a representation of the raw data present in the input, then sorting it to generate output groups, cannot accomplish this kind of restoration; the obscured data simply doesn't appear in the representation, and there is nothing in the grouping process to modify or add to the basic representation. This is perhaps the starkest illustration of the weakness of data-driven processing. The usual comparison is with top-down or context-sensitive processing, which explicitly incorporates the idea that more abstract levels of analysis can affect the information represented at lower (more concrete) stages. Top-down systems are discussed in the next section.

2.6 Advances over the data-driven approach

There have been a number of alternative approaches to aspects of the auditory scene analysis problem, some of them more or less directly in response to the problems of the data-driven systems discussed above. It is perhaps too early to present these developments as instances of a broader pattern, but we will now consider some of them individually, paying attention to the ways in which they overcome weaknesses of the data-driven systems.

2.6.1 Weintraub's state-dependent model

Weintraub's [Wein85] system was in fact the first explicit computational model of auditory organization, yet it occurred so long before the more recent systems discussed above that it is harder to relate to them. This work predates Bregman's [Breg90] book, which I have presented as the impetus for the subsequent models, although Weintraub was very definitely influenced by the psychoacoustic results then being obtained by Bregman and others which were eventually summarized in that book.

Weintraub actually describes two systems, a 'first system', and a more successful 'current system'. The first system actually falls very neatly into the data-driven framework: The target mixture (two voices, both of which are to be recovered) is analyzed by a front-end of a cochlea filterbank followed by an autocorrelation-like 'coincidence function' that identifies significant periodicities in each channel. The coincidence functions are pooled across channels to detect one or two voice-pitches, which are then used to group the 'neural events' (the low-level representation of acoustic energy in time-frequency) into tonal 'group objects'. Another kind of object collects energy at onsets, which would otherwise provide insufficient history to establish periodicity. Noise-based group objects were envisaged but never actually implemented.

Thus the first three stages, front-end, representation and grouping are readily identified. No details are given of the system output, since it is described mainly to indicate what didn't work. Weintraub states that the system was incapable in many cases of detecting the presence of two voices, and made errors at too high a rate.

His second system sought to establish a more methodical 'decision framework' for integrating the various information determining the character of the two voices. This system is structured as a strictly left-to-right progression through a sequence of processing stages (reflecting the batch-processing restriction imposed by the computers of the time), but is marked out as different from other auditory models by having a central 'hypothesis' concerning its belief about the current state of the two voices assumed to comprise the input. Each voice can be in one of three stable states (silent, periodic, nonperiodic) or four transition states between them. It is prohibited for both voices to be labeled 'in transition' simultaneously, so the joint hypothesis for the two voices has a total 33 distinct states (3 stable states x 7 possible states for the other voice x 2 for the symmetric cases minus 3x3 duplicates of the states where both voices are stable). This state is chosen on the basis of the coincidence analysis of the signal and an exhaustive combined pitch tracker for both pitches based on dynamic-programming. The current state determines the subsequent processing invoked to recover the spectra of each voice; for instance, voice spectra recovery is trivial when either voice is considered to be in its 'silent' state, since this implies its spectrum must be zero, and hence any observed energy belongs to the other voice.

This centralized global state of the signal hypothesis is very different from the more local, less restricted sound objects grouped by the other auditory organization models. Nothing in Brown's system, for instance, limits the number of objects that can be extracted, although in practice it is only a single foreground voice that is generated as output [Brown92]. Weintraub's global hypothesis amounts to a high-level abstraction that governs subsequent lower-level signal processing. However, his fixed algorithm results in the slightly awkward sequence of pitch tracking that leads to the state hypothesis that leads, in turn, to spectral recovery – there is no opportunity, for instance, for the state hypothesis to cause a revision of the pitch tracking because pitch tracking is 'complete' before the state hypotheses have been calculated. This one-way processing is typically data-driven, but the alternating levels of abstraction are not.

Weintraub exploits the restrictions implicit in enumerating every possible state for the pair of voices by using a training corpus of voice mixtures, hand-labeled according to his voice states, to derive his model's parameters. Thus transitions between possible state hypotheses are weighted by the measured *a priori* frequency of occurrence of such transitions in his 38 three-second examples. Similarly, the iterative mixed-spectrum separation stage is driven by a set of histograms that store the likelihood of voice energy ratio (in a single frequency channel) conditioned on three samples from the coincidence function: at the two detected pitches and at their difference. (Each coincidence function value is quantized to five levels, giving 125 different histograms derived from the training data for this one parameter). Perhaps the most difficult aspect of constructing this system was finding the right normalization methods and attributes upon which to condition the histograms. A more open-ended space of central hypotheses would require an exponentially increasing training set, even if such globally-conditioned processing were possible.

Weintraub is quite clear that, in his view, any complete model of auditory organization will necessarily involve more than just bottom-up, data-driven processing. He cites examples such as the improvement of comprehension with familiarity and the benefit of visual cues as evidence that auditory source separation must be intimately bound up with the process of recognition, with plenty of top-down control flowing from the latter to the

former. Dissatisfaction with the performance of his first, purely bottom-up system led him to a strikingly innovative state-based approach that managed to include a measure of abstraction-guided processing without sacrificing the computational efficiency of fixed, sequential processing. However, his results from testing the second system as a front-end to a speech recognizer are inconclusive, and he concludes that far more emphasis needs to be put on top-down, analysis-derived controls in future systems.

2.6.2 Blackboard systems

The majority of the problems ascribed to data-driven systems – difficulty in incorporating new cues, fixed dependence on particular features and inability to perform restoration – stem from the limitations of the data-driven, procedural approach to solving the problem. This approach is the dominant paradigm in signal processing, where algorithms are described by data flowcharts showing the progress of information along chains of modules. But this may be inadequate for modeling the more complex, adaptive, context-dependent processing accomplished by the auditory system.

As a result, much of the interesting recent work deals with alternative approaches to controlling the execution of the computation involved in auditory organization systems, that is, different processing architectures. Perhaps the most sophisticated of these is the blackboard architecture, which has been developing for more than twenty years as a foundation for sensor interpretation systems and other abductive reasoning problems. A particular collection of blackboard-based processing techniques has been developed at the University of Massachusetts, Amherst and Boston University [CarvL91] and has been applied to the very relevant problem of identifying different sound sources in domestic environments [CarvL92a], among other domains. The IPUS system, which extends this sound-understanding example, is described in more detail below.

It is fitting that the blackboard paradigm should find new value in the computational auditory scene analysis, since it was initially conceived as an approach to solving a problem of auditory perception – namely, speech recognition. The Hearsay-II speech recognizer [LessE77] is usually considered to be the original blackboard system (For a review of the development of blackboard architectures, see [Nii86]). Its defining characteristics were:

- The **blackboard**, which is a global database of hypotheses, both supportive and competitive, comprising the results of all the inferences and predictions performed. The hypotheses on the blackboard constitute the entire 'state' of the analysis system, in a public, explicit form available to all action modules.
- **Hierarchic organization of hypotheses**. Blackboards are normally divided into well-demarcated layers, where hypotheses are linked between layers by "supports" (lower to higher) or "explains" (higher to lower) links. Sets of hypotheses linked together by such relations form coherent partial explanations or 'islands of certainty' that may be consistent or competitive with other such groups on the blackboard.
- '**Knowledge sources**', or action modules, which create and modify hypotheses. Knowledge sources are independent domain-specific modules that can create or modify hypotheses. (Creating hypotheses at a higher level is 'explanation'; at a lower level, it is 'prediction'). Each knowledge source is attuned to a particular set of circumstances which it recognizes

and then develops. Since the state of analysis is completely and transparently encoded in the blackboard hypotheses, it is a simple matter to add new action modules to extend an implementation.

- An **opportunistic control system** which chooses the action modules to execute at each moment. The control system typically has a certain amount of knowledge about the likely consequences of each module, and can also estimate the usefulness of different action modules in the current blackboard context, typically through a module-supplied rating function. Although many approaches have been used [CarvL92b], perhaps the simplest is Hearsay II's 'agenda' system, where the agenda is a list of potential modules to run, sorted by their rating score. The control system simply executes the action at the top of the agenda, then refills, re-rates and re-sorts it.

Blackboard systems have had the most success in applications where the potential search space of hypotheses is enormous, but the fine-grained adaptive control system finds solutions with only a small amount of exploration. It is in the control systems that most of the theoretical refinements to the model have been made since Hearsay II. The RESUN system of [CarvL91] is a particularly refined example, where the knowledge sources can be not only action modules that complete in a single step but also plans that consist of a sequence of subgoals to be achieved over several steps, incorporating the power of backward-chaining reasoning systems into the blackboard paradigm.

The benefits of blackboard systems for models of acoustical scene analysis lie in their provision for top-down, hypothesis-directed processing, in contrast to the bottom-up, data-driven systems discussed in the previous sections. The structure of the blackboard makes little distinction between explanatory and predictive operations; predictions made on the basis of a partial abstract analysis can arbitrarily bias and reconfigure the lower stages to which they apply. Thus, restoration and inference phenomena, where information derived implicitly from the context is processed as if it had been extracted directly from the raw data, fit very naturally into this processing structure. A blackboard system is intrinsically extensible, since a significant part of the design consists of defining hypothesis representations and summaries of partial-blackboard states terms of general goals that are comprehensible to any relevant knowledge source whether or not it currently exists. Extending a blackboard system is as simple as defining a new action module in terms of the condition it responds to and the results it generates. When suitably registered with the control system, it will be called automatically as appropriate. Such extensibility also affords tremendous flexibility for the researcher to investigate how the analysis proceeds with different subset of action modules. This flexibility of blackboard systems has been identified as particularly valuable in the stage of system development characterized as "exploratory programming" [Shiel83].

2.6.3 The IPUS blackboard architecture

Recovering causal explanations from sensor data has long been studied in the artificial intelligence community as a useful computer application without any particular concern for modeling the human performance of this function. The IPUS system [NawabL92] [LessNK95] is a particularly interesting example in this literature, both because it is one of the most sophisticated approaches of this kind, and also because an example implementation described in many of the IPUS papers is an environmental sound recognition

system, the “Sound Understanding Testbed”, that has essentially the same goals as the auditory scene analysis function that is our concern.

The paramount idea motivating the IPUS approach is that signal interpretation should be a dual search: In common with other such systems, there is the search in the abstract representational space for the explanation that accounts for the data most effectively. However, their premise is that this search must be conducted in conjunction with a second search for the signal-processing front-end configuration that best exposes the evidence for this interpretation. Moreover, this second search can be based on models of the signal processing units' behavior, derived from the domain's underlying signal processing theory.

The assumption is that, given the 'correct' abstract explanation of the signal data, there exists an optimal configuration of the signal- processing front-end which, when applied to the signal, generates 'correlates' that confirm unambiguously the correctness of the explanation. However, since neither the explanation nor the configuration are known initially, the system must iteratively search for both, reprocessing the data with each new suggested configuration until an explanation has been confirmed with satisfactory confidence.

In IPUS, this interaction between abstract interpretation and signal-processing algorithm configuration is accomplished by four key modules emphasized by the authors: Discrepancy detection, discrepancy diagnosis, signal reprocessing and differential diagnosis. The first three are the stages necessary to extend a hypothesis abstraction system to adapt its numerical front-end to the context: Starting with some initial front-end configuration, discrepancy detection tests a collection of rules whose violation indicates that the configuration needs to be improved. These discrepancies may arise from a priori constraints within and among the front-end processors (e.g. that an overall energy increase in the time envelope must correspond to new energy peaks in time-frequency analysis), or there may be discrepancies between the system's model-based expectations (predicted correlates) and the actual observations.

The next stage, discrepancy diagnosis, seeks to explain the explicitly- labeled discrepancies in terms of specific causes. When a prediction has not been confirmed, the cause may be mistakes in the system's abstracted signal model (the current interpretation) – the usual motive for exploration of explanation-space in signal interpretation systems. However, in IPUS there is the alternative diagnosis that the front-end configuration is at fault and should be improved. The classic example involves the underlying Fourier theory of time-frequency uncertainty: If the model-based predictions are that two simultaneous sinusoidal components will occur close together in frequency, but the observed correlate is only a single peak in the spectrum, the diagnosis may be that the frequency analysis signal-processor was configured with too coarse a frequency resolution, a problem that can be corrected by using a longer time window for that unit. Discrepancy diagnosis implements an 'inverse map' between discrepancy symptoms and reprocessing strategies which are then implemented in the third module, signal reprocessing. In this stage, the original signal data is processed by the modified front-end configuration to generate a new set of correlates which may then be tested for discrepancies.

The fourth IPUS module described by the authors is differential diagnosis, a well-known concept in signal interpretation systems that becomes

significantly more useful when, as in IPUS, the search domain is extended to front-end configuration. Even if configurations have been found that satisfy the discrepancy checks, there may still be a choice of several possible credible explanations. In the terminology of the RESUN [CarvL91] blackboard system, which is the foundation of the IPUS systems and which was specifically designed to provide for differential diagnosis, the overall solution has *uncertainty* arising from possible alternative explanations. The RESUN system detects this situation explicitly, thereby permitting the execution of routines specifically intended to discriminate between the particular alternatives. This is in contrast to traditional evidence-aggregation blackboard systems that will gather information that generally supports a particular hypothesis without focusing effort on the particular information that will distinguish between the alternatives under current consideration. In the context of IPUS's search for a front-end configuration, differential diagnosis becomes particularly powerful as the discrimination techniques can include reprocessing of particular portions of signal data with processing units configured solely to resolve the ambiguity of the competing explanations.

The IPUS architecture represents a new level of sophistication in knowledge-based signal interpretation systems, offering unprecedented flexibility in the knowledge-based control of the signal-processing front-end, as well as indicating how the theoretical foundations of a particular problem domain (such as the Fourier uncertainty) can be used as the knowledge to guide that control. There are perhaps some weaknesses to the underlying assumption that an ideal front-end configuration will exist if only it can be found. The assumption behind diagnosis and reprocessing is that each object to be detected in the signal-generating universe will produce a set of unambiguous distinct features which will be resolvable into corresponding correlates by the correct configuration of signal processing units. However, referring again to the Fourier domain example, there will be signal combinations that cannot be resolved by manipulating the time-frequency resolution tradeoff, such as two objects that happen to produce simultaneous sinusoid components at the same frequency. An interpretation system can detect this circumstance and verify that observed correlates are consistent, but working within the limitations of the front-end is contrary to the IPUS philosophy of removing uncertainty by correcting the front-end; accommodating front-end limitations is the resort of fixed-configuration systems. Another possible weakness is the assumption that an object is characterized by some canonical set of features which are mapped to a specific set of correlates by the particular signal-processing configuration. The signature of a particular object can only be defined with reference to a particular analysis procedure, and though the underlying theory can define how to translate results from one configuration into those expected of another, the idea of distinct a-priori features falls down, for instance in the case of a single amplitude-modulated carrier that may, in other situations, appear as a cluster of four or five closely-spaced harmonics: Which representation is to be used as the 'canonical' feature? Systems with fixed front-ends know in advance the form of their signal correlates, and thus have a much easier time choosing internal representations.

The artificial-intelligence signal-interpretation literature embodied in the IPUS architecture has had a deep influence on the model of perceptual information processing presented in this thesis. It is interesting to contrast the two approaches: my primary interest is to model the way in which the auditory system extracts information about the real world, whereas the

motivation for IPUS-like systems is simply to analyze complex sensor data to achieve specific goals by whatever means are most successful. Since artificial systems cannot yet approach the capabilities of the human auditory and visual systems, we might expect some convergence of state-of-the-art automatic signal interpretation systems to the architectures embodied in our perceptual physiology. However, the reverse assumption may not hold: It would be foolish to assume that the best current interpretation algorithms are those being used by the perceptual system; more likely, the perceptual system uses far superior algorithms that we have yet to develop in artificial systems. However, the model-based, prediction-reconciliation architecture of IPUS and related systems suggests intriguing explanations of several known perceptual phenomena, as discussed in chapter 3, which have convinced me that it is the best direction for current perceptual information processing models.

2.6.4 Other innovations in control architectures

Other researchers have pursued different approaches to the problems of flexibility and adaptability in the control of sound-understanding systems. One of the most intriguing is the agent-based approach pioneered by Okuno and his co-workers at NTT Basic Research Labs in Tokyo [NakOK94] [NakOK95]. The philosophy behind this approach is to re-formulate the sound-understanding problem in terms of a collection of co-operating, but largely independent agents, each specializing in a particular task. In [NakOK94], the system operates by creating 'tracker' agents which attempt to follow particular sets of harmonics in an input mixture as they vary in frequency. Supervisory agencies handle the creation of new trackers when there is excess signal energy to be accounted for, and deleting trackers whose signal has disappeared, as well as censoring degenerate conditions. The flexibility of the approach is vindicated in [NakOK95] where the same system is extended with agents specializing in grouping harmonics according to their binaural-spatial characteristics, and a further agent to account for steady background noise. Their demonstrations of separating and streaming discontinuous mixtures of voices and tones are very impressive.

The independent, locally-specific agents created to track aspects of the input signal correspond to an amalgam of knowledge sources and hypotheses in a blackboard system, but the difference in metaphors results in a different flavor of solution. A hypothesis of a harmonic signal would probably be matched in a blackboard system with a rule to anticipate the continuation of that signal in the future; in contrast, the creation of an agent specifically assigned to that harmonic signal is tightly bound to an active search process for continuation through time – these agents serve a kind of object-oriented role of linking together all the knowledge associated with the detection, characterization and tracking of a particular kind of signal. As with blackboards, the details of representation and information flow between agents is not dogmatically specified, leaving unresolved some questions of how best to combine the actions of agents using different principles to contribute to the grouping of the same signals (for instance, the co-operation between spatial and harmonicity agents mentioned above). In this respect, it is possible that the probabilistic formalization of the blackboard system, where alternative signal hypotheses may be rated on their conformance to a collection of criteria, has a conceptual advantage.

A further innovation in the NTT work is recognized in its title, the "residue-driven" architecture, meaning that the residue signal – the result of taking

the input and subtracting the currently-tracked signals – is analyzed for evidence of newly-appearing sounds. Although quite different in detail, the prediction-driven approach described in the next chapter was much influenced by this concept.

Another important issue in all such systems, that of combining evidence from disparate aspects of a signal analysis. There is no simple solution to the problem of combining the confidence scores of a collection of pieces of evidence into the score that should be conferred on an abstraction they support, and the possibility of 'don't know' (highly uncertain) data complicates matters further. This problem was considered carefully in the music understanding/transcription system of [Kash95], which was able to integrate knowledge and expectations over a huge range, from the typical behavior of harmonics in signals to the common patterns of chord progressions in western music. This was accomplished using the techniques of Bayesian belief networks, which permit the integration of all known information in a hierarchic tree of supporting (and hence conditional) hypotheses. Application of such principled evidence integration becomes increasingly critical as models of auditory information processing expand in scope and complexity.

2.6.5 Other 'bottom-up' systems

The cues or features employed by grouping principles are the basic determinants of the ultimate success of the approach; without the appropriate information captured at the lowest level, no amount of clever inference and abstraction is going to be able to reconstruct the sound. However, the difficulty in developing cue-detection schemes is that their full value may only become apparent when they are used in conjunction with the correctly-attuned element formation and grouping strategies. This is perhaps why the majority of effort at the moment is going into developing processing architectures; refining the low-level cues may need to wait for a good general architecture to emerge.

Nonetheless, there have been some interesting recent developments in this area. Much attention has been paid to the way that the binaural cues of interaural time and level difference are used by listeners to judge the spatial position of a sound and to separate it from interference. One goal of this general area of research is prostheses for hearing-impaired listeners that use the same or similar cues to amplify only voices from a particular direction – since this is a common problem area for sufferers of hearing loss, even when compensated by amplification. The pure gains of beam-forming microphone arrays are applicable [SoedBB93] [StadR93], but human performance appears to exceed this, leading to proposals of nonlinear algorithms exploiting steadiness of spatial location and the harmonic structure of speech [KollPH93] [KollK94] [Woods95].

These techniques have been specifically applied to the problem of auditory scene analysis in Blauert's lab in Bochum, Germany. Using a sophisticated inhibited-cross-correlation model of azimuth perception that models the combination of timing and level cues [Gaik93], they use Wiener filtering to boost signals from a particular 'perceived' azimuth in an optimal fashion [Bodden93]. Such a system is ripe for combination with other grouping cues for improved signal discrimination [GrabB95]. Onset is known to be particularly influential in the perception of location [Zurek87], so the inclusion of common-onset-type cues should be particularly beneficial.

Although it was effectively a component of the systems of both Brown and Mellinger discussed above, the correlogram deserves separate mention as a distinct approach to the problem of detecting common periodicity and common modulation [DudaLS90] [SlanL92] [SlanNL94]. The correlogram is a time-varying two-dimensional function of frequency channel and modulation period, generated as the short-time autocorrelation of the intensity in each frequency channel of a cochlea model. Several aspects of this structure are very neurally plausible and are even supported by physiological evidence. Animated displays based on the correlogram analysis convert important acoustic properties – such as micro-modulation or jitter – into highly detectable *visual* phenomena of correlated motion, suggesting that correlogram processing might form a link between these two dominant sensory modalities. The correlogram forms a rich basis for process-oriented representations such as the *weft*, which can be used for separating mixed voices ([EllisR95], also described in chapter 4).

Correlogram analysis is often characterized as 'time-domain', since it detects periodicities in the signal using autocorrelation (delay-and-multiply) rather than a frequency transform. Another time-domain approach, also motivated by the perceptual importance of small period fluctuations (or 'jitter') was described in [Ellis93a]. The 'glottal-pulse synchrony' model attempted to detect repeated patterns of time alignment of energy maxima across frequency, as would appear if the different frequency channels were being repeatedly excited by the same source such as the vocal chords of a particular individual. Unusually, this model looks only at the timing skew *within* each energy burst and does not take account of the time *between* energy bursts (the voicing period) that is the usual basis for detection. In the model, jitter promotes segregation of independent voices because a random displacement of the energy pulse will affect all peaks from a given source equally, but disrupt any chance alignment they may have shown to independent signals in the mixture.

2.6.6 Alternate approaches to auditory information processing

So far in this section we have discussed developments in control architectures and in cue detection, both of which are recognized components of the general computational auditory scene analysis model introduced in the preceding sections. Other work does not fit so comfortably into this framework but still clearly belongs in this survey. We will now consider some of these.

A typical psychoacoustic experiment will repeat a short sound-event, or alternate a pair of events, giving the listener multiple opportunities to extract whatever information they can from the signal. The idea that information from each repetition is being overlaid onto a single, composite record lies behind Patterson's model of the Stabilized Auditory Image [PattH90]. In this model, cochlea and nerve models generate a firing pattern as a function of time and frequency, which is repeatedly copied into a buffer, taking its alignment from a particular threshold or trigger that should ensure synchrony between successive copies. Repeated instances of the sound can then be accumulated according to this synchronization. Patterson has had notable success in explaining some rather esoteric perceptual phenomena with this model [Patt94]. However, the model as it stands does not really address the problem of separating components in real-world, nonrecurring mixtures, our primary focus in this survey.

2.6.7 Neural network models

As with many complex pattern detection problems, auditory scene analysis has attracted the application of artificial neural networks, stemming from the neurophysiological speculations of [vdMalS86]. Unlike static image analysis, sound processing suffers the complication of having ever-advancing time as a variable rather than a fixed dimension, but several systems have been proposed to accommodate this situation [Wang95]. Perhaps the most interesting model of this kind is the one proposed by Brown and Cooke [BrownC95], who have both previously worked on symbolic sound analysis systems as discussed above. In a seemingly radical break from rule-based sound organization, they have recently proposed a network of chaotic neural oscillators, excited from the energy emerging from a cochlea filterbank model. Since a signal that exhibits common onset will deliver a synchronized 'kick' to the oscillators across all its regions of spectral dominance, these oscillators will tend to fall into correlated oscillation. Overlap between the passbands of adjacent cochlea-model channels and 'inertia' in each oscillator give the system a certain amount of smoothing in time and frequency. This, in combination with a model of grouping by correlation of channel oscillator, can successfully reproduce properties such as common onset and proximity grouping, as well as streaming by frequency proximity – an impressive range of phenomena for a fairly simple structure that is not overtly designed for these tasks.

Reproducing properties of high-level auditory processing by the 'emergent' behavior of biological-style networks presents some interesting problems in the philosophy of research and modeling. It is very likely that the particular solutions to detecting grouping cues employed in the auditory system will be efficient, multipurpose structures similar to these deceptively simple network models. On the other hand, the reductionist, explicit approach of characterizing and implementing individually each inferred operating principle may be a better first step towards understanding the 'how' and 'why' of auditory operation. Recall the distinctions made by Marr between implementation, algorithm and 'computational theory'; while we are still struggling to understand the computational theory, we might want to stick to inefficient, literal implementations. Once we are sure we understand just what there is to be calculated, we are well served by clever, integrated models. The (somewhat anti-Marrian) objection to this approach is that in all likelihood there are many aspects of the information processing in the auditory system, particularly the detection and combination of cues, which are far more easily understood in terms of the specific implementations of their detection than by some abstract theory – imagine trying to account for color salience and ambiguity without knowing about the three classes of color-sensitive detectors in the retina. Ultimately, convergence and cross-over between implementation-oriented modeling and computational-theoretic modeling will furnish a definitive solution to the problem of audition.

2.7 Conclusions and challenges for the future

It is only recently that computational modeling of human hearing has gained enough momentum to deserve consideration as a distinct field, yet as we have seen it already contains a broad range of approaches and outlooks. In this chapter I have tried to impose some structure on the work that is most directly relevant to the current project, systems that sought explicitly to model the process of auditory scene analysis. As they stand, these models are vehicles for testing implicit theories of how information processing in the brain might be structured, rather than specific solutions to engineering problems (although, for the particular problem of speech recognition in the presence of interference, they come quite close). However, the main lesson thus far is that the auditory system is an extremely subtle and robust mechanism that requires a great deal of delicate care in its emulation. In particular, we saw how the single-path, procedural 'data-driven' models of early researchers experienced a host of problems when exposed to ambiguous or obscured data.

These difficulties led us to consider a range of properties that we might look for in more successful auditory organization models, including:

- The detection and integration of all the cues critical to human hearing, as we become better aware of just what they are.
- Processing architectures that permit adaptive use of available and contextual information, and which intrinsically provide for the addition of new principles and cues.
- System outputs of a form useful to specific goals of a system, be they resynthesis of source signals or identification of particular attributes (such as verbal content).

Several of the recent developments in the field have incorporated these refinements, particularly the systems based on the innovative control schemes of agents and the blackboard architecture. The next chapter will propose an approach to modeling auditory function that arises from observations of this kind, and an implementation based on the approach will then be described and assessed.

Chapter 3 The prediction-driven approach

In the last chapter we saw the typical approaches that have been taken to computational auditory scene analysis and discussed some of their more problematic limitations. I will now present my proposed solution to these problems, the prediction-driven architecture. This chapter reviews the main functional motivations behind the architecture and provides a theoretical overview of its structure. The following chapters will examine the implementation of the system that has been produced.

3.1 Psychophysical motivation

In the background chapter we saw how the process of sound organization in the auditory system has typically been modeled by a data-driven structure, where specific features in the sound signal are used as the basis for representational elements, which are then grouped into larger entities forming the analogs of perceptual events or sources. However, we also saw a number of situations where this left-to-right processing paradigm was insufficient to reproduce the kind of sound analysis actually performed by human listeners. It is these shortcomings that have motivated much of the subsequent work in this area, including the system described in this thesis. In order to explain the particular qualities of the prediction-driven architecture I am about to describe, let us look once again at some specific aspects of auditory perception that cannot be handled by a data-driven model.

Inference of masked or obscured information: Listeners are able to cope with situations where certain features are difficult or impossible to recover from the input signal by making estimates of the missing information based on other aspects of the signal. The most extreme example of this kind of phenomenon is the auditory induction of [Warren70], where an entire syllable that has been replaced by noise is semantically restored without conscious awareness. However, I would contend that similar 'filling-in' occurs at less dramatic scales in almost every listening task.

My attachment to this idea comes from the sometimes bitter experiences of several researchers in trying to recover adequate acoustic features from sound mixtures in the construction of data-driven models [Ellis93b]. Many sound mixtures, despite being 'transparent' to the human ear, present a signal that seems hopelessly confounded and disrupted when analyzed by conventional signal processing techniques. There are two possible interpretations of this: Firstly, that signal processing techniques are inappropriate or too crude to pick out the kinds of details that the ear is using (no doubt true to some extent). The second possibility is that the ear cannot, in fact, extract information directly from these messy mixtures either, and it is abandoning direct analysis of these convoluted regions to rely instead on guesses based on information inferred from more evident features of the signal. Circumstantial evidence in favor of this second interpretation is that we already know that the auditory system can perform very considerable feats of auditory induction, all without conscious trace of their intervention. Thus we should *expect* to have the experience of 'directly' hearing the pieces

of a sound mixture, even if they have in fact been supplied by inferential guessing. The only way to distinguish what has been 'truly' perceived from that which had been inferred is to construct trick stimuli where the inference and the actuality are different (such as the [Warren70] phrases, where the implied syllable had been deleted prior to the addition of the masking noise, and thus was not strictly present, despite being 'heard'). All this tells us is that inference can occur in auditory processing and we should not expect to have any introspective indication of when it is being employed.

Other psychoacoustic phenomena illustrating this kind of effect include various 'continuity illusions' [Breg90], where a tone alternating with a burst of noise will, under suitable conditions, be perceived as continuing through the noise, in spite of the fact that the sound stimulus was not constructed that way. (Note that we do not say that the tone isn't there, since the illusion only occurs when the noise energy is sufficient to make an objective assessment of whether the noise contains the perceived tone mathematically ambiguous).

The kinds of mixtures that prove difficult to analyze by data-driven models are often less ambiguous than this. One particular problem scenario for the sinusoidal models such as [Ellis94] is the interference between harmonics of different sounds. If the harmonics fall into the narrowband region of peripheral frequency analysis, where they would normally appear as individual sinusoids, but are close enough together in frequency to exert constructive and destructive interference on one another, the resulting 'beats' will confound any sinusoid-extraction algorithm. What, then, are we to conclude about the mechanism at work in the human listener who is able to interpret the mixture as the crossing of two sinusoids without any uncertainty or confusion? My inclination is to attribute this perception to an inferential mechanism that recognizes when the two tones are getting sufficiently close together to confound the output of peripheral frequency analysis, and simply checks that some minimal conditions of the combined amplitude until the components move apart again. Such a mechanism might possibly be confirmed with psychoacoustic tests on the perception of an artificial signal that appeared to be the crossing of two tone but that replaced the region of their interference with some artificial substitute, preserving the larger-scale spectro-temporal characteristics of their interference but distinguishable by a more detailed analysis that predicted the exact phase and amplitude of each sinusoid. A stimulus of this kind would be delicate to construct, and I am not aware of any investigations of this kind.

Even if it turns out that human audition is more sensitive to sinusoid interference than I am predicting, the general principle that many important features are masked at the periphery, and that our commonplace perception involves considerable 'restored' inference, can hardly be doubted in view of our ability to understand, with effort, telephone calls made from noisy bars, or to recognized the music being played in the apartment upstairs, despite the fact that the greater part of the sound information is absent in both cases. Thus we must seek to build models of auditory organization that incorporate abilities of inference and restoration.

Context sensitive interpretation: A common flaw in auditory processing models is that a given local acoustic feature will always be processed the same way. This is in contrast to a wide range of examples in human audition, where a particular acoustic feature may have radically different interpretations depending on its *context*, i.e. the other features with which it occurs. We term this variation in treatment 'context sensitivity', although

there are many different faces of this basic idea. For instance, a single resolved sinusoidal component may serve as a harmonic of various fundamental periodicities, but even a data-driven system will only assign it the role of a particular numbered harmonic in the context of a pattern of other harmonics that occur simultaneously. However, other examples of context sensitivity are beyond the scope of a data-driven structure.

One set of instances arises from the perceptual concept of 'priming', i.e. the ability of certain stimuli to bias the interpretation of subsequent events. One of the best-known demonstrations in auditory scene analysis is the Bregman-Pinker experiment [BregP78] in which a single harmonic may either *fuse* with a co-occurring sub-harmonic (simultaneous grouping), or *stream* with a temporally-alternating tone at approximately the same frequency (sequential grouping). This competition may be manipulated in favor of the sequential grouping by desynchronizing the onset of the simultaneous tone. This example is usually used to illustrate the distinction between simultaneous and sequential grouping and how they can be made to compete. However, it may also be viewed as an example of *priming*, since the presence of the isolated tone 'primes' the perceptual system to interpret the matching portion of the two-tone complex as a separate object. None of the full-blown organization systems described so far carry enough context to differentiate the treatment of the target tone depending on whether a priming tone has preceded it; their architecture cannot easily accommodate such plasticity of processing, which is a key weakness. (However, perhaps owing to the ubiquity of this example, a number of more limited models have displayed the 'appropriate' sequential grouping - see [BeauvM91] [Gjerd92] [Wang95] [BrownC95], all of which use proximity in a two-dimensional time-frequency plane to reproduce the effect. Note also [GodsB95] which specifically sets out to provide context-sensitivity in these kinds of problems using blackboard architectures – as will I).

Another challenging example of context sensitivity in auditory organization occurs in [DarwC92]. Their experiments consist of manipulating a mistuned harmonic in a complex, and measuring the perceived pitch. The frequency of the mistuned harmonic is kept constant, but its influence on the matched pitch of the whole complex varies according to features known to influence grouping. For instance, when the harmonics all occur simultaneously, a 4th harmonic mistuned by 3% will shift the pitch of the entire complex by 0.5%. However, if the mistuned harmonic starts 300 ms before the rest of the complex, its influence becomes negligible; we infer that the large onset asynchrony has excluded the mistuned harmonic from the fused complex, and the perceived pitch of the complex is based upon the frequencies of the exactly-tuned harmonics only.

This is a hard enough phenomenon to reproduce in a data-driven system, since we are now saying that rather than just inspecting the pattern of harmonics present at a particular instant, we must also keep track of when they started, and apply a grouping penalty for asynchronous onset. However, Brown's system does precisely this: when the algorithm considers grouping two harmonics, it checks for activity in the onset maps at their temporal limits to confirm that the start of the harmonic as represented is a genuine energy change in the signal. If simultaneous onsets are detected, it adds a fixed bonus to the 'grouping score' between the two components. However, the next experiment performed by Darwin & Ciocca adds a complication that defeats this approach: By adding a set of aharmonic components to the 300 ms of 'exposed' mistuned harmonic, that stop as soon as the target complex begins, the influence of the mistuned harmonic on the pitch of the complex is

restored. Their motivation for this experiment was simply to dispel the possibility that the reduction in influence was due to low-level adaptation to the mistuned harmonic rather than high-level grouping. However, it presents a neat example of context-sensitivity: the influence of the second half of a sinusoidal component on the pitch of a harmonic complex depends on the presence or absence of a cluster of components that can group plausibly with the first half, thereby 'freeing' the second half to group with something else.

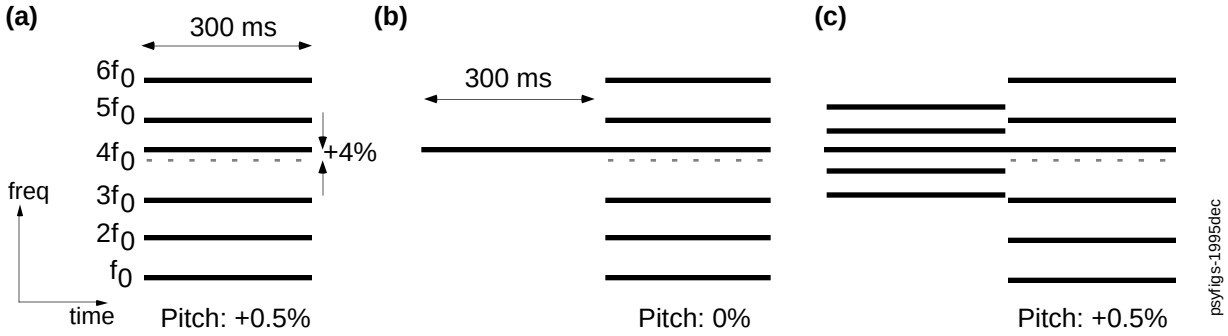

Figure 3.1: The Darwin/Ciocca experiments. In panel (a), a 3% mistuning of the fourth harmonic leads to a shift of 0.5% in the perceived pitch of the entire complex. In panel (b), starting the mistuned harmonic 300 ms before the rest of the cluster eliminates its influence on the cluster's pitch. In panel (c), adding a group of inharmonic partials simultaneous only with the exposed portion of the mistuned harmonic release the remainder and restore its effect on the pitch.

We conclude that the organization of auditory patterns is, in general, a global operation that cannot be accomplished exclusively by bottom-up, local operations but requires some kind of influence from the higher analysis. This parallels many results in visual perception and would appear to be a general feature of advanced interpretation systems. Particular aspects of the inadequacy of bottom-up processing in vision are discussed in [ChurRS94] and related to hearing in [Slaney95].

Ambiguity and revision: The examples of context-sensitivity mentioned above, where the interpretation of a particular acoustic element can take on several different forms depending upon the surrounding elements, is a kind of ambiguity. But there is another kind of ambiguity posing yet more difficult problems for successful models, where a single element has an interpretation that *changes* over its duration as the balance of evidence shifts in favor of an interpretation different from that initially adopted. This is evident in the McAdams-Reynolds oboe example [McAd84] described by Mellinger [Mell91]. Indeed one of the unique features of Mellinger's system, in contrast to other data-driven systems, was that it could 'change its mind' about the organization of harmonic partials into auditory objects as time progressed.

This is a particularly thorny problem for computer systems. The experience of listening to the McAdams-Reynolds example is that the sound initially resembles a bright, oboe tone. However, as the note is held, frequency modulation of steadily increasing depth is applied only to the even-numbered harmonics (e.g. $2f_0$, $4f_0$, $6f_0$ etc.) and this ultimately causes the perception of two sources – a hollow, clarinet-like tone comprising the odd, unmodulated harmonics, and, an octave above the clarinet, a tone like a soprano voice corresponding to the modulated harmonics. At some point in time there is a perceptual shift when the brain reconsiders its initial assumption that the

harmonics all belonged to a single source and instead understands them as the combined result of two independent sources. It is as if the auditory system concedes that its initial organization was in error, that the sound should have been heard as two sources from the very beginning; it is not clear what this implies in terms of the mental representation of the portion of the sound that has already been 'heard'.

Another example of retroactive modification of the interpretation of a sound can be found in the alternating noise stimuli mentioned in [Breg95] (originally reported in [Warren84]). Here, short segments of noise with the same power spectral density but different bandwidths are alternated (e.g. 400 ms of 0-1 kHz followed by 200 ms of 0-2 kHz, with no silence in-between). Regardless of how the stimulus was actually constructed, the inclination of the perceptual system is to interpret this sound as a continuous noise signal in the 0-1 kHz band, to which 200 ms bursts of noise between 1-2 kHz are periodically added. Bregman cites this as a classic example of the 'old-plus-new' organizing principle, where under a wide range of circumstances a change in the structure of a sound will be interpreted as an *addition* of a novel component rather than a modification of existing, stable sources. The concept of revision becomes relevant if the cycle begins with an instance of the *wider* noise-band. Without any narrow band of noise to constitute an 'old' comparison, the initial wide band of noise is interpreted as a single source. However, this interpretation becomes unsatisfactory as soon as the second band of noise appears, since this change is not a particularly likely modification of the wide noise band; the preferred interpretation becomes to regard the initial noise burst as the combination of a high noise band (which terminates) and a lower band (that continues as the narrow noise band). This interpretation does not arise until *after* the initial noise burst has finished.

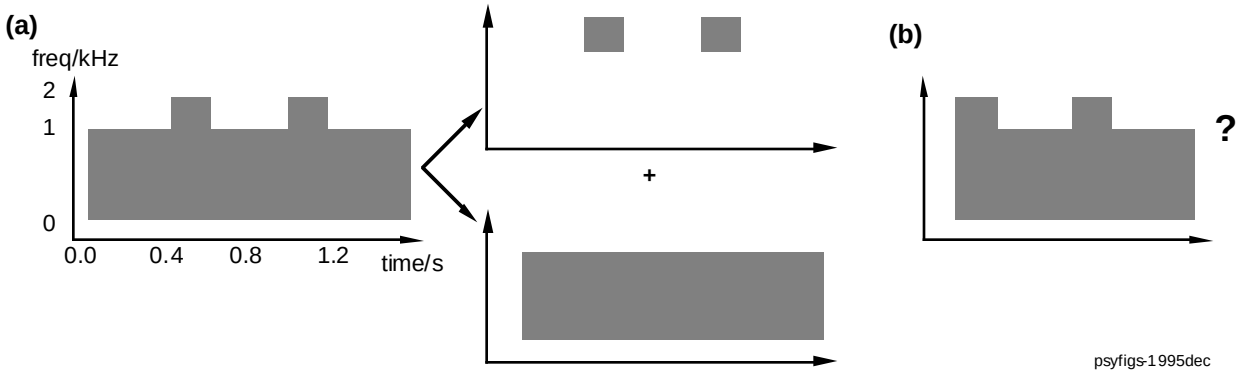

Figure 3.2: The alternating noise-bands of panel (a) are irresistibly interpreted as a continuous lower band of energy with periodic bursts of a higher-frequency band. In panel (b), starting with the wider band of noise may cause the initial burst to be interpreted as a single noise rather than a combination of bands, necessitating a subsequent 'revision' of the interpretation.

Thus we have seen several varieties of commonplace auditory organization – inference, context-sensitivity and ambiguity – which present insurmountable challenges to a bottom-up, data-driven analysis. This tension has spurred the development of a new approach, the prediction-driven architecture, which is now described.

3.2 Central principles of the prediction-driven approach

Ultimately, auditory organization modeling must aim to present a definitive model for these puzzling perceptual phenomena. In order to reach that eventual goal, we need an architecture that can at least plausibly accommodate these kinds of processing – something that is not provided by the data-driven approach. This is the purpose of the prediction-driven architecture. There are, of course, many influences on the structure of the architecture, some based on clear theoretical requirements, others derived from poorly-defined intuitions resulting from previous experiences; it would be futile to attempt to enumerate them all. Having set the scene by highlighting the problematic phenomena, I will being the presentation of the architecture by introducing some of the key concepts related to its operation:

World model: The first concept to introduce is the idea of a world model. To use Marr's term, the 'computational problem' of audition is the construction of a simplified representation of the external world that accounts for the received acoustic stimuli while simultaneously satisfying other constraints, such as sensory information from other modalities and underlying assumptions about the environment. For our purposes, a world model is characterized as a collection of independent objects whose aggregate behavior explains – and, critically, predicts – the observed signal. (The choice of this definition was made because it captures what I believe is the most important aspect of our perceptual interpretation of the world, namely that we understand the world as being constructed from the combination of independent entities).

Interestingly, Churchland *et al* [ChurRS94] come out strongly against this concept as a basis for modeling visual perception, preferring goal-specific processing that derives only the variables of use to the task at hand, rather than wastefully extracting details that will not be needed. Their arguments are powerful, but open to the rebuttal that as the number of special-purpose tasks served by a perceptual system becomes very large, the economics shift towards a single, general-purpose analysis; the approach best suited to the human visual system may be qualitatively different from the goals and constraints applicable to that of a fly. It is also possible that audition is significantly different from vision in that it has no limited 'field of view', nor any obvious analog of the fovea; it is practically impossible to 'ignore' certain aspects of auditory stimuli, whereas resource limitations in the visual periphery make selective attention a central and early principle of operation.

Avoiding the derivation of features other than those specifically required for a task is one alternative to a world-model centered approach. Another distinction may be drawn between systems based around world-models and systems which, while having no particular ideological objection to general-purpose representation of the world, still lack a well-delimited portion of their structure fitting this description. This 'delocalization' of the world model might occur for one of two reasons: On one hand, many systems have implicit narrow assumptions about the structure of the world, e.g. as a single pseudo-periodic target signal amidst non-target. In this situation, the 'world model' consists of the parameters of the identified target and an undifferentiated 'everything else'; the limited expressive capacity of such a representation leaves it poorly described as a world-model. Another possibility is a system that calculates all the information to make a world model, but chooses not to organize it around representations of external objects – perhaps focusing instead on one cue at a time without drawing connections between the different cues as manifestations of the same cause. While this may appear to

be only superficially different from a world-model based system, the different perspective is likely to have a profound influence on the types of computation and reasoning performed.

Consistency: How do we connect a world model with a stream of sensory data? The data-driven model presents one answer, which is to construct successive levels of abstraction, each stage founded solely on the identifiable features in the data. The alternative here is to generate the abstractions by some other means, then to require only *consistency* between the model and the stimulus. This concept arises as a result of the range of uncertainty introduced by the abstraction process. For a given perceived signal, particularly if we consider sensory noise, there may be any number of possible explanations. How do we choose among them? The data-driven answer is to be maximally conservative, only to postulate entities whose presence cannot be doubted. In the prediction-driven framework, the model itself is obtained by a wider range of mechanisms (i.e. predictions from the existing components), and the 'connection' is limited to ensuring that the model falls somewhere in the space of uncertainty. Depending on how model and stimulus uncertainty is represented, there may be a wide range of possible matches, with a continuum of resulting confidence or quality metrics, rather than a single, brittle yes/no comparison.

To provide a precise definition for consistency we will have to define our basic beliefs about the use of cues by the auditory system, something inevitably open to debate given our current state of knowledge. A provisional list of the most basic stimulus properties would include:

- Energy in the time-frequency plane: Activity in a certain area of the auditory nerve is directly indicative of acoustic energy in a certain frequency region, and the basic task of the auditory system is to account for any such deviations from absolute silence. Absence (or inadequacy) of perceived energy is also a very strong constraint: any explanation that requires energy in time-frequency regions where no energy is in fact detected is clearly inconsistent with the observations. (This explains why it is easier to understand speech interrupted by noise bursts that if the interruptions are silent – the noise-burst corrupted signal is *consistent* with a continuous speech stream, but the silences are not [Breg95]). Thus close adherence to the time-frequency energy envelope extracted by the peripheral filtering appears to be an important condition of consistency.
- Periodicity: A second very important cue is the presence of consistent periodicity along time and across frequency. This is detected by listeners at surprisingly low signal-to-noise ratios [Colb77] and has a strong influence on the percept, suggesting a particular mechanism for and perceptual emphasis upon this feature. We may conclude that 'consistency' must include an account of any such periodicities detected in the signal.

Prediction as a starting point: The third introductory concept for the prediction-driven architecture is the role of the predictions themselves. A world model defines a representational space, and the criterion of consistency allows us to decide when our model is acceptable, but how do we come up with a candidate model, especially in view of the underdetermined nature of our criterion? The search space is huge and the solutions are not unique, but rather than continually starting from scratch (the data-driven approach), we can use *predictions* as the first estimate for the representation. Ignoring for

the moment the bootstrapping problem of starting from a situation of no knowledge, if we already have some beliefs about the sound-producing entities in our environment as embodied in our current world-model, we can exploit that knowledge to predict what will occur next. (In practice, this encompasses simple first-order continuity prediction through to complex inference on the basis of similarity to remembered sound-patterns). If our prediction is then consistent with new observations, we do not need to invoke any further computation to account for the input signal; our analysis is complete. Even if the prediction is not immediately acceptable, we have a starting-point that we can seek to modify to achieve consistency.

An interesting corollary of this approach is that the analysis may overlook valid alternative explanations if it already has a satisfactory explanation pre-empting the search for others. This is a serious issue in artificial reasoning systems based on this kind of approach – the 'termination problem' of [CarvL92a] – but has a nice parallel with certain auditory illusions, where a sound is unwittingly misinterpreted as the result of a misleading suggestion (for instance, the fusion of musical instruments, sometimes perceived as a single source owing to their carefully aligned onsets and harmonics, yet in reality several different sources).

The key point about this process is that the prediction, originating in the internal abstraction of the world model, can include many subtleties and details that would not necessarily be derived directly from the observed features, but are none the less consistent with them. Thus our analysis system has the potential to 'detect' the kind of detail that we typically experience in real audition but which generally eludes current, data-driven systems. In this way, a prediction-driven system opens the possibility, subject to sufficiently sophisticated prediction mechanisms, for inference of obscured data and context-dependent treatment of features.

Model-fitting as constraint-application: The process of delineating a representational space – the world model – then choosing a point within the space to correspond to observed data has been discussed so far as model-fitting. A subtly different perspective is to look at it as the application of constraints: the essence of any model is that it assumes a certain underlying structure in the source, leading to codependence in the observations. Finding a model, at whatever degree of sophistication, that can be fit to part of the observed data indicates that the data obeys the implicit constraints of codependence.

It is these constraints that permit the model to make predictions. For a region where observations are hidden or ambiguous, the internal dependencies of a model that has been fit to the observable parts of the data permit reasonable estimation of the missing data. For unobscured data, the constraints operate to select among all possible model instances the one that is appropriate; where data is obscured, their role changes to that of the source of best-estimates of missing information. The prediction-driven architecture manages to unify both of these roles in a single predict-and-verify loop, but the different interpretations are illuminating.

The constraints that are actually applied, i.e. the kinds of model that the system employs, are intended to mirror the organization and sensitivities of the auditory system. These constraints might be viewed as arbitrary empirical limitations arising from our overall aim to reproduce human sound organization. But of course the ear has evolved to be particularly well adapted to kinds of sounds and sound mixtures that are actually encountered

in the real world. To the extent that we can successfully devise a representation that duplicates the saliences and limitations of the auditory system, we may assume that the constraints being applied are the useful and powerful assumptions that listening must employ in order to perform its astonishing feats of scene analysis.

3.3 The prediction-driven architecture

We can now examine how these concepts are assembled into the architecture. For the purposes of the explanation, the prediction-driven system will be divided into four main pieces: the generic sound elements at the core of the world-model, the higher-level abstractions, the front-end processing, and the prediction-reconciliation engine that connects them all. The arrangement of these pieces is illustrated in figure 3.3, and each part is discussed below.

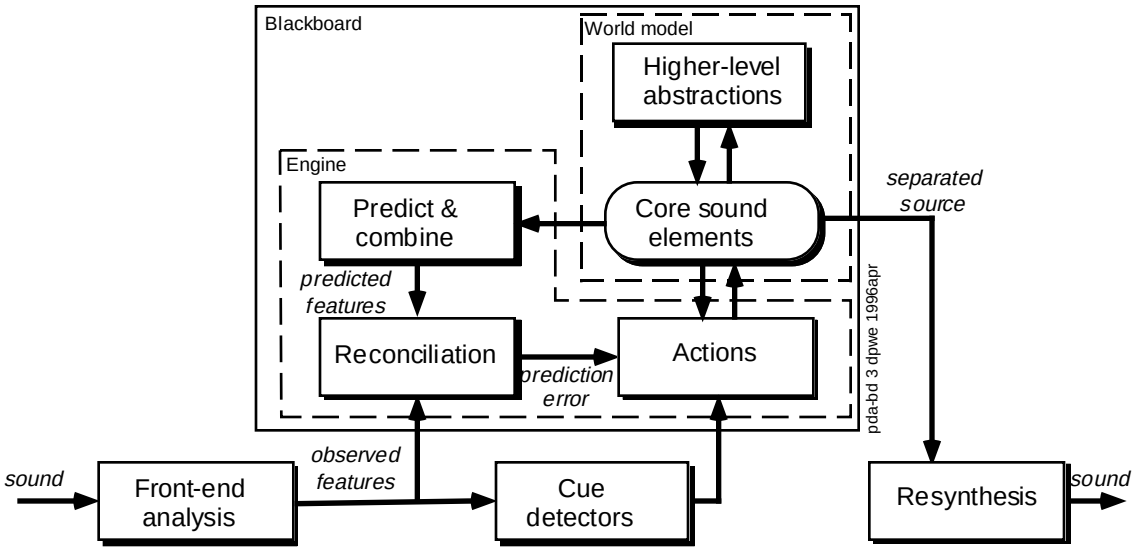

Figure 3.3: Basic layout of the prediction-driven architecture, showing how the engine manages explanation of and reconciliation to the front-end features by the core world model, as well as updating and deriving predictions from the higher-level abstractions.

The core world-model: We have already introduced the concept of a world-model in abstract terms, and the core of this system is such a collection of representations of independent sound sources perceived as existing in the environment. So far I have been rather vague about the concept of a representation, but at this lowest level the meaning is quite specific: the core consists of a collection of generic sound elements, instances of representational entities that express the perceptually-relevant aspects of the kinds of sounds that occur in the world by capturing all the overtly-dependent energy from a single source into a single element. At present there are three classes of elements, whose structure has been chosen to give a satisfactory expression of the different kinds of subjective experience of sound, to be a reasonable basis for modeling arbitrary signals, and to embody some of the real-world constraints apparently used in auditory organization [Ellis95b]. The elements are:

- **Noise clouds:** The simplest explanation for the detected signal is as unstructured noise energy. Noise cloud elements have relatively few

constraints on their structure, but they also encode rather little information about the assumed source, weakening their ability to predict. The rationale for having this kind of element lies in the everyday perception of certain sounds as 'noise-like', e.g. wind blowing through trees, water running through pipes, or the fricative sounds in speech. When we hear a sound as noisy, we largely ignore its fine structure; it is mainly the average properties (such as the overall spectrum) that are significant. Thus a noise element is parameterized by its intensity over a range of time-frequency, but this surface is rather sparsely sampled, reflecting an assumption of smoothness in the underlying expected time-frequency intensity. Noise clouds can, however, start and stop abruptly, as with all elements.

- **Tonal elements (wefts):** Another very broad and important class of sounds is those with a perceptible periodicity, generally experienced as pitch. Such sounds are represented by an element we have christened the 'weft' [EllisR95]. Wefts have the same kind of smooth time-frequency energy envelope that characterizes a noise cloud, but in addition they have a period-track, recording the detected periodicity of the energy whose extent is described by the envelope. The periodic sounds that are the domain of wefts were also the exclusive focus of the previous sound organization models described in chapter 2; however, those models tended first to extract individual sinusoids, then group them together on the basis of their lying in a harmonic pattern. The weft avoids this perceptually suspect separation of individual harmonics by representing all energy associated with a particular periodicity in a single element even at the lowest level. (Note, however, that the analysis procedure for a weft is derived from the 'auditory objects' of [Brown92]. The difference lies mainly in the order of operations: Brown carves up his time-frequency plane, calculates a pitch-track for each segment, then groups segments whose pitches match. In weft analysis, the segmentation is based on an initial pitch-track extraction from the entire signal, allowing, incidentally, the segments to overlap.) The details of weft analysis are presented in greater detail in appendix A.
- **Transients:** The third representational element is designed to account for very rapid bursts of energy that are too short to have any pitch, and also too short to be sensibly modeled as some kind of steady-state noise process. The brevity of such a transient event means that its spectral detail is low, and it is characterized by a broad spectrum and some indication of its duration and decay. The class of perceptual events that this element is seeking to model are the clicks and cracks of the everyday sound environment.

The elements are intended to be the most generic components needed to express any sound in terms of its perceptually important qualities. Thus the approximate nature of the noise model, where most of the detail of the signal is deliberately ignored in favor of the average energy over some large time window, would be entirely inappropriate for applications where information may be present in that detail. However, because we are dealing with human perception of sounds, we are justified in treating certain large classes of sound as essentially equivalent. (The precise formulation of the noise element does not exactly achieve this – however, the intention is to have an element that precisely describes the experience without recording any unnecessary detail).

Higher-level abstraction hierarchy: The generic sound elements of the core world model form the lowest level of the system's representation of the sound-world. They are the basis of an abstraction hierarchy wherein elements are explained as the outcome of more sophisticated causal structures (e.g. repetition of remembered patterns, instances of detailed classes such as speech). In theory, this abstraction could be arbitrarily complex and refined (speech is a good example), just as long as it can be realized into a collection of these bottom-level generic elements. This hierarchy is an important part of the overall architecture, although it has not been very much developed in the current implementation. It is responsible for two of the key advantages of this approach: the ability to infer corrupted or obscured information, and the possibility of extending a particular instance of a sound understanding system with new knowledge in the form of additional abstractions.

Front-end feature analysis: This consists of the signal processing primitives applied to the input sound waveform to derive the attributes that the world model must match. I have already mentioned the two kinds of feature that must be explained – the overall energy envelope and the signal periodicity. The energy envelope is calculated as the rectified, smoothed output of a fixed, linear approximation to a cochlear filterbank, and provides a general indication of the signal energy at each time-frequency location after removing the fine detail. The periodicity detection takes the short-time autocorrelation of the signal in each peripheral frequency band and integrates this across channels to detect strong periodicities in the signal. (Greater detail on this processing is provided in the next chapter). Both of these cues are 'indispensable', that is, a world-model cannot be deemed consistent unless it accounts for these cues with a high level of accuracy.

It is possible to have additional forms of cues that are consulted only when they are needed. Onset is an interesting example: Many auditory models acknowledge the demonstrated importance of onset in the formation of perceptual wholes by having specific onset-detector circuits. (This tendency is encouraged by direct physiological evidence of such detectors [HewM91]). Such detectors are somewhat redundant in a system that is already following the energy envelope of a signal, since an onset will be reflected in that domain too; it may be that the special perceptual importance of onsets is as much an aspect of deeper processing as of the peripheral circuitry. However, a system that can explain a rapid increase in energy either as a brief click (the transient sound element) or as the beginning of a sustained energy burst (noisy or tonal) would benefit from some kind of peripheral detector to distinguish these two conditions, perhaps by suitable filtering of the channel envelopes. This additional onset-discrimination cue would not be included in predictions and would not be consulted in determining the consistency of a prediction. However, under certain conditions it could be used by the explanation process to guide the creation or modification of model elements.

Prediction-reconciliation engine: Tying these pieces together is the engine comprising all the control logic to maintain the world model, supervise the construction of the abstraction hierarchy and the formation of predictions. It must then check the consistency of the predictions with the front-end cues, reconciling any prediction error by modifying the representation as necessary.

From the perspective of a real-time system, where sound is being explained on-the-fly, the force driving the progress of the analysis is new information becoming available from the peripheral cue detectors. The first step is to

compare this with the current predictions. Assuming the stimulus is behaving consistently with the current world model, these predictions should be adequate. There will be slight differences between the predicted and actual observations, corresponding to the residual unpredictability of the external sources, but these are reconciled by making minor modifications to the current elements. (Apportioning of prediction error between overlapping objects is rather delicate, as discussed later). This additional confirmation of the current representation is propagated up the explanation hierarchy, which then permits modifications of the current predictions to descend the hierarchy and be reflected in the base representation. In this case, the loop is ready to start again.

However, the observed signal may not be consistent with the predictions, in which case the engine must make more radical modifications to the world model. Where there is significantly less energy in the observed signal than in the prediction, the engine looks for a world-model element to terminate, such that its removal would restore consistency. On the other hand, excessive observed energy is handled by the creation of one or more putative explanatory elements (perhaps on the basis of additional cues as discussed above). These changes to the bottom level of the explanation hierarchy will trigger modifications at higher levels, according to specific rules attached to each abstraction, and possibly resulting in further changes at the lowest level. The net result, however, is a new prediction with which to continue the analysis.

The actual implementation of the analysis engine is as a blackboard system and is discussed in chapter 4.

3.4 Discussion

The basic exposition of the architecture raises many issues, including relating it to the perceptual phenomena introduced at the beginning of this chapter. Some of the relevant qualities of the approach are discussed below.

Inference: Under the right conditions, a prediction-driven system can perform inference on incomplete data where some cues have been obscured. As discussed above, model fitting is a form of constraint application: Each level at which a model can be constructed reflects additional constraints that the input data has been found to satisfy. Assuming that there is sufficient uncorrupted evidence to indicate an appropriate abstraction with little ambiguity, the resulting analysis will be confirmed when the predictions of that model are found to be consistent with the input – even in the area of obscured data, since although a corrupting signal may have prevented cue extraction in the first instance, it will also preclude the *refutation* of predictions in that region. Thus the combination of a specific explanatory hierarchy, an analysis engine that searches the solution space, and a success condition that demands only consistency, leads directly to a system capable of inference-style phenomena, where missing cues appear in the internal representation just as if they had been directly perceived. The tricky part is, of course, the model hierarchy – expressing the possible configurations of cues with sufficient flexibility to match the full range of sounds that occur, yet with sufficient precision to be able to make useful predictions of the cues that are absent (or at least to extract the same level of information despite the missing cues). In order to perform the phonemic restoration examples – where a linguistically-significant syllable is inferred – the internal analysis needs to extend all the way up to the semantic constraints on the content of

the sentence. While we are still a few years away from this degree of automatic language understanding, examples of inference at lower levels – such as the perceived continuity of tones through noise bursts – is quite feasible.

Context-dependence: A similar argument can be made to explain how this approach to analysis results in treatment of particular cues that depends heavily on their surrounding features. The constraints represented by the current context will determine the space of explanation uncertainty within which a particular cue must be accommodated, and hence the role served by that cue may be radically different in different situations. This seems almost too obvious to state given this kind of approach to analysis, but it should be contrasted with a purely bottom-up analysis: In systems of that kind, there is no attempt to build a high-level abstraction on the basis of part of the raw data, then adapt the processing of the remainder conditional upon that abstraction; the data-driven systems, as I have characterized them, work only by synthesizing data at each level of abstraction, applying all possible required processing at each ascending level of representation without the benefit of previous, more abstract analyses.

Ambiguity: A particular set of information may not uniquely specify an abstract interpretation. To use a very high-level example, we may hear a sound and be unsure as to whether it is the doorbell or the telephone. In such situations, the system could create a range of plausible hypotheses and make distinct predictions for each. These hypotheses will be maintained as long as their predictions can be reconciled to the input signal, the hope being that some later information (such as the regular ringing pattern of the telephone compared to the increasingly frustrated repetitions of the doorbell) will eliminate the incorrect conclusions. This approach of handling ambiguity by pursuing multiple hypotheses is a common feature of blackboard systems [Nii86] [CarvL92a], as well as having proved its value elsewhere [Rutt91] [Rosen92]. However, difficulties arise when the number of hypotheses that are produced becomes unmanageable: Using flexible models to produce fewer, more highly parameterized fits to a given signal helps to restrict this problem, but there is a fundamental combinatorics problem between independent ambiguities for which a solution better than enumeration has yet to be found for this work.

Revision: The second aspect of ambiguity mentioned at the start of this chapter is the situation where an analysis at one stage of processing is rejected sometime later on – for instance, the broad noise band that suddenly loses a big chunk of its spectrum, suggesting that in fact it had originally been two spectrally-adjacent noise signals and that one of them has stopped. If there had been any prior hint that this was the case, the system might have been maintaining two alternative hypotheses – the hypothesis that the noise was a single-band, which would ultimately have been rejected, and the alternative hypothesis that the signal is a combination of bands which would have sustained. However, assuming that the possibility of a combination had not been postulated, the collapse of the single-band hypothesis poses interesting processing choices. To make the historical analysis consistent with its revised beliefs, the system would need to go back in time and re-represent the broad band of noise in light of the new information. Nothing about the architecture prevents this kind of backtracking, however, the focus has been on processing new information as it arrives. It is not clear how much access or attention should be given to past events once they have been processed, particularly since we assume that progressively less information is held for events further in the past. On the other hand, the context

comprising these past events will influence the current and future processing, thus the system needs at least to revise the more abstract levels of its model to incorporate its changed interpretation of the passed noise burst, for instance to 'prime' the analysis to interpret added noise in the future as a repetition of the retrospectively-constructed second band.

Competition between hypotheses: In the previous paragraph I invoked the idea of the system maintaining multiple explanations as an interim measure to handle the situation when the cues received are insufficient to distinguish alternative interpretations. Given the possibility of a large number of competing hypotheses, we need some approach to choosing which of them to develop, and a way to dispose of past possible interpretations that have become very unlikely. To achieve these aims, each hypothesis has an associated quality rating which reflects the confidence with which that abstraction accounts for the relevant aspects of the input data, and can be used to choose the most promising from among a collection of hypotheses for further processing effort. This rating is derived from the goodness-of-fit between the particular constraints of that model and the data it is seeking to explain. At higher levels of abstraction, it is also influenced by the ratings of its supporting hypotheses. Since the progress of the analysis is guided by these ratings, they demand some care in their construction. They can be related to probability, since a successful explanation for a given set of data with a low *a priori* probability provides the most additional information about the situation, and is thus the most promising analysis to pursue. A more constrained hypothesis, such as a tonal sound, should be rated above a more general alternative like a tone burst, all other things being equal. The rating of hypotheses, according to a probabilistic or minimum-description-length principle, is described in chapter 4.

Prediction-led search: As we observed above, the predictions provided by the world-model provide a starting point for the search for explanations of new data. If the prediction is adequate, no search is needed at all. But even when some kind of exploration is required, having a specific starting point that is presumably quite close to a good solution is a considerable advantage over a system that starts from scratch to account for each new piece of data. From the perspective of searching the solution space, the use of predictions makes it feasible to have a more intricate, ambiguous space and still arrive at good solutions with limited computational effort; data-driven systems tend to require simple solution spaces with unique solutions that permit the direct, bottom-up analysis of input data. The explanatory redundancy of the generic sound element representation, where a certain patch of energy might be a single noise element, or the overlap of several, or perhaps a tonal element instead, would be intractable without some way rapidly to locate and choose between the possible alternative explanations.

Resynthesis: There are many practical reasons to wish for a path to resynthesis in a model of sound organization. Perhaps the most important at this point in the development of the field relates to assessing the outcome of analysis – listening to a resynthesis based on a system's analysis often gives the most insight into system behavior, and opens the way to subjective tests of the system (as discussed in chapter 5). The definitions of the generic sound elements ensure that they include sufficient information to permit a resynthesis of an actual sound matching the recorded characteristics; because the elements are essentially concerned with the physical properties of the sounds they represent, such resynthesis is relatively straightforward. For noise elements, it is simply a matter of amplitude modulation of noise signals filtered into the different frequency bands, and tonal elements can be

similarly reconstructed by starting with a pulse train. More details of the resynthesis are provided in the next chapter.

Encoding uncertainty in predictions: Throughout the system, signals are represented not simply as levels, but as probability distributions – in most cases by associating a variance to each 'expected' signal level. Where the value is not intrinsically uncertain, such as the observed signal coming from the front-end, this variance can be zero, although a value indicating sensor noise might be more appropriate. This stochastic formulation is necessary for instance in the predictions made by noise elements, since the noise is modeled as a steady average power which predicts the distribution, but not the individual values, of the samples observed at the front end. The stochastic representation of predictions has other uses, such as permitting the expression of uncertainty in model parameters. It also provides a convenient domain for the combination of predicted levels, even those contributed by disparate elements. One particular example is in the combination of tonal elements: if two narrow-band periodic signals overlap in frequency, the likely result is constructive and destructive interference between their waveforms – known as 'beating' in the case of close, steady sinusoids. The precise amplitude contour of this beating is sensitive to the exact phase alignment between the two signals, yet the representation, following perception, is not particularly concerned with phase and does not predict it. However, beating can be accommodated by recognizing the situations where it is likely to occur, and widening the variance of the amplitude prediction in that area to encompass the full range of intensities that can result from the interference.

Error allocation by parameter uncertainty: In the description of the engine's processing loop, I mentioned that when a prediction is found to be adequate, there will still be a small prediction error, which is then divided among the contributing elements to trim their parameters to match the known input as closely as possible. This distribution can be troublesome, for instance in the case where more than one object has contributed to the predicted level in a particular time-frequency cell: what error should be assigned to each one? One approach might be to pro-rate the error according to each object's contribution to the prediction; a better approach is to weight this allocation according to the *parameter uncertainty* of each element. An element that has been successfully and accurately predicting the observations should not be asked to accommodate as great a proportion of the prediction error as a recently-created element whose precise identity has yet to be quantified, even if the more stable element contributed a larger value to the prediction. To permit this kind of behavior, each element provides an 'error weight' indicating its parameter uncertainty and hence its preparedness to accept prediction error, and the error is apportioned on this basis.

3.5 Conclusions

The prediction driven architecture is presented as an alternative to conventional data-driven systems. It is better able to emulate some common hearing phenomena that bottom-up analysis is hard pressed to reproduce. We have reviewed some of these phenomena, then considered the underlying concepts and the main modules of a prediction-driven system. The ways in which the architecture helps to address the motivating problems have been developed in more detail. There remain a couple of broader issues, mentioned

in the introduction, which should now be considered: extendibility of the architecture, and its plausibility as an auditory model.

Extendibility: This was a principal motivation behind the development of the architecture, different from the consideration of problematic phenomena. At this early stage in the field, it is unrealistic to aim to build a model of auditory organization that is anything like complete, or even that is largely correct in the details it does address. Recognizing that the process of continued research in computational models of auditory scene analysis will involve many modifications and additions to the systems of today, it would be preferable to work with an architecture that is amenable to changes without entailing major changes to every component. This is a luxury; it is difficult enough to build a model of any kind, and imposing additional constraints unrelated to the actual function of the immediate goal may not be feasible. On the other hand, the robust adaptability to different conditions shown by the auditory system suggest that it is highly modular and plastic, and thus this should be an attribute of a model for functional, as well as logistic, reasons.

Serving the twin goals of suitability for future development and embodiment of modular, adaptable theory of auditory organization, the prediction-driven architecture offers significantly improved flexibility for extension compared to previous auditory organization models. To a certain extent, this extendibility arises from the fact that the architecture in itself does not comprise a complete, operational model: it requires the addition of abstractions and their associated rules to provide the structure in the explanation hierarchy. However, the framework that such abstractions fit into provides for the addition of such innovations with the very minimum of programming.

The second aspect contributing to the extendibility of the system is the way in which the engine incrementally develops hypotheses according to their current rating scores, a common feature of such blackboard systems. Rather than having a system where the sequence of analysis steps is explicitly specified in advance by the programmer, the system as described selects, from among the possible hypotheses and rules, the one that is currently rated as most probable. Thus if a new class of abstraction is added, all it must do is ensure that the rules it provides give themselves a sufficiently high rating when the circumstances are right for their instantiation. The engine will then automatically invoke the new methods, even though it was previously unaware of this kind of abstraction.

Finally, new rules may also employ information from new kinds of front-end cue detector, which they will know to invoke. There are, of course, limits to the adaptability of the basic engine: there is no way to add new 'indispensable' cues without rewriting the prediction-reconciliation rule. However, there is no barrier to adding 'advisory' cues to guide the creation and development of new kinds of element or other representation.

In the face of such broad claims to flexibility, it might be useful to review what the architecture actually provides over a general-purpose programming environment. Apart from the basic hypothesis hierarchy and rule system framework, the key aspect of the system is the emphasis on analysis by prediction. Added abstractions must provide predictions so that the basic analysis-prediction loop can continue, allowing the advantages of inference and context-sensitivity to accrue. The implementation provides a base set of rules and abstractions around which larger systems may be built. The specifics of the rule system developed so far are presented in chapter 4.

The prediction-driven architecture as a theory of auditory perception: In the introduction, I emphasized that the goal of this work was to model human auditory organization both in terms of its functional output (so far as that can be established) and in terms of its internal processing by which it arrives at those results. But the presented architecture is based on specific computer-science ideas, such as the blackboard structure, that have only the most tenuous relation to human information processing. Has the goal of modeling the internal operation of the auditory system been abandoned?

Well, not deliberately. Systems described as seeking to duplicate the internal operation of the brain often involve models of neural circuits: the operation to be duplicated is conceived at a very literal level. However, as discussed in chapter 2, and as powerfully argued by Marr [Marr82], there are many levels at which to model the function of a perceptual system. While symbolic reasoning systems such as the prediction-driven architecture raise profound questions over how they could be implemented with neurons, the fact remains that such systems often provide an intuitively satisfying reproduction of the more abstract aspects of our cognition (for instance, the classic planning systems of [NewS72] or the broader theories of mind in [Minsky86]). The prediction-driven architecture is offered in this vein: I do not venture to suggest that exact analogs of hypotheses or the flow of abstraction will be found (although they might exist). However, the strength of the architecture, and its claim as a model of perceptual processing, lie in its ability to accommodate actual phenomena of restoration and inference; it is difficult to think of any system for which this could be true other than one more or less based on the concepts that have been presented.

At this stage we have seen the kinds of phenomena motivating the new model architecture, and discussed several aspects of why the architecture that has been presented is an appropriate response to these challenges. These rather abstract ideas will become more concrete in the next chapters, where the architecture is examined through the description of an implementation that has been developed to test these ideas.

Chapter 4 The implementation

This chapter describes the computer implementation of a system that follows the prediction-driven principles of chapter 3. This 'sound organization' system analyzes real sound scenes into an abstraction founded on generic sound elements through a process of incremental prediction and reconciliation between the observed sound and the internal representation. Multiple hypotheses concerning the explanation of the actual sound compete through their ability to make accurate predictions.

The current system cannot be considered complete, certainly not as a model of human auditory organization, nor even as an implementation of the prediction-driven architecture. However, it does serve to make concrete many of the concepts involved, and despite its limitations it exhibits interesting and useful behavior in the analysis of real-world sounds. These will be described in chapter 5, Results.

4.1 Implementation overview

4.1.1 Main modules

Before describing each system component in detail, a brief overview of the broad system structure and operation will give a context for the more specific descriptions. Reflecting the structure introduced in chapter 3 and repeated in figure 4.1, the implementation divides into four pieces:

- **The front-end:** Fixed numerical signal-processing algorithms are applied to the raw acoustic signal to translate it into the domains in which the prediction and explanation occur. The main pieces of the front end are a filterbank to model the frequency-selective decomposition of the cochlea, and a correlogram-based periodicity-detection scheme acting as the foundation for periodicity-based signal grouping and explanation. There is considerable evidence for some structure broadly comparable to the correlogram in the early auditory physiology [Lang92].
- **Core representational elements:** These are the pieces from which a model explanation of the observed sound is constructed. As introduced in the previous chapter, there are three different kinds of elements, specifically aimed at noisy sounds, transient sounds and pitched sounds. Each kind of element is characterized by a set of internal constraints, which, in conjunction with the parameters for a particular instance, define the sound-object to which it corresponds. Resynthesis into sound for individual elements can be accomplished based on this information.
- **The prediction-driven engine:** The blackboard-based collection of rules and procedures comprises the engine by which the internal model of the sound is constructed, and its predictions are reconciled to the correlates of the observed sound. This portion includes the management of competing hypothesis-explanations and addresses issues such as the allocation of prediction error between overlapping elements.

- **Higher-level abstractions:** Larger-scale organizations of the basic elements to accommodate more complex structures present in the real sound. Although rather inadequately developed in the current implementation, this open-ended explanatory hierarchy gives the prediction-driven system the potential to recognize and track arbitrarily complex signals even when they are obscured – the kind of phenomena observed in auditory restoration.

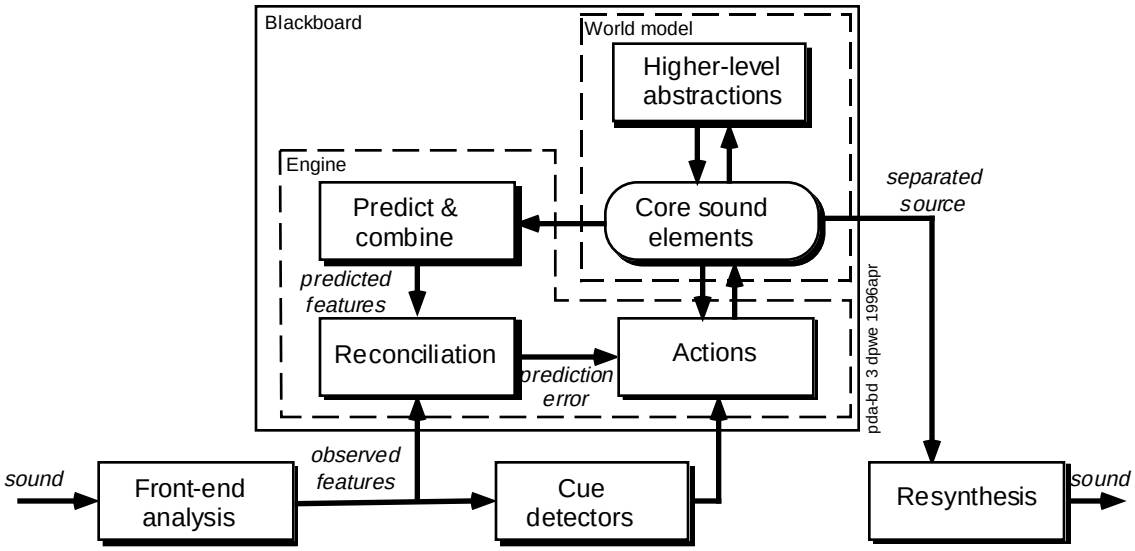

Figure 4.1: The main modules of the prediction-driven architecture (as fig. 3.3).

4.1.2 Overview of operation: prediction and reconciliation

The prediction-driven architecture specifies that an internal model should be developed through reconciliation of its predictions with external observations; in practical terms, this proceeds as follows: At a given moment in time, the internal configuration of sound elements and the abstractions they support allows a prediction to be made of the observed signal for the next instant. For example, a noise element, in the absence of a more sophisticated interpretation, will predict that the next time slice will contain the same noise spectrum that was the model for the current slice, and a click element will predict that the energy from its transient will have decayed from the current values at the rates assigned to each channel. Where the elements are the support for higher-level abstractions, the predictions could be correspondingly more sophisticated.

The predictions that are made are probabilistic, that is, the *expected* value of the signal is given along with a margin of uncertainty, encoded as separate *deviation bounds* for positive and negative differences from the expectation. This variance in the prediction arises both from uncertainty in the parameters of the predicting model, and from the fact that certain models cannot exactly specify the future form of their element's signal (such as noise clouds, which contain an nondeterministic component). At a given time, there may be several competing hypotheses as to the correct explanation of the observed signal, each with its own set of elements. Within each group, the predictions of each element are gathered and combined according to rules specific for each group; generally the elements are considered to be incoherent, and thus their predictions add linearly in the power domain.

When the information for the new time slice arrives, it is compared with the prediction that has been made. If the actual signal is within the error bounds specified, the prediction is deemed to have been adequate, and no changes to the set of elements is needed. The engine simply notifies the elements of the actual signal observed (to allow them to update their parameters), then continues. On the other hand, if there is too much energy in the observed signal to be accountable with the prediction, the engine seeks to create an additional element to 'soak up' that energy. If the observed signal contains less energy than the minimum predicted, the engine will try to terminate some of the existing elements to permit a consistent explanation.

In the remainder of this chapter, each of the main modules is considered in detail, giving a complete picture of the overall implementation.

4.2 The front end

The numerical processing of the front-end converts the raw acoustic signal (one-dimensional pressure variation as a function of time) into the domains in which explanations must be provided. As the first stage of a model of the auditory system, it has the strongest correspondence to an identifiable portion of the auditory physiology – namely the periphery of the cochlea and immediately succeeding neural centers. However, even in this most literal subsystem, the correspondence is inevitably stylized, reflecting the beliefs implicit in this model about the relative importance of various aspects of peripheral processing. The arrangement of front-end processing modules is illustrated in figure 4.2 and each is discussed below.

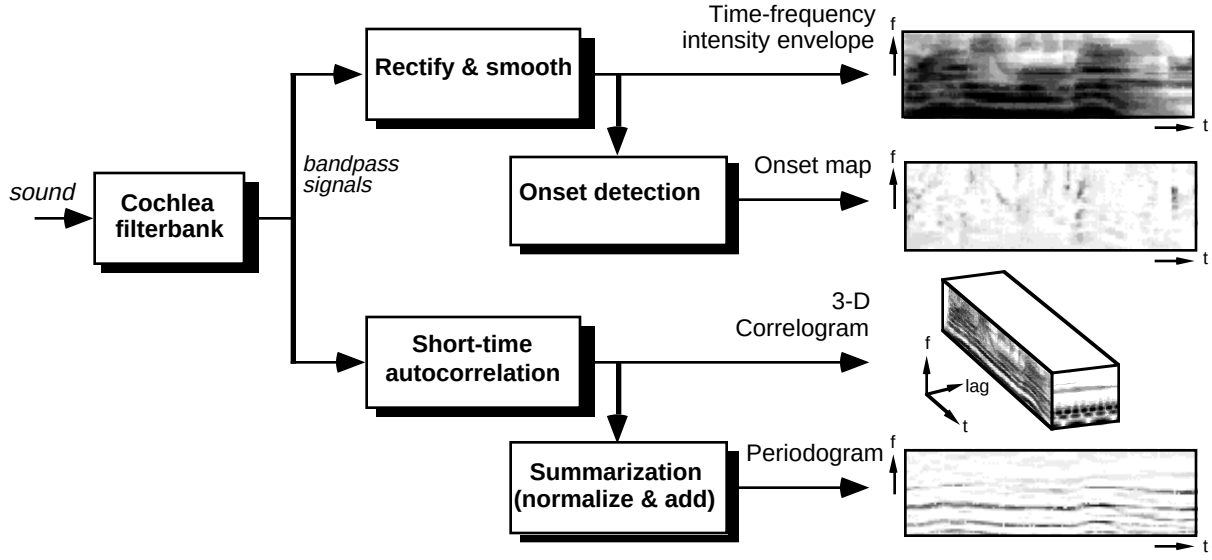

Figure 4.2: The various components of the front-end processing in this implementation.

4.2.1 Cochlear filterbank

In common with other models of auditory processing, the analysis begins with a frequency decomposition corresponding to the frequency-selective resonant properties of the cochlea. The particular implementation used here is based on the model of [PattM86] [PattH90] which has been widely used, for

instance by [Brown92] and [MeddH91] – my particular implementation is based on the work of [Slaney93]. It has the following properties:

- **Fixed, linear array of bandpass filters:** It has long been known, through measurements of auditory nerve fiber firings in response to swept tones, that the effective tuning of the cochlea is variable, tending to broaden or saturate with increasing intensity. However, for the purposes of this model, this effect is assumed to be secondary to the breakup of the signal into different frequency bands which is taken as the most important function of the cochlea; the filterbank used here models only this primary function. Note that several models have addressed the nonlinearity of frequency response [Sieb68] [Ghitza88] [Slaney88] [GigW94], but the manner in which this aspect of the physiology benefits the process of hearing is not clear. A linear filterbank, on the other hand, has the advantages of efficient and well-characterized operation, favoring this selection in the absence of a more concrete reason for its rejection. Preserving as much linearity as possible also greatly aids the inverse problem of resynthesis, although carefully-designed nonlinear elements can similarly support this [SlanNL94].
- **Simple 8th-order IIR filter structure:** The basic 'gammatone' filter structure advocated by [PattM86] [PattH90] is simply a cascade of repeated pole-pairs, in this case four pairs. Although this filter structure represents a good approximation in view of its great computational simplicity [Irino95], it lacks the very abrupt high-frequency rolloff observed in the actual cochlea transmission line (as modeled in [Slaney88]), which some researchers have argued is critical to the excellent frequency discrimination of hearing [Sham89] [Wein85].
- **Bandwidth increasing with center frequency:** In middle and high frequencies, the bandpass filter characteristics of the ear seem well characterized as constant-Q i.e. with bandwidth proportional to center frequency (when measured under comparable conditions) [MooreG83]. At low frequencies, cochlea filter bandwidths appear to plateau, meaning that a purely constant-Q model will result in unrealistically narrow filters in this region. Although experimentation failed to reveal any significant qualitative difference between analysis based on strictly constant-Q analysis and a more realistic distribution of bandwidths, imposing a lower limit on channel bandwidths does serve to keep the impulse response time-support (and hence the time-blurring) down to a manageable level in the lower bins, and was adopted for this practical reason. The filter bandwidths are those in the implementation of [Slaney93], based on the recommendations of [MooreG83], with the equivalent rectangular bandwidth (ERB) approaching a Q of 9.3 at the high frequencies, and bottoming-out at around 250 Hz, representing a good match to physiological measurements of real cochleae.
- **Logarithmic spacing of filter center frequencies:** For a constant-Q filter bank, the most natural filter spacing is in proportion to those frequencies, i.e. a uniform density of filters when plotted against a logarithmic frequency axis. This makes the responses cross at the same level relative to the maximum in each filter, and again seems to reflect both the relevant physiological and psychological knowledge. Although the filterbank used was not strictly constant-Q, exact logarithmic spacing was used for computational convenience, with the effect that the lowest frequency channels are rather more overlapped than the rest. The filter density used was six samples per octave (i.e. each filter's center frequency

approximately 12% larger than its predecessor); this gives a generous overlap between filter responses. A greater density of filters would contribute only redundant information to the subsequent processing. Real cochleae contain several thousand inner-hair cell structures collecting information on the energy in highly overlapped frequency bands [Pick88], a density much greater than this model. However, the ear has additional challenges such as the noisiness of neural encoding to overcome, which may provide an explanation for this very dense sampling.

The composite frequency response of the cochlea filter bank is show in figure 4.3 The sampling rate of the original sounds for this implementation was fixed at 22.05 kHz, thus the Nyquist rate for these filters is at 11 kHz. There are 40 filters used starting from 100 Hz, which, at six per octave, places the highest center frequency at 10160 Hz, or just below Nyquist. Although human hearing extends as much as an octave above this frequency, the assumption is that in the great majority of cases information outside this range is of no great environmental significance, and can safely be overlooked for the purposes of this work. 100 Hz was adopted as a lower limit for the pragmatic reason that the filterbank implementation becomes unstable in the octave below this, and it was again assumed that it is rare for the loss of bandwidth below 100 Hz to impair the prospects for organizing a given sound significantly. In particular, following real audition, the periodicity detecting mechanisms in this model are not disturbed by the absence of a fundamental, so the lower limit on peripheral frequency analysis does not correspond to a limit on pitch perception.

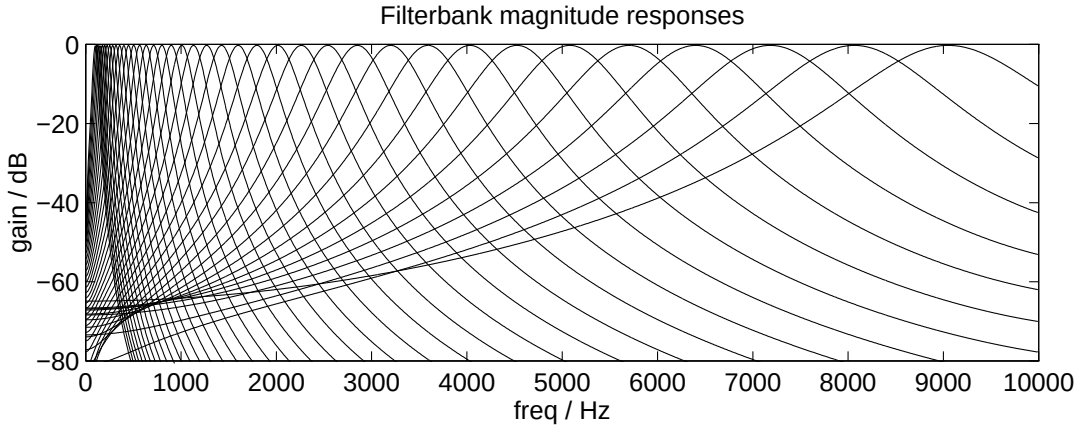

Figure 4.3: The magnitude-frequency responses of the 40 linear filters in the front-end filterbank.

The phase response of this kind of pole-zero filter is of course nonlinear. For convenience, a pure delay was added to the output of each filter so that the theoretical envelope maximum of the impulse response would be aligned in each frequency channel. It is interesting to note that the propagation delay down the cochlea introduces a delay to the lower frequencies just as exhibited by this kind of filterbank. Patterson [Patt87] found no unusual perceptual effect of pre-compensating for this delay (arranging for all the stimulated nerves in the auditory fiber to fire simultaneously); we might guess that the higher auditory centers incorporate some equivalent to these delays to remove the systematic but uninformative between-channel time differences.

4.2.2 Time-frequency intensity envelope

The multiple band-pass signals generated by the filterbank constitute a multiplication of the amount of data used to represent the raw sound signal, which is completely recoverable from those outputs; this is not yet a particularly interesting representation of the sound i.e. one that has 'thrown away' some information deemed to be unimportant [EllisR95]. The next module addresses this, by converting the array of bandpass signals at the full sampling rate into a much more sparsely sampled unipolar intensity envelope in time-frequency. This envelope is calculated by half-wave rectifying the output of each frequency channel, squaring this signal, then smoothing the result through a simple one-pole lowpass filter with a time constant of 25 ms. (These particular choices were made for compatibility with the correlogram analysis, explained below). The smoothed output is subsampled at 5 ms intervals. The sort of envelope returned by this procedure is illustrated in figure 4.4, against the corresponding band-pass signal, in this case the output of a filter centered at 4 kHz.

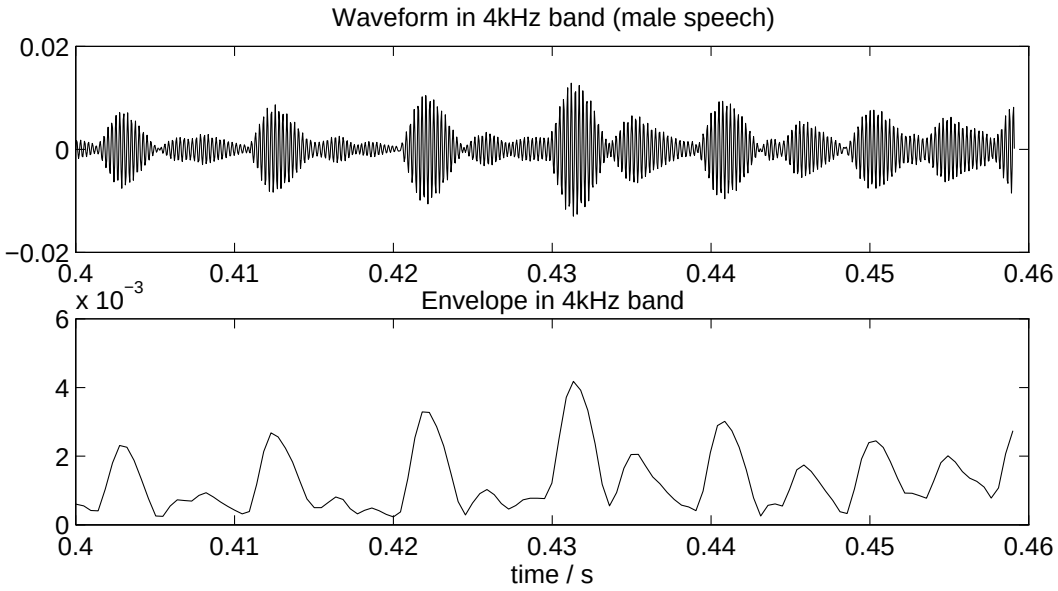

Figure 4.4: Typical bandpass signal at 4 kHz and its subsampled envelope representation.

The rationale behind this representation, as reflected by its role in the subsequent processing, is that it is on approximately this scale that energy present in the incident sound must be explained. 5 ms is comparable to the latencies and propagation delays of the basic neural hardware of the brain; any sensitivities finer than this timescale will rely on special-purpose dedicated sensors which will presumably only be employed in situations where such information is particularly useful. Such situations include spatial location through interaural time difference (which has a resolution on the order of tens of *micro*seconds [ColbD78]), and pitch detection based on temporal structure (which needs sub-microsecond accuracy to account for our pitch discrimination at fundamental frequencies up to 1 kHz and beyond). However, for higher level phenomena, such as the jnd of note onsets in music perception, a few milliseconds seems to be the upper limit [Palmer88], roughly equivalent to that afforded by the subsampled representation calculated in this module.

Although the front-end described so far has made many compromises compared with an exact model of the auditory periphery, it is still credible that a representation of essentially this kind is employed within the auditory system. Summation of all the nerves excited by a given frequency band, smoothed by the inherent sluggishness of a neuron, would give a measure of the net intensity within a given frequency channel over some smoothed timescale somewhat similar to this envelope parameter. A typical time-frequency intensity envelope for a fragment of speech is illustrated in figure 4.5.

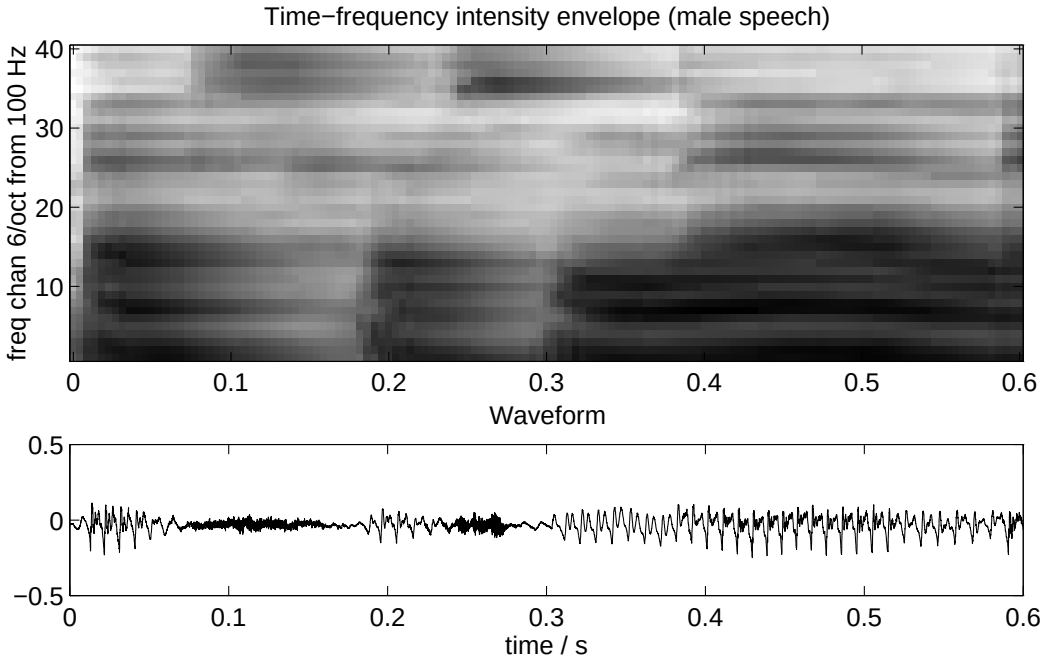

Figure 4.5: Example time-frequency intensity envelope for a short sample of male speech, "this is mere...". The speech waveform is shown beneath.

4.2.3 Correlogram

The intensity envelope discards all the information in the fine structure of the bandpass-filtered signal which is however indicated to be perceptually salient by many hearing phenomena. Thus the front-end must include additional processing to extract some of these features. Pitch is the most significant perceptual attribute that is obscured by the particular combination of broad frequency channels and heavy frequency smoothing of the intensity envelope, hence the front-end includes an autocorrelation-based processor to permit the detection and extraction of common periodicity in different frequency bands, the usual cause of a pitch percept.

Considerable success in explaining pitch phenomena has been obtained in the models of [MeddH91]. The basic idea is to measure the similarity of a signal in a given frequency channel to time-delayed versions of itself, a process generally achieved by autocorrelation (the inner product between the shifted and unshifted versions of the signal). In the three-dimensional correlogram representation [SlanL92] [DudaLS90] (used in the auditory scene analysis models of [Mell91] and [Brown92]), the short-time autocorrelation as a function of delay or 'lag', is calculated for every frequency channel at

successive time steps, giving an intensity *volume* as a function of lag, frequency and time. If the original sound contains a signal that is approximately periodic – such as voiced speech – then each frequency channel excited by that signal with have a high similarity to itself delayed by the period of repetition, leading to a 'ridge' along the frequency axis at that lag in the correlogram. Note that the relatively broad tuning of the bandpass filters, along with a bandwidth that increases with frequency, mean that the upper spectrum of a broadband periodic signal will be analyzed not as the resolved harmonics of a high-resolution, fixed-bandwidth Fourier analysis, but as a wider-bandwidth signal showing *amplitude modulation* at the fundamental period. Although a resolved harmonic would indeed show an autocorrelation peak at the modulation period (since harmonics have periods that are integer subdivisions of the fundamental, and autocorrelation gives peaks at all multiples of the period), it would not be locally distinct from adjacent peaks. However, with the amplitude-modulated broader-band energy, autocorrelation reveals the underlying common period far more clearly.

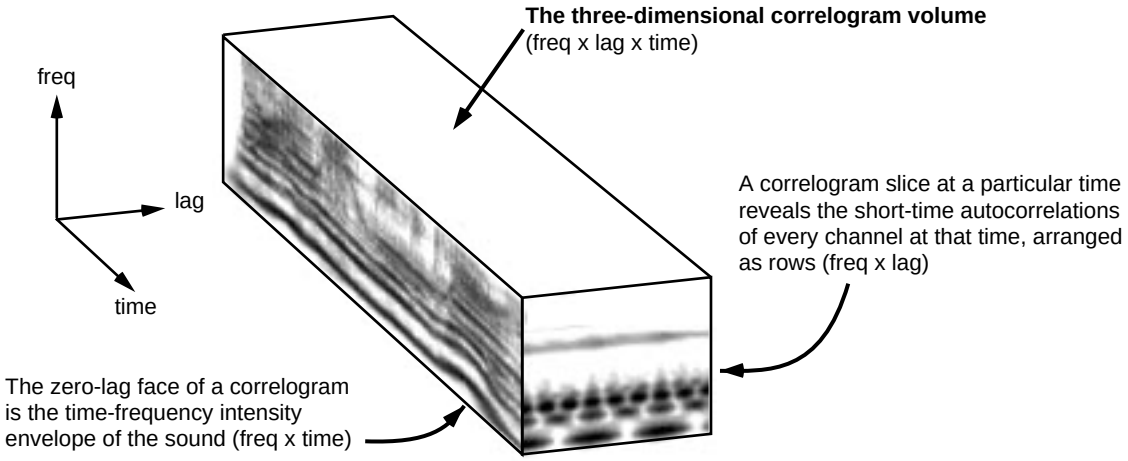

Figure 4.6: Illustration of the correlogram volume, which represents signal intensity as a function of three axes: time, frequency and short-time autocorrelation lag. A correlogram slice (a two-dimension function of time and frequency at a specific time instant) shows the periodicities present in each frequency channel at that instant; the extreme of the volume where the autocorrelation lag is zero is effectively the time-frequency intensity envelope of the signal.

In the straightforward autocorrelation of an amplitude-modulated, bandpass signal from the upper frequency region (i.e. far above the fundamental of the periodic modulation), we would see fine structure at lags corresponding to the center-frequency of the channel resulting from the autocorrelation of the carrier frequency; superimposed on this fine structure would be the broader effect of the amplitude modulation, revealing the pitch-related information which is our particular interest. If the fine structure occurs at a period outside the useful pitch range (which begins to run out at about 1 kHz, although the piano continues for another two octaves [Pierce83]), it is not of interest, and we should take just the envelope of the autocorrelation by some process of rectification and smoothing. But if this is our eventual intention, we might as well take the signal envelope *before* autocorrelation, since the autocorrelation may be more cheaply performed on the lower-bandwidth envelope signal than on the full-bandwidth frequency channel – an argument

that probably applies equally to the computer model and to the biological prototype. Thus it is the bandpass signal *envelope*, showing the amplitude modulation but not the high-frequency carrier, that is autocorrelated. The envelope is calculated by half-wave rectifying the bandpass signal, then applying light smoothing over 1 ms time window to remove supra-pitch-rate information. Half-wave rectification is important here to ensure that the autocorrelation of a resolved sinusoid shows autocorrelation maxima only at its true period and not at half that period, as would result from full-wave rectification.

The effect of autocorrelating the full-bandwidth signal and its half-wave rectified derived envelope are compare in figure 4.7 for the 4 kHz band of some male voiced speech.

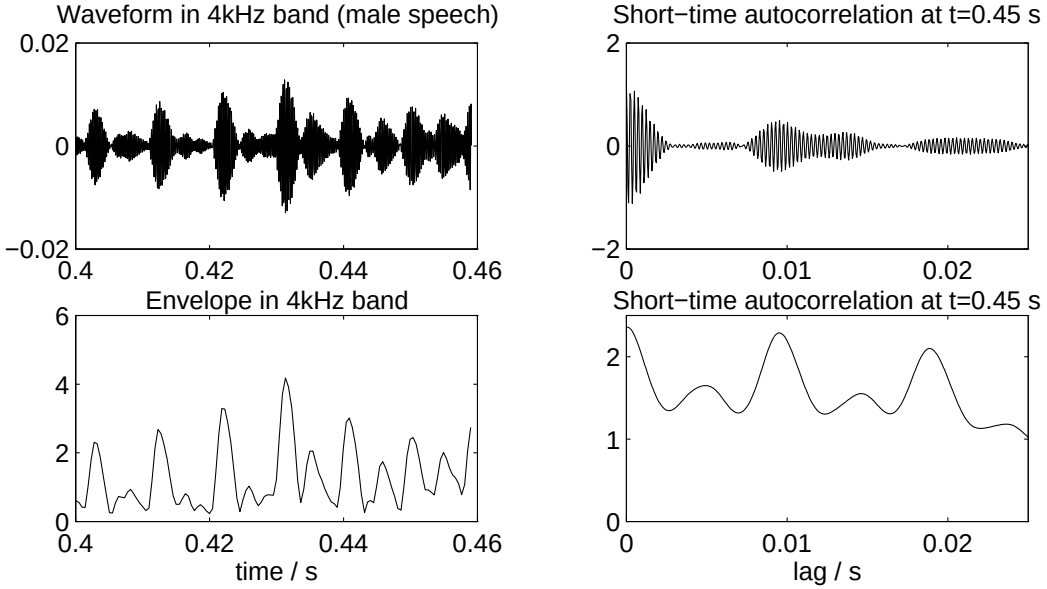

Figure 4.7: Comparison of autocorrelation of a 4 kHz frequency band excited by voiced speech with a fundamental frequency of around 110 Hz. The top half shows the unprocessed filterbank output and its autocorrelation. The lower two panes shows the envelope, downsampled by a factor of 10, and its autocorrelation. Using the envelope smoothes out the fine structure in the autocorrelation while retaining the correlates of the fundamental modulation (the peaks at 9 and 18 ms).

Sampling the lag axis

There now arises the interesting question of the lags at which to sample the autocorrelation. In discrete-time systems, it is generally convenient to calculate autocorrelations at multiples of a fixed lag such as the sampling period; this also makes sense on information-theoretic grounds, since the bandwidth of the autocorrelation signal will be the same as the original, and thus it can safely be sampled at the same density without loss of information. However, the pitch-detection goal of this processing suggests a quite different sampling: Fundamental-frequency discrimination, like the peripheral frequency analysis, is roughly constant-Q, that is, the just noticeable difference is proportional to the frequency itself over the major range of pitch perception [Moore89]. It would be natural to have an autocorrelation lag axis with a sampling density that followed the perceptual resolution for fundamental frequency, which this feature will be used to model. This would

suggest a log-time sampling of the autocorrelation function, with the shortest lag corresponding to the highest frequency to be considered a fundamental (set somewhat arbitrarily at around 1 kHz), and successive lags each some constant factor larger than their predecessors, out to a maximum lag corresponding to the lowest detectable fundamental period. This arrangement gives a constant number of autocorrelation bins per octave of fundamental frequency. The autocorrelation function will be oversampled at the low-period end (adjacent bins corresponding to fractional delays differing by less than one discrete-time sample) and potentially undersampled at the long-period extreme, unless the signal to be autocorrelated is smoothed appropriately. It has the advantage of providing a 'natural' view of the lag axis, with each octave of period detection occupying the same space. This was the scheme employed; in order to minimize the smoothing necessary to avoid undersampling, a rather dense sampling of 48 lags per octave was chosen, and the periods measured ranged over five octaves from 40 Hz to 1280 Hz. A comparison of linear- and log- time sampled autocorrelations is illustrated in figure 4.8.

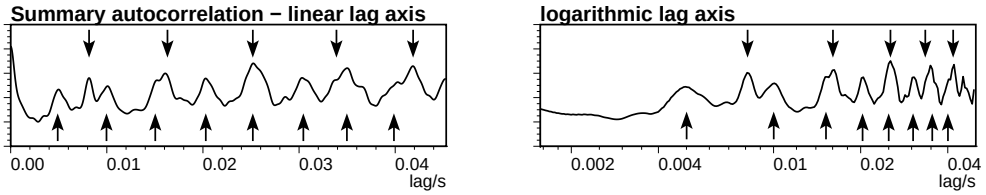

Figure 4.8: Comparison of linear- and log-time sampling of the autocorrelation lag axis. The left panel shows a single short-time autocorrelation function for a channel containing a mixture of male and female voices. Note how the two different periods present result in regularly-spaced sets of autocorrelation maxima with different spacings indicating the different periods. (The arrows above the trace indicate the peaks resulting from the lower 120 Hz male voice; the arrows underneath the trace point to the peaks arising from the 200 Hz female voice). The same function, displayed with log-time sampling in the right panel, shows each period resulting in a series of peaks that get closer together along the lag axis, but for which the difference in fundamental period results in a simple shift rather than the dilation apparent in linear sampling. Note also that 'zero lag' cannot be included on a finite-sized log-time axis.

Calculation of short-time autocorrelation

Although plain autocorrelation is uniquely defined, there are two approaches to calculating short-time autocorrelation that should be distinguished. If the autocorrelation of $x(t)$ is defined as a function of lag τ as:

$$R_{xx}(\tau) = \int_{-\infty}^{\infty} x(t)x(t-\tau)dt \qquad (4.1)$$

then one definition of the short-time autocorrelation at time t_0 would be to window x at time t_0 and autocorrelate that short time version of x, i.e. if $w(t)$ is a local time window, the windowed version of x is:

$$x_w(t, t_0) = x(t)w(t-t_0) \qquad (4.2)$$

and the short-time autocorrelation at time t_0 is:

$$R_{x_w x_w}(\tau, t_0) = \int_{-\infty}^{\infty} x_w(t, t_0) x_w(t-\tau, t_0)\, dt \qquad (4.3)$$

This however has the disadvantage that longer lags have attenuated correlation owing to the narrowing effective window, $w(t,t_0)\cdot w(t-\tau,t_0)$, in the integral. This is avoided by separating the window from the signal and applying it directly inside the integral:

$$R_{wxx}(\tau, t_0) = \int_{-\infty}^{\infty} w^2(t-t_0) x(t) x(t-\tau)\, dt \qquad (4.4)$$

Unlike (4.3), this form allows short-time autocorrelations to be calculated for lags longer than the time window without difficulty. Further, (4.4) may be calculated as the product of the delayed and undelayed signals smoothed by a filter whose impulse response is $w^2(-t)$:

$$R_{wxx}(\tau) = \left[x(t) x(t-\tau) \right] * w^2(-t) \qquad (4.5)$$

where the short-time index t_0 has been dropped to indicate the implicit time variation of the output of the convolution. This smoothing of the product of the signal multiplied with a version of itself, delayed by the lag, is the approach used to calculate the correlogram here, as illustrated in the block diagram in figure 4.9. The smoothing is accomplished by simple one-pole lowpass filters with time constants of 25 ms, determined experimentally to be a good compromise of eliminating pitch-rate variation while responding to rapid changes in the signal. The intensity envelope described above uses the same smoothing filter so that it is, in fact, equivalent to the correlogram output for zero delay (and can, incidentally, be used to normalize the correlogram rows for the periodogram, described below).

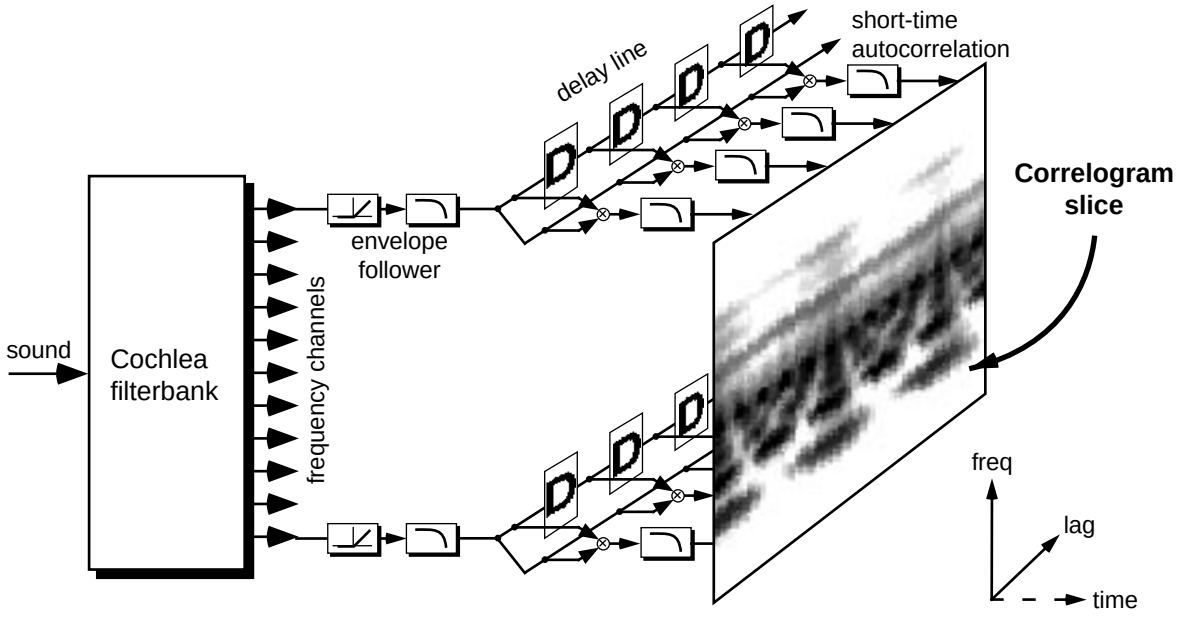

Figure 4.9: Block diagram of the calculation of the correlogram.

The disadvantage of this approach to calculating autocorrelation is that it cannot be implemented using the very efficient fast Fourier transform (FFT) technique, which is the motivation for the use of the form of eqn (4.3) in certain published auditory models [SlanL92]. On the other hand, the log-time sampling chosen for this system is not particularly efficient to derive from the linear-time sampling of the FFT method, whereas it is conveniently incorporated into the delay-and-multiply scheme by setting the appropriate fractional delay in the delay line. The delay-multiply-and-smooth structure is very physiologically plausible and was indeed first proposed as an auditory periodicity detection mechanism over four decades ago by Licklider [Lick51].

Note that this autocorrelation scheme is roughly equivalent to its physiologically-motivated antecedents[MeddH91] [SlanL92]. One distinction is that the use of nonlinear elements (rectification) has been kept to the minimum, used only to calculate the subband envelope prior to autocorrelation. Physiological models generally include an equivalent element representing the effective half-wave rectification of the inner hair cells of the cochlea, along with some kind of amplitude-range compression and perhaps an argument about the representation of the ensemble effect of many spike-carrying nerve fibers as a firing probability suitable for autocorrelation. Limits to phase-locking observed in auditory nerve firings run out at a few kHz, meaning that temporal fine structure information is presumably limited to this bandwidth, which can justify the kind of smoothing included here to remove unwanted non-pitch information; in this model, the signal is rectified as a necessary part of smoothing to obtain an envelope. Rectification of the signal prior to autocorrelation has the additional benefit of guaranteeing a nonnegative autocorrelation function, avoiding the question of how negative quantities might be encoded in nerve firings. While it is reassuring that the current structure is still close to models motivated by explicit physiological considerations, the limited concessions to known physiology in this model are based on the principle that processing should be motivated only by functional considerations, and not include aspects of the physiology whose practical significance cannot be explained.

4.2.4 Summary autocorrelation (periodogram)

The correlogram evidently contains useful information reflecting periodic modulations in the original signal, but we have yet to describe how it may be used to advantage in the analysis. As mentioned, the most significant aspect of correlogram analysis is the way in which periodic modulation that affects multiple frequency channels will be displayed: Such features will appear as a distinct vertical structures on a two-dimensional slice of the correlogram at a given instant, lying on a line of constant lag matching the fundamental period of the excitation. A structure of this kind is illustrated in fig. 4.10.

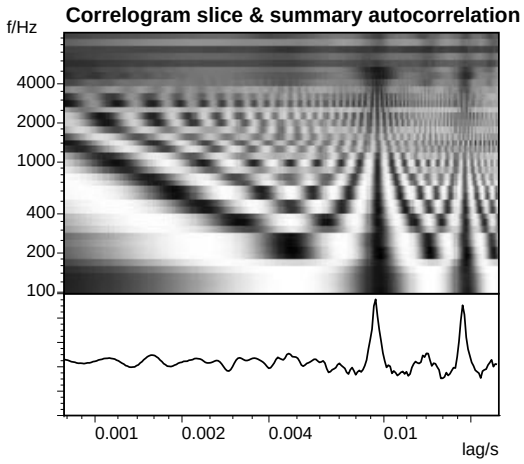

Figure 4.10: Instantaneous slice of the correlogram analysis for some voiced male speech, showing how the common modulation period of 9.0 milliseconds appears as an intensity peak at this lag across many frequency channels. This vertical structure is the basis of pitch detection. The lower trace shows the summary autocorrelation, with pronounced peaks at 9 ms and its alias 18 ms.

As an intermediate stage in extracting the perceived 'pitched' objects in a sound, it would be valuable to reduce the correlogram volume to a more manageable two-dimensional function of fundamental period likelihood versus time, known as a *periodogram*. Each vertical slice of the periodogram represents the net short-time autocorrelation information for all frequency channels at a particular time instant and is normally termed the *summary autocorrelation*. A common approach is to collapse the frequency channel axis of a correlogram slice (intensity as a function of autocorrelation lag and frequency channel) by adding all the autocorrelations together [MeddH91] [Brown92]. Some normalization is required to equalize the significance of peaks in channels whose absolute energy might be quite small compared to other remote frequency channels; their contribution should be comparable both in view of the relative indifference of the auditory system to static spectral imbalances, and on the information-theoretic grounds that the evidence for a certain modulation period in a frequency channel is not proportional to the total energy in that channel, but depends instead on its signal-to-interference ratio.

In physiological models, the nonlinear hair-cell model normally accomplishes within-channel normalization more or less explicitly [MeddH91] [SlanL92]. Further normalization of each row of the correlogram slice may be applied based on the peak value, at lag zero, which follows the total energy of the signal in that channel [Wein85]. An alternative approach is to select all local maxima in each autocorrelation function, regardless of absolute amplitude, and sum them up across channels with unit amplitude [EllisR95], although this loses useful information from the relative height of different peaks within a single autocorrelation function. In the current model, I adopted straightforward normalization by channel energy before averaging across channels. Thus the summary autocorrelation for the lag of zero (if indeed it were included on the log-time axis) would always be unity.

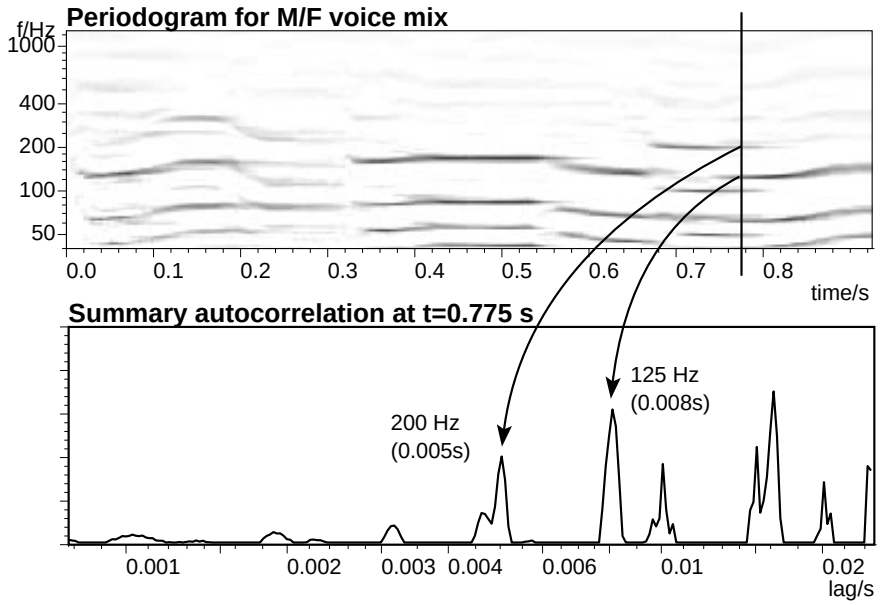

Figure 4.11: Periodogram (intensity as a function of lag period and time) for a mixture of male and female voices (fundamental periods of 125 Hz and 200 Hz respectively), along with a slice (a summary autocorrelation) at t =0.775s. The different periods of the two voices are apparent, along with their aliases at multiples of the periods. Note the log-time spacing on the lag axis is flipped in the periodogram display so that the short lags are at the top, permitting it to be labeled in frequencies instead, forming an equivalent log-frequency axis. This convenience of display facilitates visual comparison with other time-log-frequency displays such as the time-frequency intensity envelope, although the information displayed is of course quite different.

The result is a single 'periodicity intensity' as a function of period lag for each time instant, a two-dimensional function (the periodogram) illustrated in figure 4.11 along with a representative one-dimensional slice (a summary autocorrelation). The local maxima of the summary autocorrelation indicate the dominant periods in the full signal and are used as the basis for the 'weft' periodic elements as described later.

4.2.5 Other front-end processing

The approach used in this implementation, partially on practical grounds, is to conduct all the numerically-intensive front-end processing in an independent module whose results form the input to the context-sensitive explanations of the prediction-driven analysis. In the previous chapter, a distinction was drawn between indispensable and optional cues, the former being those that the abstraction must explain (such as positive energy in some region of time-frequency), and the latter being other functions of the input signal that might only be consulted in special circumstances. The autocorrelation volume, while not subject to prediction itself, is consulted in the characterization of the periodic weft elements (described below) and thus is a kind of optional cue; the periodogram however must be fully explained and is thus indispensable.

Onset map

A second optional cue is the onset map, which is used to help decide whether to create a new click element to explain an excess in observed energy. The basic form of an onset map is very simple, and is a feature common to previous auditory organization systems [Brown92] [Mell91]. Brown's approach was rooted strongly in known aspects of auditory physiology, including the existence of populations of neurons that fire in response to abrupt *increases* in the energy of a given peripheral frequency channel. His system ran a range of band-pass filters over the intensity in each channel, which, when rectified, indicate energy onsets over different timescales – to account for the salience of both rapid and more gradual energy increases in real sounds. Brown uses the term 'map' to refer to a multidimensional array of sensors that respond to conjunctions of parameters varying systematically over each dimension – such as a 2-D array of onset-detectors tuned to combinations of peripheral frequency channel and onset rise-time.

The current implementation uses a simpler scheme in which a single onset score is calculated for each time-frequency cell to indicate if an abrupt increase in energy is occurring. The relatively coarse 5 ms time-step means that for most channels simply using the increase over the preceding time step is a sufficient indication of energy increase. In order to normalize for the absolute level in the channel, this differencing is done in the logarithmic domain (i.e. on the channel energy in decibels). Consequently, if the energy in a channel doubles in the space of one time step, the onset score will be 3 dB regardless of the absolute energy in the channel. Some of the lower frequency channels have sufficiently narrow bandwidth that even the most abrupt onset is spread over several time steps; to avoid penalizing the onset scores in these channels compared to the faster-responding higher-frequency channels, the onset scores of the four temporally-adjacent time-steps are weighted and added-in to the overall score; the weights depend on the channel bandwidth compared to the frame-rate. Since the onset map is only consulted when reconciliation suggests the need for a new element, it is less important if the map registers extra features where no onset event is generated; the important quality is that it should be able to reject candidate onsets from the reconciliation that occur when the energy envelope does not exhibit a rapid increase in energy.

Figure 4.12 shows the onset map for a brief fragment of speech shown alongside its energy envelope. The onset score can be seen to follow energy onset even for the low-energy channels, and even for the more gradually-increasing low frequency channels. The effect of weighting several successive time steps in the lower channels causes the onsets to be considerably blurred in time for those channels.

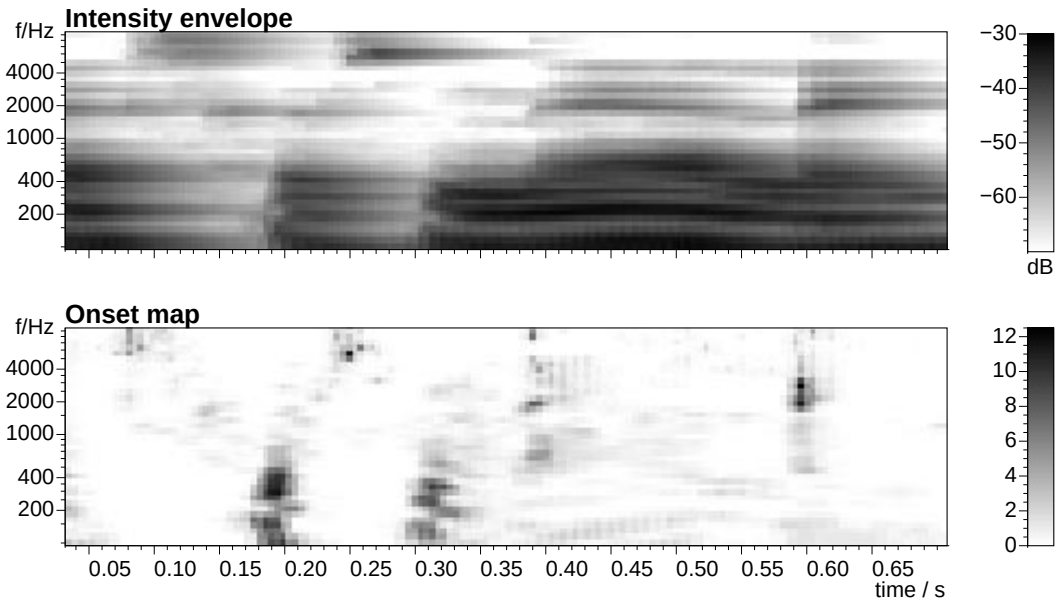

Figure 4.12: Onset score values for a brief fragment of speech (lower panel) compared to the energy envelope for the same piece of sound (top panel). The onset score is normalized for absolute channel energy, and summed across several time-steps in the lower frequency channels in order to capture the full extent of slower onsets in these channels. Thus the onset at t=0.40s is strongest in the highest channel, even though the level in this channel is relatively small.

Fixed front-end processing is of course anathema to the IPUS architecture, which is motivated by the idea that numerical processing should be determined by context inferred from the symbolic abstraction created for the signal [LessNK95]. From a different point of view, physiological results concerning the variable characteristics of the filtering effected by the cochlea have led to more signal-dependent adaptive models [GigW94], and this gains credibility from speculation over the role of the efferent auditory nerve fibers, carrying information back from the auditory centers in the brain right to the outer hair cell bundles on the cochlea. However, patients who have lost their efferent fibers are not reported as having a vastly impaired ability to hear [Scharf94], and for the purposes of this model, no particular disadvantage was noticed arising from the use of fixed front-end processing.

4.3 Representational elements

In the prediction-driven architecture, the process of analysis is governed by the goal of generating an internal representation that is consistent with cues derived from the actual sound. This representation is expressed at its lowest level as a collection of generic sound elements which capture the perceptually-important qualities of the sound, while at the same time embodying inherent constraints on the useful structure that the perceptual system can extract from a sound. As introduced in the previous chapter, the current implementation includes three kinds of generic sound element, aimed at representing noisy sound energy, short-duration transients, and pitched sounds respectively. While this is a rather limited vocabulary, one of the purposes of this implementation is to see how successfully such elements may express complex, arbitrary sounds as a way to assess the implicit hypothesis

that the auditory system employs primitives of this kind. The details of the internal constraints, analysis and synthesis methods for each kind of element are now presented.

4.3.1 Noise clouds

The goal of the noise-cloud element is the representation of areas of sound energy with no obvious local structure, but which none-the-less are perceived as, and typically generated by, a single source. By contrast, energy that has coherent intensity variation with time, or energy with some degree of periodicity, has a more substantial basis for its organization and is handled by different elements. However, a great many everyday sounds fall into this "noisy" category, from voiceless fricatives of speech, to the sound of wind blowing, to the crunching of paper. As a rule, such sounds have been neglected by previous models of auditory organization in favor of the more structured class of pseudoperiodic sounds (exceptions being the noise beds of [LessNK95] and the stochastic components of [Serra89]).

Signal model

In the noise cloud elements, sound is modeled as a white noise process to which a static frequency envelope and a slowly-varying time envelope have been applied. Thus, in the time domain, the signal is modeled as:

$$x_N(t) = h(t) * [A(t) \cdot n(t)] \qquad (4.6)$$

where $h(t)$ applies the fixed spectral coloration (expressed as $H(w)$ in the frequency domain), $A(t)$ is the slowly-varying time envelope, and $n(t)$ is a white noise process. The analysis problem is to estimate the spectrum of $H(w)$ and the magnitude of $A(t)$.

This constraint that the noise energy envelope be 'separable' into time and frequency envelopes might seem like a rather severe restriction on the kind of sounds that can be represented. For example, a noise signal whose average spectral content changes smoothly and slowly – such as the classic 'moaning' of the wind – would not be amenable to such a decomposition. However, the intention of these generic elements is to define a set of components that encode the very minimum amount of perceptually-important information about sound events; if a sound *is* largely conformal to a separable model, it would seem that the independent marginal frequency and time envelopes would comprise a salient description. Also, the generic sound elements are the very lowest level of representation, and it is intended that more complex percepts, which may ultimately be experienced as a single event, might be composed of a sequence or combination of the basic elements. Thus the rising 'whoosh' caused by a gust of wind could possibly be approximated by several spectrally-static noise clouds in succession. While this would seem like a clumsy analysis, it should be remembered that the tremendous data reduction achieved by storing only the marginal spectra rather than attempting to sample the entire two-dimensional surface might make the clumsy representation still a more efficient encoding compared to a more flexible general-purpose noise envelope. However, a frequency-modulation term would appear to be a valuable future extension to this element.

Analysis is performed in the domain of the downsampled time-frequency intensity envelope calculated by the front end, $X[n,k]$, where n is the downsampled time index (in 220.5 Hz steps) and k is the frequency channel index (proportional to the log of filter center frequency). The smoothness constraints on $A(t)$ and $H(w)$ are defined such that each can be regarded as

constant over a single time-frequency tile of $X[n,k]$, thus the intensity envelope of the model formulation (4.6) may be expressed as:

$$X_N[n,k] = H[k] \cdot A[n] \cdot N[n,k] \qquad (4.7)$$

where $H[k]$ and $A[n]$ are the discretized, piecewise-constant spectral and magnitude envelopes respectively, and $N[n,k]$ is the time-frequency intensity envelope of a white noise process.

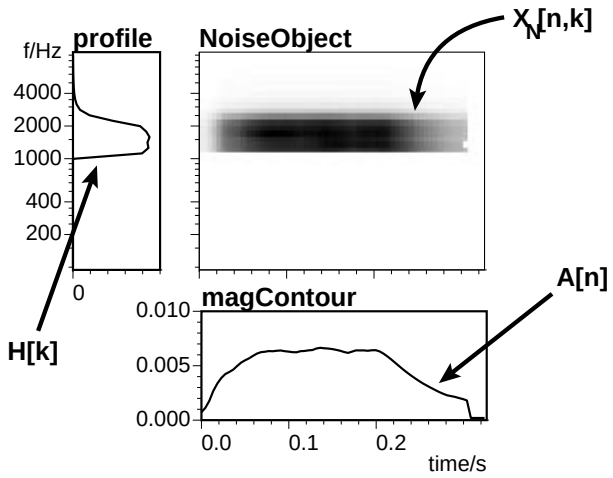

Figure 4.13: Diagram of a noise cloud, showing how the sustained, noisy time-frequency energy is modeled as the output of a random process with a smooth underlying expected power, separable into a spectral profile and a time contour.

Analysis

In order to recover $H[k]$ and $A[n]$ it is necessary to average over many samples so that the effect of the random variation in the samples of $N[n,k]$ is reduced. Consider estimating $H[k]$ from a single row of $X_N[n,k]$ where $A[n]$ is assumed known or constant and the noise has been adjusted to deliver unit power through the filterbank channel in question: the correct value of $H[k]$ would be the expected or average value of $X_N[n,k]$. If the successive time samples of $X_N[n,k]$ were independent, and their variance was known, $H[k]$ could be estimated with arbitrary accuracy by averaging enough values of $X_N[n,k]$, thereby reducing the variance in proportion to the number of points averaged. (A perceptually-inspired target of a noise estimator with a standard deviation of 0.5 dB was adopted). Although the successive values of an oversampled bandpass-filtered noise signal are not independent, the correct averaging time to achieve an estimate of a given accuracy may be derived by the following argument:

Considering a single frequency channel (k constant), $X_N[n,k]$ would be obtained by squaring and smoothing the output of a bandpass filter excited with white noise. The square of a bandpass signal has a lowpass spectrum, extending up to $\sqrt{2} \cdot B/2$ for the ideal case of a Gaussian bandpass spectrum of bandwidth B. Estimating the underlying energy of the noise process is effectively further low-pass filtering this squared signal so that the proportion of noise energy (variance) relative to the central d.c. peak (mean) is reduced to the desired level. The constant-Q nature of the bandpass filterbank means that for each frequency channel this noise spectrum will have essentially the same shape up to a dilation factor of the filter center

frequency, thus to reduce the variance-to-mean ratio in estimating $H[k]$ to a given level, the squared signal must be low-pass filtered to a bandwidth proportional to the original frequency channel's center frequency. This is done by applying a simple one-pole filter to the successive samples in each channel, where the time constant of this smoothing filter is longest for the lowest channels which require the narrowest bandpass filters to reduce the variance in the estimate. The actual size of the time constant to achieve the desired estimator variance was adjusted empirically for a prototype channel, then scaled for the other channels.

Estimating $A[n]$ involves equalizing the assumed variation of each relevant frequency channel (by dividing by their estimated mean level), summing them together, then smoothing this result to get estimated overall amplitude profile in time.

Although once established, the underlying noise process energy envelope is constrained to be smooth, the system permits noise clouds to start and stop abruptly when so required by the analysis. During the startup period of an abruptly-created noise cloud, the per-channel spectral profile estimates are simple unweighted averages of all the intensity samples seen so far, until this time exceeds the normal smoothing time constant, whereupon simple one-pole smoothing resumes.

Resynthesis

Resynthesis from parameterized noise clouds is simply derived from the underlying model. The desired noise intensity envelope $G[n,k]$ is formed by making the 'outer product' of the final estimates of $H[k]$ and $A[n]$:

$$G[n,k] = H[k] \cdot A[n] \tag{4.8}$$

This is then used as a time- and frequency-varying mask to gate a white noise signal broken into frequency bands by the analysis filterbank (cochlea model). It is necessary to precompensate for the considerable overlap of the analysis channels by 'deconvolving' each frame of intended spectral levels by the spreading functions i.e. to obtain an output spectrum of $G[k]$, the weights $W[k]$ applied to each channel of the filtered white noise excitation are obtained from:

$$G = \boldsymbol{B} \cdot W \tag{4.9}$$

where G and W are column vectors of the spectra $G[k]$ and $W[k]$, and $\boldsymbol{B}$ is a square matrix of real weights b_{ij} indicating the energy of the overlap between analysis channels i and j, i.e.:

$$b_{ij} = \int h_i(t) \cdot h_j(t)\,dt = \int H_i(w) \cdot H_j^*(w)\,dw \tag{4.10}$$

where $h_i(t)$ is the impulse response of the i^{th} bandpass filter in the filterbank, and $H_i(w)$ is its frequency response. Since the weights are being applied to energy in different frequency channels, they cannot be negative; hence the inversion of (4.9) to find W from G must be performed by nonnegative least-squares approximation.

After applying the weights $W[n,k]$, the filtered noise channels are added together to form a full-bandwidth noise signal whose time-frequency intensity envelope matches $G[n,k]$.

Prediction and error weights

The basic mode of the noise cloud element is to assume that it will continue without systematic change. Thus when required to make a prediction of the future, in the absence of higher-level guidance, the noise element simply repeats its current estimate of the average level in each frequency channel. The deviation bounds of this prediction are based on two factors, the parameter uncertainty and the intrinsic variability of a noisy signal. The uncertainty in the model parameters, which is approximated as a recursive estimate of the variance of those parameters. If the observed signal has been fluctuating considerably, preventing the model parameters from settling down to stable levels, the resulting changes to the model parameters at each time step will result in a significant variance for the history of the parameters when considered as a time-series. By contrast, if the observed sound is well matched by the assumptions of the model, the parameters should rapidly stabilize close to their optimal values, and the recursive estimate of the variance for the history of each parameter will become very small, indicating high confidence in that parameterization. (The uncertainty of the single magnitude contour, $A[n]$, is combined with that of each channel's profile level, $H[k]$, to give a net uncertainty in each channel).

This basic parameter uncertainty, expressed as a standard deviation, constitutes the error weight vector for this object, upon which basis the error between prediction and observation is apportioned between overlapping elements. However, the deviation bounds of the prediction are greater than this, because even if the model is a perfect fit to the data, the fact that it is noisy energy that is being modeled means that there will be an inevitable spread between the expected value of the intensity in each channel and its observed value at a given time. The range of this spread depends on the smoothing applied to the envelope of each channel relative to the bandwidth of the noise energy in that channel: The smoothing of the front-end intensity envelope extraction is the same for all channels, but the higher frequency channels pass a much broader band of noise. Consequently, the theoretical variance-to-mean ratio of the time-frequency envelope for static noise processes actually decreases in the higher channels, where the fixed smoothing filter has removed a larger proportion of the noise energy superimposed on the average signal level. Thus, the predictions of noise elements have a variance component, added in proportion to their level, with the constant of proportionality based on the empirical measurement of the variance in the envelope of noise analyzed by the system front-end for the particular frequency channel.

4.3.2 Transient (click) elements

Like the noise clouds, the second kind of generic sound element is intended to handle a class of sound that has been largely ignored in previous auditory organization models, namely short-duration isolated energy bursts, typically experienced as short clicks, cracks or bangs. The basic idea is that, like the noise clouds, such events are essentially characterized by a single spectrum and a separate description of their time variation. However, in the case of transients, the variation of intensity with time is more constrained: The perceptual characterization of a transient event is assumed to be an instantaneous onset of energy with some intensity profile followed by a rapid, exponential decay of that energy. Exponential decays arise from processes whose rate of energy loss is in proportion to their intensity (i.e. $d(\text{intensity})/dt = -k \cdot \text{intensity}$, the simple first-order differential equation solved by an

exponential) and occur very frequently in nature. There is some evidence for the special processing of sounds that conform to this general pattern, exemplified by the dramatically different perception of a sound consisting of a series of exponentially decaying tone-bursts and the same sound played backwards [Patt94]. Time asymmetry arising from the unavoidable constraint of causality in the lowest levels of auditory processing must certainly account for some of this difference, but the categorically different perception of sounds with rapid onsets and gradual decays compared to sounds with gradual onsets and rapid decays [Breg90] supports the idea that the environmentally commonplace pattern of a sudden energy burst smoothly dying away may involve a specialized perceptual representation. Note also that acoustic reflections will often adorn a sharp energy transient with a tail of decaying 'reverberation' energy; such tails are very often exponential in nature , and moreover listeners seem to be peculiarly able to isolate and discount such additions when they are not of interest [Beran92].

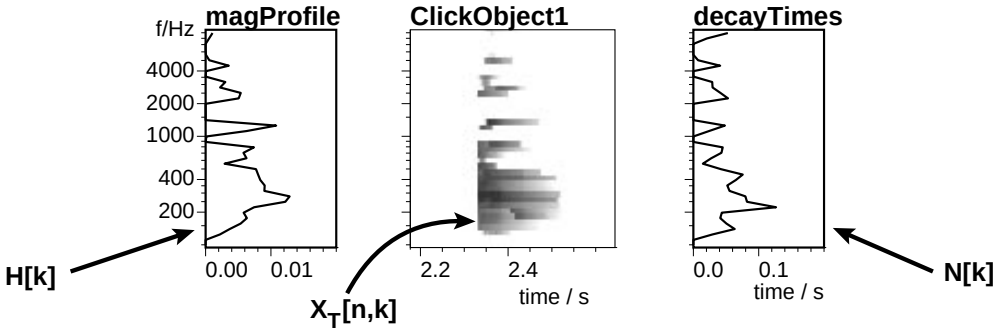

Figure 4.14: The structure of a click element: A region of decaying energy in time-frequency is modeled as an initial spectrum with a separate decay time for each channel.

Signal model

In the system, transient elements are characterized by three parameters:

- their onset time
- their initial spectrum
- their exponential decay rate within each of the peripheral frequency bands

Thus, in the same sampled time-frequency intensity envelope domain used to model the noise clouds in (4.7), the intensity envelopes of transient elements are modeled as:

$$X_T[n,k] = H[k] \cdot \exp((n-n_0)/N[k]) \qquad (4.11)$$

where $H[k]$ is the transient's spectral profile, n_0 is its onset time-step, and $N[k]$ is the per-channel decay time constant. Again, as with the noise clouds, analyzing a fragment of sound into such an element is a matter of extracting n_0, $H[k]$ and $N[k]$ from a region of $X[n,k]$; resynthesizing a transient element is a question of generating a sound whose time-frequency energy profile matches equation (4.11).

Analysis

The process of selecting a region of intensity envelope for modeling as a transient element is part of the prediction-driven analysis engine which is described later. However, once a section of energy has been targeted as a transient element, the analysis is rather simple: the envelope in each frequency channel is tracked until it reaches a local maximum; this is then the spectral intensity for this transient in this channel. Once the peak has been reached, the decay rate is formed as an average of the ratios of successive intensity samples in that channel, weighted by their absolute intensity, so that the initial few samples, most intense and therefore presumably most immune from noise corruption, dominate the exponential decay curve that is fit to them. The onset time is taken as the average of the times at which the maximum was recorded in the different channels, weighted towards the high frequency end of the spectrum, since the much broader band-pass filters in this region have the sharpest time resolution, meaning that their onset time estimates are more precise.

When considering short-duration transients in the context of a time-frequency analysis, it is important to pay full attention to the implications of the time-frequency uncertainty tradeoff. The filters underlying the time-frequency intensity envelope, and the Fourier theory underlying those filters, mean that the samples of time-frequency intensity cannot take on arbitrary, independent values, but are compelled to have a certain correlation with their neighbors. There are several possible interpretations of this condition; one is to consider the output of a filterbank as samples of an 'ambiguity function' – a kind of time-frequency distribution capable of arbitrarily sharp time and frequency boundaries, but with other undesirable characteristics – that has been smoothed by a 'blurring' function that removes the unwanted (nonsocial) artifacts, but causes a certain redundancy between adjacent time and frequency samples [Riley87]. The precise shape of the blurring function can be chosen arbitrarily, however, its total volume must exceed some minimum, imposing a joint lower limit on combined resolution in time and frequency. Equivalently, the redundancy between adjacent samples may be seen to arise from the overlaps in their support; as discussed above, the bandpass filters have quite considerable spectral overlap, so energy exciting one will most likely excite its neighbors to some degree. Similarly, the spacing between time samples may be significantly smaller than the time support of a given filter's impulse response, so successive samples will share a portion of exciting energy.

In the higher frequency channels, the smoothing and subsampling of the time-frequency intensity envelope occurs at a time-scale much larger than the intrinsic limit on timing detail imposed by these broad bandpass filters. Thus, other than the smoothing involved in constructing the envelope, successive time samples are independent, and a very brief transient signal will result in an equivalently brief transient in the intensity envelope. At the lower frequencies, however, the narrow bandwidth of the peripheral filters (only a few tens of Hertz for the lowest filter centered at 100 Hz) means that most rapid variation that can occur in the envelope of that channel is significantly slower than the 220 Hz sampling of the envelope is capable of representing. Thus even the briefest click would result in an envelope with a measurably slow decay in the lower channels, which is one reason why the decay time is estimated separately for each frequency channel: otherwise, artifacts of the analysis filterbank might obscure the underlying characteristics of the signal.

A second, related reason to measure decay rate separately in each channel is that many real-world transients do exhibit slower decay in the low frequencies, even when analyzed by a fixed-bandwidth filterbank that has no intrinsic bias towards sluggishness in low frequency. The reverberant impulse responses of rooms, while not the most obvious 'transient' in themselves, are a good example, showing much more rapid decays at high frequencies than at low [Beran92]. This can be attributed to various aspects of the real-world, such as processes that lose a certain proportion of their energy in each cycle (meaning that energy loss is slower when the cycle-rate – i.e. frequency – is lower, an equivalent definition of constant-Q), absorption constants that increase steadily with frequencies (meaning that lower frequencies will persist through a greater number of reflections), the ability of lower-frequency sound waves to diffuse around obstacles of a certain sound (meaning high frequencies are more likely to be shadowed or otherwise removed by objects in space). It is interesting to speculate that perhaps the variable-bandwidth structure of frequency analysis in ears evolved not simply for reasons of efficiency of construction or compromise among detection characteristics, but because we live, in some sense, in a 'constant-Q' environment.

Resynthesis

The time-domain constraints of the analysis filterbank must be taken into account in the resynthesis of transient elements also. Our representation of the element gives a well-defined shape in the intensity-envelope domain as described in equation (4.11); we could simply use the same amplitude-modulation of narrowband noise channels that was used to generate noise matching a given intensity envelope for the noise cloud elements. However, because the timescales of transient elements are likely to be short compared to the filterbank constraints, it may be important to factor out the time-blurring implicit in each analysis filter to generate a sound whose re-analysis will have the same decay time as represented in the model. Thus, although the resynthesis technique is essentially the same as for the noise clouds, some attempt is made to precompensate the decay envelopes applied to the noise channels for the blurring that would occur on re-analysis.

Predictions

The predictions of the basic transient element are simply the extension of the current spectral profile and per-channel decay rates by another time step. Both the deviation bounds and the element error weights are based on recursive estimates of the parameter uncertainty, which rarely have the opportunity to stabilize. Unlike noise elements, no additional margin is added to the deviation bounds to account for intrinsic signal unpredictability.

4.3.3 Weft (wideband periodic) elements

The third element to be described is in many situations the most important, but it is presented last as it is also the most complex. The concept of the *weft* element (whose name comes from the Anglo-Saxon word for the sets of parallel threads in a woven fabric through which the warp is threaded) arose from a comparison [EllisR95] between the discrete narrowband elements of the analyses used by [Cooke91] and [Ellis94], and the more promising, but unsegmented, correlogram representation of [SlanL92] [DudaLS90]. The simple idea is to use combined information from short-time autocorrelations within each of several frequency channels to detect the presence of periodic signal in a sound, then to recover as much as possible of the information

about the spectrum of that sound by looking into the autocorrelations of each channel. Rather than considering the periodic signal as a collection of resolved Fourier components, as would result from a narrow, fixed bandwidth analysis, the assumption is that each of the peripheral frequency channels is so broad as to encompass several harmonics. The constructive and destructive interference between these harmonics will cause the energy contribution of the wide-band periodic excitation in that channel to be reflected in an autocorrelation peak at the fundamental period. Of course, in situations where the frequency channel is sufficiently narrow to pick out a single Fourier harmonic, this harmonic would also give a peak at the fundamental period, since its own period will be an integral division of the fundamental, and autocorrelation produces aliases at period multiples.

Signal model

The signal model which weft analysis seeks to extract is a periodic wideband excitation with a somewhat smoothly-varying period, shaped by a time-frequency envelope also subject to smoothness constraints. Thus the weft signal,

$$x_W(t) = \left[e(\tau) * h_W(\tau; t) \right](t) \tag{4.12}$$

where $h_W(\tau;t)$ is the time-varying filter, and $e(t)$ is the pseudoperiodic excitation. The model is thus very reminiscent of the tradition source-filter model of speech, where the pseudoperiodic glottal-pulse-train is shaped by the resonant cavities of the vocal tract [RabinS78]. The pseudoperiodic excitation is defined as:

$$e(t) = \sum_i \delta(t - t_i) \tag{4.13}$$

– a train of impulses $\delta(t)$ at times t_i given by:

$$t_i = \arg_t \left\{ \int_0^t \frac{2\pi}{p(\tau)} \, d\tau = 2\pi \cdot i \right\} \tag{4.14}$$

where $p(t)$ is the time-varying instantaneous period of the excitation. (Since information about $p(t)$ is only obtained at each impulse occurring in $e(t)$, it is further constrained to be smooth on that time scale).

The time-varying spectral modification is defined as a scalar weighting surface sampled at the particular discrete-time and -frequency matrix used in the system, as for the noise elements. Thus in the sampled time-frequency domain of (4.7) and (4.11), the weft element is:

$$X_W[n,k] = H_W[n,k] \cdot E[n,k] \tag{4.15}$$

where $E[n,k]$ is the transform of $e(t)$, and the time-varying spectral envelope $H_W[n,k]$ is constrained to be sufficiently smooth to be approximated by a constant value within each time-frequency cell. The task of weft analysis is to recover the smooth spectral envelope, $H_W[n,k]$, and the instantaneous period track, $p(t)$, for a particular wide-band periodic element detected as present in a mixture.

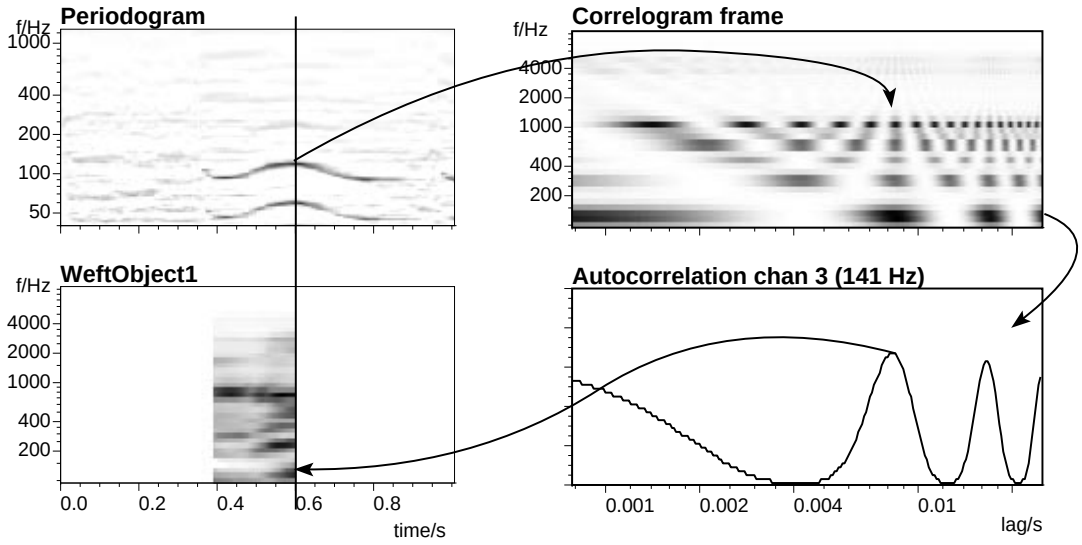

Figure 4.15: Diagram of the analysis of weft elements. Ridges in the periodogram (top-left panel) indicate the presence of periodic modulation across many frequency channels visible as a vertical structure in the correlogram slice (top right panel) and give the period, $p(t)$ for the weft element; an individual unnormalized autocorrelation function for a given frequency channel and time slice (lower right pane) is inspected for a local maximum at the appropriate lag; any such peaks found are used as the basis for an estimate for one cell of the weft energy envelope, $H_W[n,k]$ (lower left pane).

Analysis

The approach, which was touched upon in the description of the correlogram and periodogram front-end features, is to first extract a likely period track from the periodogram, then to estimate its spectral envelope by looking at the appropriate lag-time in the short-time autocorrelations of each frequency channel from the correlogram. The periodogram was devised for the single goal of exposing the presence of wide-band periodic excitations in a signal, so the starting point for extracting a weft is a peak that exceeds a simple threshold in a slice of the periodogram (a summary autocorrelation). Since the periodogram is already normalized, there is no need to adapt the threshold; it was set through experiment to give a reasonable discrimination between desired and spurious periodic components.

When the analysis has decided that there is indeed a wideband periodic modulation present at the current time-frame with a certain, known period (one point in $p(t)$), its amplitude in each frequency channel must be obtained to 'fill in' the column of $H_W[n,k]$ for the current time-step. This is done by looking at the short-time autocorrelation for each channel at that time step; if a given channel's autocorrelation contains a local maximum very close to the fundamental period for that time slice, and assuming the periodic excitation dominates the energy in that channel, the square-root of the unnormalized autocorrelogram sample is taken as the time-frequency envelope value for that cell. (If there are interfering simultaneous signals, the value is compensated as discussed below). The rationale here is that if the frequency channel contained only a purely periodic signal, the autocorrelogram sample would be the smoothed output of the product of that signal multiplied by itself delayed by exactly one period. Since for a periodic

signal the delayed waveform would exactly match the undelayed signal, the autocorrelation value would be just the time-smoothed square of the (rectified) filterbank output, whose square-root is the overall time-frequency intensity envelope used by the system as the basis for energy explanation. Although factors such as time-variation in the period and amplitude of the signal make the true situation more complicated than this, the assumption is that a value based on these simplifications will give a reasonable indication of the energy in the frequency channel due to signal components modulated at that period.

In channels that do not have a clear local maximum, the weft's envelope is recorded as a 'don't know' to permit later interpolation if it turns out that periodic energy in that channel has been masked by another signal. The actual autocorrelation value is recorded as an 'upper limit' on the envelope in that channel, since if the periodic component had been more intense than would correspond to this level, its autocorrelation bump would presumably not have been masked.

When there is periodic energy from more than one source in a given channel, there will be interaction in their autocorrelation signatures. If the signals have similar periodicity, there comes a point when nothing can be done locally to separate their contributions – the best we can do is add some later processing to notice the collision, and perhaps fix it up by extrapolation from better-distinguished time frames. A comparable situation occurs if one signal has a period which is close to an integer multiple of the other; although their 'fundamental' autocorrelation maxima may be well separated, the fundamental bump of the longer period may be distorted by the higher-order aliases of the shorted period's bump, since a correlation between a signal and itself delayed by a lag of D will usually be accompanied with a very similar correlation at a delay of $2D$. In fact, if we assume that the signal is static over that timescale, the bump at $2D$ will be a precise replica of the local maximum at the fundamental delay D. This opens the way to a kind of compensation: When the weft analysis detects a signal at some short period D (corresponding to a high fundamental frequency), it can generate a 'mask' for the predicted higher-order aliases of the autocorrelation maxima in each channel by copying the fundamental maximum at shifts corresponding to multiples of the fundamental period. This mask predicts features expected to arise from this short-period signal at larger periods in the correlogram slice, and can be subtracted from the full autocorrelation functions in an attempt to remove the effect of higher-order correlations on the traces of other periodic signals in a channel. (Since a given bin in our autocorrelogram scheme has an average level significantly greater than zero, this subtraction is performed in a domain where the normalized average level for each bin has been subtracted). Thus the search for local maxima in the summary autocorrelation proceeds upwards in period (i.e. starting from the highest fundamental frequency considered), and each recognized periodicity causes its higher-order aliases to be factored out of subsequent processing.

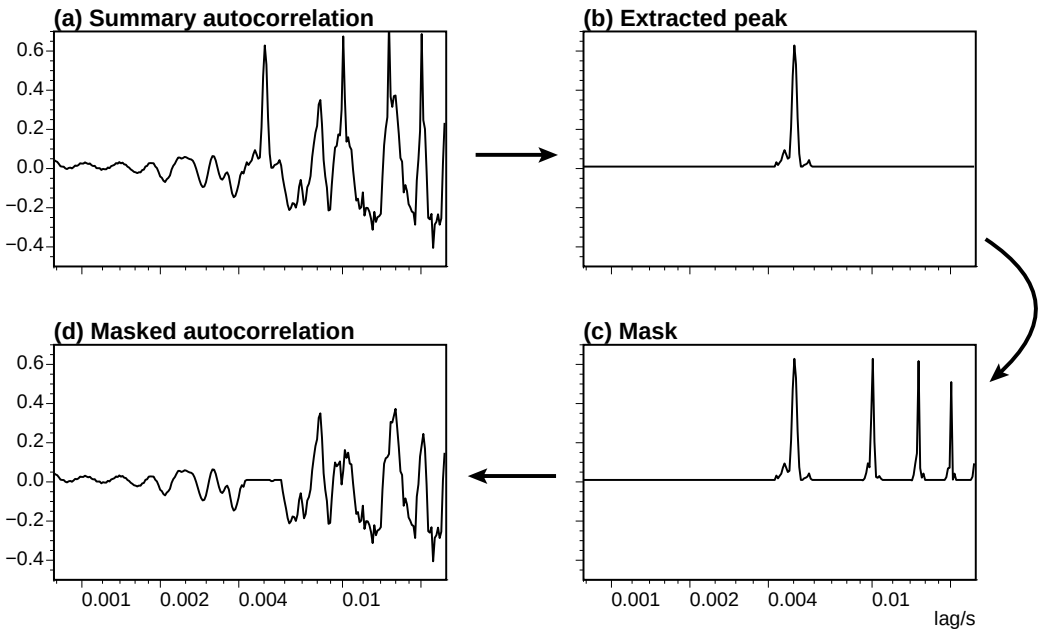

Figure 4.16: Illustration of the removal of higher-order aliases for an identified 'bump' in a summary autocorrelation. Panel (a) shows the initial summary autocorrelation (on a log-time axis) with a clear periodicity bump at a lag of 5 ms. Panel (b) shows the isolated bump extracted by the analysis. In panel (c), the bump has been shifted by multiples of its average lag time (i.e. to 10 ms, 15 ms etc.); the log-time sampling makes each alias appear successively condensed. Panel (d) shows the residual summary autocorrelation after the aliases have been removed, from which other periodicity features (e.g. the bump at 8 ms) may be extracted.

The effect of non-periodic interference (i.e. added noise) on a channel's autocorrelation cannot be handled so neatly. Channel noise tends to boost the autocorrelation at every lag in a nonlinear fashion, leading to overestimates of periodic energy. Some compensation for this effect is accomplished by comparing the ratio of peak and average correlogram values to that expected for noiseless periodic signals, and reducing the signal level estimate if an abnormally large average level indicates added noise. The basis of this calculation is presented in more detail in appendix A.

Resynthesis

The weft analysis procedure can extract multiple period tracks and associated time-frequency envelopes from the periodogram (summary autocorrelation) and the 3-D correlogram volume. Resynthesis of a weft from its period and envelope characterization is straightforward and analogous to resynthesis of the noise elements; first, a pseudoperiodic impulse-train excitation is generated from the period track $p(t)$, then it is broken up into frequency channels, each of which is modulated by smooth, scalar envelope derived from $H_W[n,k]$, where the derivation incorporates compensation for the differing amounts of energy contributed by the excitation to each frequency channel (in proportion to the channel's bandwidth) and factoring-out the overlap between adjacent channels by non-negative least-squares approximation.

Predictions

As the only elements capable of accounting for features in the periodogram, the predictions of weft elements are made both in the intensity-envelope domain (as for the other elements) and for the next summary autocorrelation (periodogram slice). The basic prediction principle is to assume that the element will continue unchanged into the next time slice, in the absence of any higher-level model which might, for instance, be tracking pitch slope. The deviation bounds are set as a small proportion of the intensity. However, an interesting problem arises compared to the noise and transient elements: Whereas overlapping noise elements have an essentially unconstrained problem when apportioning observed signal energy between them, a weft element is less accommodating: The process of extracting a spectral envelope from the correlogram has produced a separate estimate of the actual energy associated with this modulation period, regardless of the other elements present. While the correlogram extraction mechanism has a certain amount of error associated with its results (reflected in its associated variance values and carried through to the prediction), it is less able to participate in the apportioning of error between predicted and observed signal energy, since it is separately constrained to follow the levels derived from the correlogram. Thus the intensity variations absorbed by weft elements, and consequently the error weights they offer, are rather smaller than for other kinds of elements.

4.4 The reconciliation engine

The third component of the prediction-driven architecture is the engine that manages the creation, modification and termination of the internal model of the sound-scene to match the external signal. At the lowest level, this is a question of maintaining the set of elemental sound representation objects to be consistent with the cues supplied by the front-end analysis. It also encompasses the updating of the higher-level back-end hierarchy of explanations for these sound elements. Before going into the detail of how this is achieved, we examine the blackboard system through which this operation is accomplished.

4.4.1 The blackboard system

The engine implementation is based on a blackboard system, as discussed in general terms in chapter 2. The actual code is derived from the ICP (IPUS C++ Platform), whose features are described in [WinN95] [LessNK95], and many of whose key features are based on the RESUN architecture of [CarvL91], although this specific lineage is not crucial to the flavor of the system. Using a blackboard architecture is largely a programming convenience as it provides a structured foundation for a system that consists of numerous competing hierarchies of hypotheses at different levels of abstraction. Blackboard systems address the issue of flow-of-control in such situations by using an explicit rating to choose and develop the hypotheses that seem the most promising. This is particularly valuable in systems where analysis can proceed in various directions, some of which may be fruitless wastes of effort.

This blackboard implementation is characterized by four aspects: the levels of the hypothesis hierarchy, the basis for rating the quality of each hypothesis, the idealizations of solution state in the problem-solving model, and the actions defined for developing hypotheses.

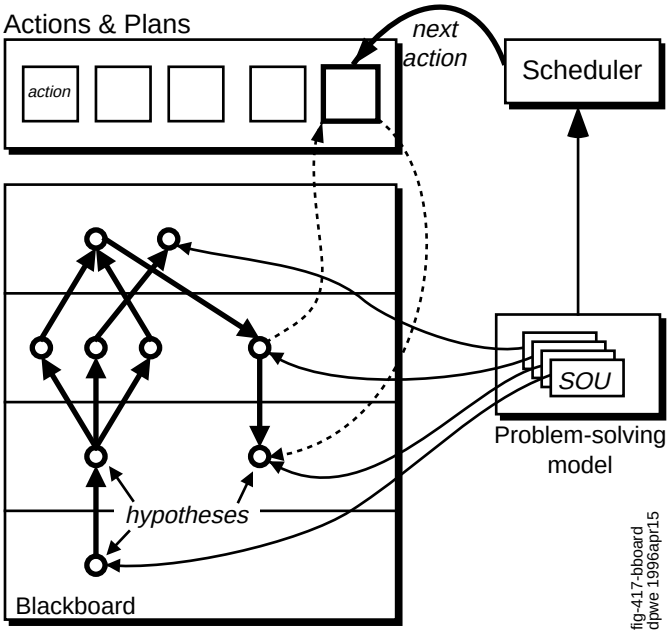

Figure 4.17: Diagram showing the major components of a typical blackboard system.

The hypothesis hierarchy

The sound organization system defines four initial levels on the hypothesis blackboard, although the intention of the system is that further levels, representing more abstract analyses, may be added. The blackboard architecture assumes a hierarchic relationship between the levels of the blackboard, with hypotheses at lower levels providing 'support' for hypotheses at higher levels, which correspondingly provide an 'explanation' of their support. The four levels defined for this implementation, starting from the lowest (least abstract), are:

- **Surface**. This level contains only one 'hypothesis', standing for the raw input data obtained from the front end. This is the grounding support for all the other hypotheses in the system. The actual data here is both the time-frequency intensity surface and the periodogram; the goal of the analysis is to provide satisfactory 'explanations' of both these data sets. Although the data in the surface is in fact read into memory in a single operation, it is analyzed incrementally via a steadily-advancing time horizon, as if the new data were arriving in 'real time'.
- **Subscene**. The immediate explanation of the surface is provided by the subscene hypotheses, each of which represents the lowest level of a putative complete explanation for some patch of the surface (i.e. a limited range of time). When a prediction is made by querying each element of a candidate explanation, it is within the subscene hypothesis that the sum of the predictions is recorded, to be compared against actual data from the supporting surface hypothesis in a later time-step. Each set of objects proposed to explain a portion of the sound thus has a single subscene hypothesis at its base.
- **Element**. Directly above the subscenes in the explanation hierarchy are the sound element hypotheses which correspond to single elements of the types specified in the previous section. A given subscene may contain any

number of elements depending on the complexity of the explanation it is constructing. It is by gathering the individual predictions of each of its explaining elements that the subscene constructs its aggregate predictions for the entire surface.

- **Source**: As explained in the discussion of the internal sound model, elements are merely the most concrete aspect of descriptions of sound-producing entities in the external world that may be arbitrarily abstract. In this implementation, there is just one level of explanation for the elements; in theory, there could be multiple levels of explanation, each adding higher-order constraints on the predicted behaviors of the elements at the base of its support. Note that there is a 'fan-out' between subscenes and elements, with a single subscene potentially supporting many elements. Above the elements, however, there is a 'fan-in', with each element permitted to support only one source-level hypothesis, and potentially several elements being controlled by each source.

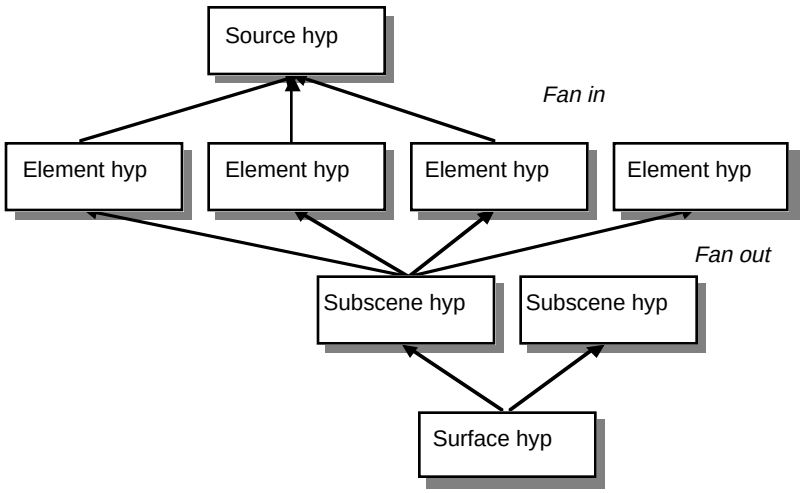

Figure 4.18: The various hypothesis classes present on the blackboard, and an illustration of the fan-in and fan-out on each side of the element hypotheses.

The rating of hypotheses

In order to make the choice of where to apply processing effort, a blackboard system must have a method by which to compare the 'promise' of the different partial solutions represented by its hypothesis. A common problem, not avoided in sound organization system, is that of devising a rating system that can be usefully compared between entities of different types. The goal is to produce some kind of estimate of the likelihood that a particular partial solution will lead eventually to the correct (or one of the acceptable) analyses of the input.

In the sound organization system, the common principle adopted was the information-theoretic concept Minimum Description Length (MDL), where a particular explanation is 'scored' according to the number of binary digits it requires to account for the observed data [QuinR89] [Riss89]. MDL is used in situations where a choice needs to be made between radically different models which do not permit any direct comparison except at this most basic level. The difficulty or art in applying MDL lies in devising the specific representations whose size is to be measured, and in calculating the description-length penalties to be associated with the 'model specification' part of a description (which specifies the particular competing model to be

used) as distinct from the 'model parameters' part (the actual coefficients for that model to account for the observations). MDL has a close relationship to probability theory, particularly Bayesian analysis, where the model specification lengths equate to Bayesian priors associated with each model. Overall, the description length score of a particular explanation may correspond to a likelihood that it is correct according to Shannon's equation:

$$b = -\log_2(p) \tag{4.16}$$

which relates the number of bits, b, in a code for a symbol (or outcome) to the probability, p, of that symbol occurring.

In the sound organization system, the rating of a hypothesis is intended to reflect the average number of bits that hypothesis requires to account for each time-frequency sample it covers. In this way, hypotheses that have been able to account for a sound with the simplest set of parameters will be preferred over more complex explanations. However, there also needs to be some accounting for the accuracy with which the model has predicted the signal: Recall that a putative explanation provides a probabilistic prediction of each future input sample, i.e. an expected level and a tolerable deviation bound around that level. If the model is explaining the signal as a random noise process, this prediction is necessarily uncertain since the precise fluctuation of the noise energy cannot be predicted. However, if a competing hypothesis makes a much narrower prediction of the signal level (one with much smaller deviation bounds) which turns out to be correct, the competing hypothesis should possibly be rated as superior to the high-variance noise prediction, even though it might have employed more bits of parameterization to achieve that prediction.

Such a comparison is afforded by calculating the MDL scores as the number of bits required to specify the input signal to a fixed level of accuracy. By this mechanism, a 'loose' (high-variance) prediction is penalized, since in order to encode the signal to the specified accuracy it would need additional information to specify, to the required level of accuracy, where within its wide prediction range the observed signal fell. A model providing a very narrow prediction range might need no additional information if it already predicts the level to within the required accuracy, and the amount of extra information required increases with the variance range. Thus broad predictions are penalized compared to more specific predictions, and the advantage of the MDL approach is demonstrated by the ease with which the concept of the rating may be extended beyond the characterization of an explanation itself to encompass its accuracy too.

A modification to the basic MDL rating was required to handle the case of very small signals. Since deviation bounds are normally roughly proportional to prediction levels, a model predicting a very small signal in a given channel will also have a very small deviation bound for that channel. Such predictions run the risk of a larger but modest signal appearing in the channel whose level is none-the-less very remote from the prediction in terms of the tiny deviation bound. The MDL rating for the object incorporates the negative log-likelihood of this level under the probabilistic model of the prediction (according to (4.16), and the huge score for this situation would effectively doom the object to immediate guaranteed abandonment. This is rarely the correct treatment, so to avoid this undesirable situation, an upper ceiling was imposed on the number of bits calculated to be required for any single sample. This ceiling was set at the word-length required to store the sample in a linear binary representation at some fixed resolution: The

rationale for such a ceiling is that this linear encoding achieves the greatest accuracy required of any time-frequency sample, and if encoding under a model would require a longer description, the linear code would be used in preference.

Note, of course, that the description of the signal whose bit-length is being considered is never actually constructed. It is merely that the ratings of element and subscene hypotheses are calculated as the length of an imaginary code that would be required precisely to represent the observed signal under the models they comprise, since the length of this theoretical encoding forms a suitable basis for comparison.

The problem-solving model

The RESUN architecture of [CarvL91] is based on the view that signal analysis in blackboard systems is essentially a process of eliminating uncertainty in the explanation until a sufficiently likely answer is obtained. To this end, it includes a summary of the state of analysis called the *problem-solving model* or PSM. The PSM keeps track of the ways in which analysis may be continued by tracking all the *sources of uncertainty* (SOUs) attached to current blackboard hypotheses. Every hypothesis type has an associated range of SOUs that might apply to it; as each fundamental analysis action is applied to a hypothesis, it will add and remove SOUs to indicate how the state of analysis for that hypothesis has changed. For example, when the initial 'surface hypothesis' is created, containing the raw data from the front-end processing, it is tagged with a 'no-explanation' SOU, indicating that the most obvious immediate reason that the system has not yet produced an answer is that no explanations have been suggested to account for the bottom-level data. Although their specific interpretation depends on the type of hypothesis to which they are attached, the basic SOU types defined include:

- 'no-explanation' and 'no-support', indicating that the hypothesis hierarchy needs to be extended, either in the bottom-up, abstraction direction, or the top-down, prediction, direction respectively.
- 'partial-extension', indicating a hypothesis capable of being extended into future time-steps.
- 'inadequate-explanation' and 'inconsistent-explanation', which are used to flag predictions providing too little or too much energy relative to the actual observations.

When an action is applied to a particular hypothesis, it is also responsible for altering the SOUs attached to the hypothesis to indicated the new analysis state. Typically, this will involve removing the SOU which was actually responsible for invoking that action. Hypothesis development can be viewed as a sequence of states, each labeled by an SOU, with particular actions comprising the transitions between those states.

A given hypothesis may have several attached SOUs. SOUs have an intrinsic priority to determine which should be resolved first, all other things being equal. Thus, an 'inconsistent-explanation' SOU, indicating that a particular prediction has failed to match up to the observed data, is addressed before a 'partial-extension' SOU, whose resolution advances that hypothesis to the next time step.

The analysis actions

Having represented the state of the solutions offered by each hypothesis, the system has to decide what action to take to make progress in the analysis. This must be based on a combination of choosing the most promising hypothesis, and knowing what may be done to improve it. Ratings provide the first part of this by indicating which of the existing hypotheses has been most successful thus far, and might therefore be a promising bet for further development. Through the specific 'sources of uncertainty', the problem-solving model facilitates the second part of this choice, by explicitly labeling the various deficiencies in a particular hypothesis that the system may choose to resolve. The RESUN framework actually incorporates a subgoal-based planner, which achieves more abstract goals (such as 'solve-problem') by breaking them down into sequences of smaller goals (such as 'select-SOU' then 'resolve-SOU'); each possible action that the control engine may invoke has a particular subgoal that it can solve, along with a precondition that tests whether it is applicable. Thus the action to extend a subscene hypothesis into the next timestep meets the 'resolve-SOU' goal, but only in the situation where the selected SOU is of the 'partial-extension' class attached to a subscene hypothesis. All this information is attached to the action itself, thereby giving the control engine sufficient knowledge to proceed through the analysis.

Some examples of the actions involved in the analysis are:

- 'solve-no-explanation-sou-element-hyp': A sound element hypothesis has been created, but as yet there is no more abstract source hypothesis that it is supporting (indicated by the 'no-explanation' SOU attached to it on creation). This action will attempt to find or create a higher-level (source) explanation for this element, and will remove the SOU if it succeeds.
- 'solve-partial-extension-sou-subscene-hyp': A subscene (a collection of elements explaining a particular patch of the observation surface) has been reconciled with only a limited ranged of the observed time -steps; this action will widen its purview to include the next time step, thereby triggering the creation of predictions by all the elements and sources it supports. This, in turn, will cause an 'incomplete-support' SOU to be attached to the subscene to indicate that it contains predictions that have yet to be reconciled to the underlying support. Only after this reconciliation (and the modifications to the explanatory hierarchy it may entail) will the refreshed 'partial-extension' SOU, also added by this action, be invoked to advance the analysis still further in time.
- 'solve-inadequate-explanation-sou-subscene-hyp': A prediction for some part of the observed signal surface failed to account for all the energy that was actually observed; this rule will attempt to create a new element to add to the explanation of this subscene that can account for the excess energy. Creating a new element will firstly look for current source hypotheses that might predict possible additional elements of a specific form, then fall back on explanation as default noise or transient elements. Very often, there will be several choices for the additional explanation which each seem quite plausible given the information to date. In this case, the subscene hypothesis will 'fork' into several alternative versions, each pursuing a different possible explanation. Each candidate explanation will then have its own chance to account for the actual observations; the less appropriate solutions will rapidly fall into neglect as their ratings fail to compete with their more successful siblings.

4.4.2 Basic operation

The interaction of the various components of the blackboard system is best understood by walking through a typical sequence of analysis operations. Although this suggests a central analysis 'loop', one of the defining features of blackboard implementations is that nowhere is such a loop directly encoded; rather, it emerges from the set of possible actions available to the system, the states which they connect, and the particular characteristics of the data under analysis. One motivation for using a blackboard architecture is that the structure of the problem frustrates attempts to specify an analysis loop in advance; a blackboard system can proceed with an analysis regardless.

Prediction and reconciliation

Since the analysis is incremental, new time steps are considered only when all the data to date has been satisfactorily explained. The first action to occur concerning a new timestep is the construction of a prediction for the indispensable domains – the energy envelope and the periodogram. A subscene hypothesis that is sufficiently highly rated compared to its neighbors will have its 'partial extension' source-of-uncertainty selected for resolution, and the resulting action will gather and combine the predictions of the attached sound-element hypotheses. The elements themselves will generate predictions based on their existing context (i.e. their history) and on any source-hypothesis they support.

Predictions are combined within each class of element, and then into single overall predictions of the energy and periodogram slices for the new time slice. The predictions are all probabilistic, consisting of expected values along with positive and negative deviation bounds. Each class of element has a separate rule for combining predictions with those of its siblings; for noise elements, this is simply a matter of adding the expected levels in the power (amplitude-squared) domain, and adding the variances, taken as the squares of the deviation bounds. A weft rule may have more complex logic to account for more slowly-varying phase interactions between resolved harmonics, although this was not included in the current implementation. Combining the predictions of the different element classes into the single overall prediction is again done in the power domain i.e. assuming incoherent phase interaction between the different elements.

A complete prediction is flagged by a 'partial-support' SOU on the subscene, which triggers the process of reconciliation of that prediction to the observed input. For the time-frequency intensity envelope, the difference in each frequency band between the predicted and actual amplitudes is scaled by the deviation bounds of the appropriate polarity; the norm of the resulting vector of scaled differences is equivalent to the Mahalanobis distance between the prediction and the actual [Ther92]. The positive and negative components of this vector are treated separately: If the norm of the positive vector elements exceeds a threshold, then there is energy in the input that exceeds the prediction by more than the prediction's own deviation bounds, and an 'inadequate-explanation' SOU is generated, motivating the addition of a new element to the explanation. If the norm of the negative vector elements becomes too large, the observed energy in those channels is significantly smaller than predicted, and an 'inconsistent-explanation' SOU is created which will trigger a search for an element that may have terminated.

Similarly in the periodogram domain, the actual summary autocorrelation for this time step (i.e. the periodogram slice) is searched for significant peaks that were not predicted by the element ensemble. These are flagged as

explanation inadequacies. It is not, however, necessary to flag predicted periodogram features that are not observed, since each weft element will consider its own termination during parameter update if an autocorrelation peak at the desired period cannot be seen. Also, if a periodic source has terminated, it is likely to be reflected by a change in the energy envelope too, and resolution of 'inconsistent-explanation' SOUs for the energy envelope may also terminate wefts.

Hypothesis creation

Handling the 'inadequate explanation' SOU that results from excess input energy is one of the trickiest parts of the system. It is also an important window of opportunity through which higher-level abstractions, including future system extensions, can influence the analysis. As a well-defined hypothesis-SOU conjunction, any newly-added actions that address this situation will automatically be invoked as appropriate. The current action operates as follows: Firstly, the periodogram surface is consulted to see if there is cause to generate a new weft element. An increase in signal level may result from the addition of periodic or aperiodic energy, and the periodic explanation takes priority (since periodogram features cannot otherwise be explained). If a new summary autocorrelation bump is found, a weft element is created for that period and added to the subscene's explanation.

If the excess signal is not accounted for by periodic energy, it is passed on to be explained as a noise or click element. These elements are rather similar at onset, and the choice between them may not be possible until the signal has been observed for several more time-steps. At a later time their different handling of the post-onset behavior of energy (decaying in click elements, sustained in noise elements) will distinguish the preferable choice. In creating these new elements, the following considerations also apply:

- The onset map (described earlier in this chapter) is consulted to check for evidence of a rapid onset. From this map, an average local rate of energy increase is calculated over all the channels indicated as contributing to the observed energy excess. If this value (already normalized for absolute signal level) exceeds a threshold, the energy is judged to have arisen from a genuine onset event, and new elements are created. If the threshold is not reached, however, the observation of excess energy is considered spurious and ignored. This can happen for several reasons: If the deviation bounds attached to the prediction become extremely small (because predictions have gained confidence through past accuracy), a modest deviation in the observed energy will become greatly magnified when normalized by the deviation bounds. This can lead to an 'inadequate-explanation' event which should in fact be ignored, with the effect that the worsening prediction performance of the existing elements will cause wider variation bounds for future predictions. Another source of spuriously inadequate explanations is the enforced decay of a click element, which, if attempting to fit a non-decaying energy source, will provide successively diminished predictions that leave a growing shortfall compared to the actual input energy. In this case, refusing to add a new element to absorb this shortfall is an important step in letting the rating of the associated subscene hypothesis deteriorate, reflecting the inappropriate match of the explanation to the data, and leading to the eventual abandonment of the hypothesis.
- Abstract explanations of previous observations may result in the anticipation that a particular element or conjunction will occur at some

point in the future. At present, this is limited to the situation of a terminated noise element generating an anticipation that a noise element with the same spectral profile may reappear; in general, a more sophisticated abstraction of the signal could generated more complex anticipations when a particular pattern or sequence is recognized. These anticipated elements are checked for compatibility with shortfalls in the prediction, and may be realized if they provide a promising explanation. By starting with parameters that are already attuned to a particular form of energy, an anticipated element will out-compete a completely new element on the basis of ratings, provided of course that the new energy does in fact conform to the anticipation. This mechanism may be thought of as analogous to the concept of 'priming', usually considered in relation to vision, where a particular context can increase a subject's sensitivity to a particular pattern, predisposing them to interpret new information in a specific way.

- There are several special-case rules to help with the correct operation of the example implementation. Bootstrapping at the start of a sound is one such situation: The system is obliged to add a noise element to an inadequate scene that contains no other elements, ensuring that there is a continuous 'background-noise' explanation within the subscene. The large amount of onset energy at the very beginning of a sound example would otherwise favor a click element, but such a hypothesis, decaying away again to nothing, is not worth pursuing.
- It was advantageous to inhibit the creation of multiple click elements in rapid succession by imposing a minimum time interval between the creation of such elements. Without this constraint, a single transient might generate several elements as its initial spectrum developed in the first few time steps to confound the element that had initially been created. By refusing to create additional elements, the first element is forced to conform to the true form of the transient. This situation might be more properly addressed not in the element creation logic but with more sophisticated predictions of click elements at onset.

Typically, several alternative elements can account for the excess energy, and, lacking evidence to choose between them at this stage, the subscene hypothesis branches to generate several alternative continuations. These variations on the original hypothesis will then be able to compete for development further into the future on the basis of their success at predicting the observations, as reflected in their ratings. Branched hypotheses inherit copies of all the active explanation elements in the subscene, which must be distinct entities since they will develop independently henceforth. This branching of hypothesis versions during the resolution of inadequate-explanation events is the only mechanism that generates the multiple alternative hypotheses on the blackboard, at least in the current implementation.

Once a new hypothesis has been created including the additional element intended to account for the excess energy, the prediction and reconciliation for the current time-step are repeated. This should always result in a successful prediction of the current observations, since the new element has been added specifically to achieve that end. The new hypothesis will then be ready to advance forward and predict the next time step.

Hypothesis termination

The previous subsection described the handling of the situation in which the energy observed in certain frequency channels exceeds the predictions for those channels, suggesting that new elements should be added to the explanation. In the converse situation, where the energy observed is in fact significantly smaller than predicted, the system will consider eliminating elements to reconcile the 'inconsistent-explanation' SOU that has been generated. In principle, this is a simple matter: For each current element, predictions are made based on a subscene that has had that element removed, and these reduced predictions are compared to the actual observations. If the exclusion of a particular element leads to a better match with the observations (according to the Mahalanobis, or variance-weighted, distance metric), then that element is terminated and the subscene continues without it.

In practice, there are a couple of modifications to this basic principle. Firstly, this comparison is made in the time-frequency intensity envelope domain. This could result in the elimination of a weft element whose *energy* contribution was unnecessary, while at the same time providing a crucial explanation of a feature in the *periodogram* domain. Thus each tonal (weft) element is only considered for termination if its periodogram support is relatively weak; otherwise, eliminating the element in one time-step would simply result in its regeneration on the next, as the newly-exposed peak in the summary autocorrelation demanded explanation. Secondly, it was found to be necessary to provide 'protection' for certain hypotheses: For instance, a relatively low-level background noise explanation might be eliminated to provide a short-term benefit when a more intense foreground element predicted a little too much energy. Later on, when the foreground element decayed, the previous termination of the background would leave the hypothesis in a difficult situation. This situation could be handled through competition between alternative hypotheses, generating branches on element termination with and without the excess element, and relying on the ratings to eventually favor non-termination when it was the correct choice. Pragmatic considerations made it preferable to 'nudge' the system towards the correct choice at the decision point; this was achieved by prohibiting the termination of elements that had existed for more than a certain amount of time.

The handling of inconsistent-explanation SOUs is only one way in which an element may be terminated. In addition, each element has some logic for self-termination based on its parameter updates. Click and noise elements monitor their levels relative to the peak level in each channel and will drop out of the explanation if they have become small relative to their past and to the current total signal. Weft elements are governed by the presence of periodogram features that they explain; if their fundamental-period pulse disappears for more than a couple of time frames, they too will terminate.

Apportioning prediction error

Even when a particular combination of elements provides a satisfactory prediction of the observed signal energy, there will still be some residual error between the actual and the prediction. This to should be passed to the elements to allow them to trim their parameters for improve future predictions. Simply informing each element of the overall error will not work: Consider a channel in which the prediction was a little smaller than the actual observation; each element, notified of prediction error, might increase

its level in that channel to compensate. But if there were many elements, the combined effect of the individual compensations might over-predict the next time step, leading in all likelihood to oscillation. The ideal solution would be to inform only the 'right' element – the one which really ought to have made a larger prediction – and none of the other. Of course, there is no obvious basis upon which to decide which element is 'right'.

Instead, the total error is divided up so that each element involved is passed only a fraction of the overall error, with the fractions summing to the whole error. This division is made on the basis of the error weights through which every element rates its own confidence in the prediction it is making. The error weight is often the same as the deviation bound (i.e. the 'standard deviation' of the probabilistic prediction), although for noise elements a very low error weight may be attached to a prediction whose variance is still quite large, reflecting the intrinsic variability of noise signals. The error weight itself follows both the magnitude of the prediction (so that an element contributing a greater proportion of the prediction will similarly received a greater proportion of the prediction error, all other things being equal) and the parameter uncertainty for that element (so that a newly-created click element that has had little opportunity to gain confidence in its model parameters will 'soak up' the major portion of the prediction error).

Error allocation involves summing the error weights for all the currently-active elements, allocating prediction error to each (in the magnitude-squared domain) on the basis of their proportion of the total error weight, constructing a new 'target magnitude' for that element, then backing off any error allocation that would take the target magnitude for a certain element in a particular channel below zero. In this way, an element making a low-magnitude prediction with a large error weight will not inadvertently absorb a large negative error which it cannot accommodate. Since the total prediction is essentially the sum of the element predictions, and since the observed level will always be nonnegative, the prediction error can never become so negative as to be incapable of absorption by the elements.

Each element is passed its apportioned error, which it then uses to update its own parameters, possibly propagating through to a higher-level source explanation. The modified element is then up-to-date with the current time-step and ready to make a prediction for the next instant.

Note that the question of error-apportioning would not arise in a data-driven system, where the explanation would have been abstracted directly from the observation. By contrast, the prediction-driven approach can arrive, through context, at a situation where more than one element is overlapped to predict a single observed value – commonly the correct interpretation in a real sound scene, and thus one that the a successful system must be able to construct, but one that carries with it the problem of apportioning deviations from prediction between basically inseparable components.

Calculation of ratings

Each element hypothesis is given a rating score notionally related to the number of bits required to represent the actual signal in the context of the model, as described above. When a subscene consists of a single element predicting the observations, the rating may be calculated unambiguously. When several elements are acting in combination, it is necessary once again to divide up the difference between actual and prediction between each of the elements, constructing the 'target magnitudes' of the preceding subsection, which would have predicted the actual observations exactly. Each element's

rating within the mixture is then the appropriate description-length score for the respective target magnitude in the context of the prediction and deviation bounds. Note that if the error weights are just the deviation bounds (as they are in the click and weft elements), then the deviation-based component of the rating will be balanced across all the elements (although the individual ratings will vary owing to the aspects of the rating calculation that reward more specific predictions at a given level of accuracy).

The overall rating of a subscene hypothesis is simply the sum of the ratings of the elements it supports, in keeping with the idea that the rating measures the number of bits required for a partial description of the sound, and the entire sound is described by combining all the parts. A lower rating therefore corresponds to a superior hypothesis, indicating that it described a sound by using fewer elements, or with fewer bits in addition to its elements because the elements made an accurate prediction.

Startup, termination and competitive behavior

After re-rating the hypotheses, the system is ready to continue on to the next time step, first by making a prediction, and then by reconciling it to the actual data; the central 'loop' is complete. Some special cases warranting further discussion are startup and termination of entire explanations, and also the typical behavior of competition between hypotheses. Starting the analysis from scratch is in some senses an artificial problem affecting only the model: A 'real' sound-organization system will have some kind of current context at all times. As mentioned above, the very beginning of a sound is treated as a special case by the element-creation code – the newly-created subscene will have no elements with which to account for the observations. In this case, a noise cloud element is constructed to act as the underlying noise floor, and that element is allowed to stabilize for a short while before creating additional explanatory elements. In the examples used to assess the system, stabilization required only on the order of a hundred milliseconds.

Analysis termination is in general a complex issue in blackboard systems, since the *first* complete explanation generated might not be the *best* – some of the other partially-complete hypotheses might, when fully developed, turn out to provide better overall explanations of the data. Empirically, the current implementation can take many tens of time-steps (i.e. hundreds of milliseconds of input sound) to settle on its favorite hypothesis after an event such as the appearance of a new sound object. This did not, however, lead to any disruptive ambiguity, at least in the examples employed. Although it was not necessary in the current implementation, the system could insure against overlooking a temporarily-disadvantaged explanation by developing to completion *all* partial explanations of comparable rating after the front-runner has exhausted the input data. Then the best solution could be picked based on the final ratings.

The general pattern of competition between different hypotheses tends to operate as follows. When a potential solution diverges (i.e. at an onset event whose correct explanation is ambiguous), the two forks are initially very similar, having almost equal ratings, and will be developed in an interleaved fashion for a while. Within a few tens of time-steps, one of the branches will often turn out to be a better fit, and will begin to be awarded more of the analysis effort on the basis of its superior rating. Normally, within a hundred time-steps or so, the 'losing' branch will have been completely abandoned, since, as the leading hypothesis advances into the future, its rating will improve with the cost of adding a new element being averaged over a larger

number of time steps. Occasionally, however, a hypothesis will obtain a short-term advantage (e.g. by eliminating a low-level background element) which will turn out to be a major disadvantage later on (when the masking foreground element decays); in these cases, the rating of the front-runner will start to grow as the inappropriate fit becomes apparent, until the rating of the languishing alternative is once again worthy of consideration. In the examples explored, it has never been necessary to go back more than one second (i.e. approximately 200 time-steps) to find an alternative hypothesis, and indeed for reasons of practical efficiency hypotheses that fall more than a couple of hundred time-steps behind the front-runner are 'retired' – removed from future consideration.

4.4.3 Differences from a traditional blackboard system

Although I have placed considerable emphasis on the value of using a blackboard architecture as the basis for a hypothesis-oriented, prediction-driven analysis scheme, there are several aspects in which the system as described is not a particularly compelling instance of a blackboard implementation. One curious oddity is the way in which the subscene hypotheses act as a point of convergence for multiple higher-level elements and source explanations, reversing the more common pattern in abstraction hierarchies where a single higher level element explains and combines many lower level elements. This may reflect the slightly unconventional goals of the analysis system: Rather than trying to answer a specific question, such as "can I hear a car approaching?", the sound organization system is trying to construct a *complete explanation* for all the sound it receives. In a traditional sensor-interpretation blackboard system – for instance, the helicopter signal tracker of [CarvL92a] – the ultimate answer is a single, abstract causal explanation ("reconnaissance mission") for the interesting signal data. By contrast, the desired output of the sound organization systems is not a single causal account, but several high-level abstractions to reflect the several, causally-unrelated external sound sources that have overlapped to create the total sound scene. From this perspective, the low-level subscene hypothesis takes on great significance as the point at which the otherwise independent source hypotheses are combined and finally reconciled with the observed sound. It is at the level of the subscene hypothesis that the final complete 'answer' is chosen, which then indicates the set of preferred source-level explanations by association.

Probably the greatest advantage argued by proponents of blackboard systems is their ability to switch between analysis procedures at widely differing levels of abstraction on an opportunistic basis. However, until the abstraction hierarchy of the current system is deepened through the addition of higher-level explanations, there is a rather shallow range of abstractions with which the system is dealing, and the 'choice' of which level at which to pursue analysis is normally trivial. None the less, a blackboard architecture has turned out to be a very convenient approach to handling the competition between hypotheses and the distinct levels of explanation involved in this problem. Moreover, it is an approach very well suited to the future extensions that any approach to computational auditory scene analysis must expect to require.

4.5 Higher-level abstractions

The final major component of the prediction-driven architecture is the collection of higher-level abstractions which provide explanations and interpretations for the basic sound elements created at the lower levels. Much of the theoretical attraction of the architecture relies on this part of the system and its ability to disentangle confounded or masked cues through context-dependent interpolation. Regrettably, the most glaring shortfall of the current implementation is that this part of the architecture is hardly developed.

In the description of the blackboard's hypothesis hierarchy, the 'source' level was introduced as containing hypotheses that were supported by one or more sound elements that together formed a larger, interrelated pattern. These source hypotheses embody a more sophisticated level of knowledge about the sound-world, incorporating both general rules about sound patterns from distinct sources, and more specific recognized conjunctions that permit detailed predictions. A system that is able to recognize these larger-scale and thus more highly constrained patterns will also be able to make more specific predictions and thereby accomplish a more efficient analysis.

Presently, the only object in this class is the rather humble model of repeating noise bursts. When a noise-cloud element is created, a noise-burst-source hypothesis is also created as its explanation in the source level of the blackboard. Although in general source hypotheses will modify the predictions and parameterizations of their supporting elements, this particular source allows the noise element to develop without additional constraints. However, when the noise element is terminated, the source hypothesis remains active and creates a specific anticipation that a noise element with the same spectral profile will reappear at some time in the future. This anticipation is embodied on the blackboard as a noise-element hypothesis that supports the noise-burst-source hypothesis, but without any support of its own and without a specific time range. This anticipated element is ignored until an inadequate-explanation situation arises, meaning that the underlying subscene has failed to predict enough energy and needs to consider adding a new element. At this stage, any 'anticipated' elements are considered in preference to a completely unparameterized new element, as described in the section on hypothesis creation above. Assuming that the new energy does indeed arise from a recurrence of something very similar to the previous noise burst, the anticipated element will immediately provide a good fit to the observations, and will form part of a successful and highly-rated hypothesis. When the new noise element terminates, the source hypothesis will generate yet another anticipation, and so on.

The noise-burst-source hypothesis is a very simple example of one kind of source-level hypotheses dealing with extended temporal patterns. Other possible functions of hypotheses at the source level include:

- Implementation of grouping heuristics. Gestalt principles of similarity, continuity and common fate that are not captured directly in the sound elements may be detected at the level of source hypotheses. Sound elements whose onsets are aligned, or whose onsets coincide with the offsets of other elements, are probably related, as are elements with similar characteristics which occur close together in time. Manual simulation of this kind of grouping was implicit in the combination of several weft hypotheses to form single 'car horn' source-events in the city-sound analysis example, described in the next chapter.

- Short-term acquisition and anticipation of patterns. I believe that a very important aspect of context-sensitivity in sound organization is the way in which a particular sound pattern will have an acute influence on the auditory system to interpret any closely-following similar sounds as repetitions of that pattern. The noise-burst-source hypothesis was one example of this behavior, but a more general implementation would encompass conjunctions of *multiple* sound-elements, perhaps grouped by Gestalt principles, which would then provide a short-term bias and specific anticipations if parts of the pattern were seen to repeat.
- Recognition of previously-learned patterns. The idea of short-term predisposition extends over longer time scales to a system that has a repertoire of known sound patterns, again possibly biased by context, which can be brought to bear on the current sound scene. These can suggest highly specific interpretations, and correspondingly constrained predictions, for specific low-level sound elements.

Ultimately, hypotheses at the source level can themselves support still more abstract explanations, conceptually extending all the way to very abstract explanations such as “the mail being delivered” or “Jim talking about his roommate”. The benefit of these additional layers of explanation lies in the extra constraints they bring to bear upon the supporting sound elements; if the constraints are successful in matching the sound observations, then the predictions have been both specific and accurate, and the system has made a much more sophisticated interpretation of the data. In practice, problems might arise from an exponentially-growing number of alternative interpretations as the number of abstraction levels increases, but this is precisely the kind of computational challenge that a blackboard system handles by dynamically limiting its search to the promising areas of hypothesis space.

Higher level abstractions are a very important part of the overall vision of the prediction-driven architecture and represent the crucial dimension for the development of the sound-organization system described in this chapter and other such implementations. However, even despite the extreme limitations of the current implementation, it has been able to accomplish some interesting and valuable analyses of otherwise problematic, dense sound scenes, using only the unguided development of the low-level sound elements themselves. These results are described in the next chapter, along with a comparison to the performance of human subjects listening to the same sound-scenes, and subjective quality ratings by those listeners of the system’s output.

Chapter 5 Results and assessment

The previous two chapters presented an approach to automatic sound organization, and described an implementation of a computational auditory scene analysis system based on this approach. In this chapter, we will examine the performance of this system from two viewpoints. In the first half, the behavior of the system in responses to various illustrative sound examples will be considered. This will show qualitatively how the basic components of the system – such as the different generic sound elements and the hypothesis-blackboard system – operate in practice.

The second half of this chapter makes a comparison between the behavior of the implementation and the prototype of which it is supposed to be a model – the auditory system. This is accomplished through listening tests with human subjects. I will described the test that was devised, emphasizing the qualities that make it particularly appropriate for working with an intrinsically-defined phenomena such as auditory scene analysis. The results of the experiment provide some kind of answer to the question of whether the model is truly modeling the human organization of complex sounds.

5.1 Example analyses

In this section the various features of the implementation will be illustrated by considering the results of the system when processing some simple sounds. Although the system was developed with a specific orientation towards dense sound mixtures of the kind used for the subjective tests described later, it is instructive to see how it handles less complex examples. Also, the finer detail discernible in less dense sound scenes makes them in some senses a more difficult domain, compared to rich ambient sounds where the majority of detail is possibly masked.

5.1.1 Bregman's alternating noise example

The first example is a completely synthetic stimulus, cited by Bregman to explain his concept of 'old-plus-new' auditory analysis [Breg95] – i.e. that the auditory system is inclined to view a significant change in sound quality as the *addition* of a new sound element, rather than the complete *replacement* of one source by another, whenever the acoustic evidence is consistent with this interpretation. (Clearly this principle has a sound ecological foundation, i.e. one source starting requires less co-ordination in the environment than a source that starts at the same instant as another finishes, constituting a simpler, and hence preferable, explanation).

The stimulus, which was discussed in chapter 3, is illustrated in the top panel of figure 5.1 by its cochlea filterbank spectrogram – i.e. the time-frequency intensity envelope calculated by the model's front end – as well as its periodogram, uninformative in an aperiodic sound of this kind. The sound consists of bursts of low-frequency noise (below 1 kHz) alternating with a broader band of noise (up to 2 kHz), with the spectra of both matching below 1 kHz. Bregman makes the point that although the signal could have been

constructed either by having two different noise bands that alternate in time, or by having a continuous sub-1 kHz noise to which bandpass noise of 1 to 2 kHz is periodically added, it is only the latter interpretation which is available to the human listener; the continuity of noise energy below 1 kHz between the two bursts guarantees that the auditory system will arrive at the 'simpler' interpretation that the episodes of broader noise energy result from an addition to a continuous band of low noise, rather than a complete substitution, conforming to the 'old-plus-new' principle.

This is also the natural result of analysis by the prediction-driven system. The model's outputs are illustrated in the lower panels of figure 5.1. The system models the sound as a continuous, smooth background noise from 0 to 1 kHz, to which additional noise bursts between 1 and 2 kHz are periodically added. This simple and rather obvious result illustrates numerous aspects of the implementation, which are now discussed.

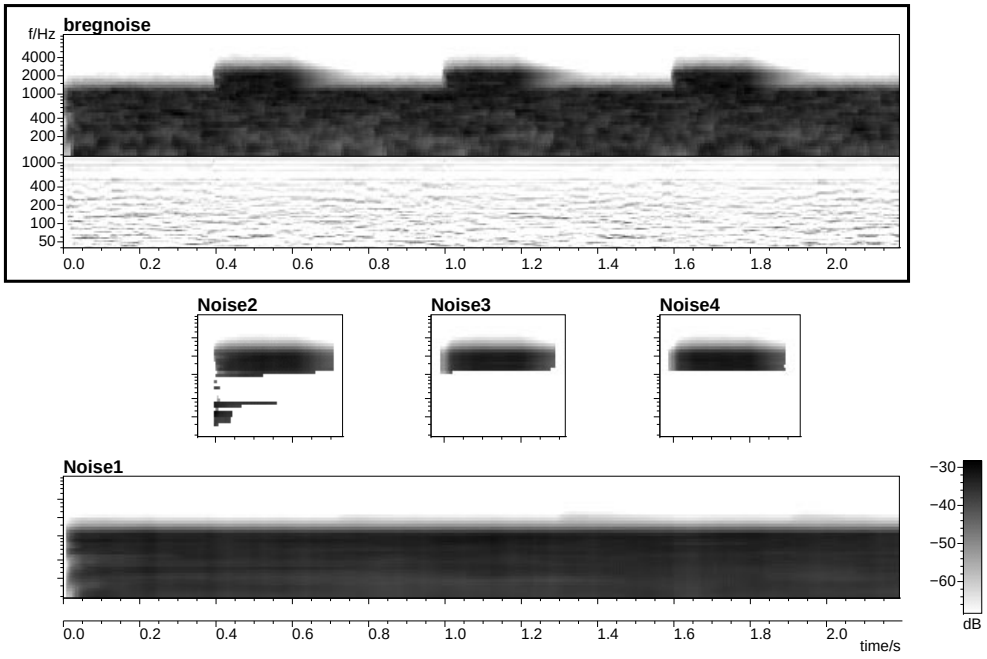

Figure 5.1: The Bregman alternating noise stimulus: The top pane shows the time-frequency intensity envelope and periodogram of the original sound. The three middle panes and wider bottom pane are the envelopes of the four noise clouds by which the system explained the example.

Modeling noise

This is a noise-based stimulus without any perceptible periodic component, and the elements with which the system describes it are all noise-clouds. (The periodogram feature and the weft elements will be discussed in the next example). The difficulty in modeling signals as noise is that the relationship between observed characteristics and model parameters is statistical rather than direct, and the parameters must be estimated over a time window. As explained in chapter 4, the noise cloud model is a steady underlying noise function whose spectral profile and total magnitude fluctuation are separately derived. The smoothing windows applied to the actual signal values in order to estimate the underlying expected level vary according to the effective bandwidth of the frequency channel; the broader, high-

frequency channels require less smoothing to achieve a given level of stability in the estimates than the more slowly-varying narrow low-frequency channels. Figure 5.2 shows a slice along time through the intensity envelope for a single frequency channel, showing the fluctuations of the envelope of the noise signal and the evolving estimate of the underlying level extracted by the system. The smoothing time is proportionally smaller at the very beginning of a new element to permit more rapid accommodation during onset. Note that the model's level at each time step is a product of the value for this frequency channel in the 'normalized' spectral profile with the scalar overall magnitude value for that time step; thus, the level can vary in response to across-spectral variations in intensity even after the normalized spectral profile estimates have become very stable.

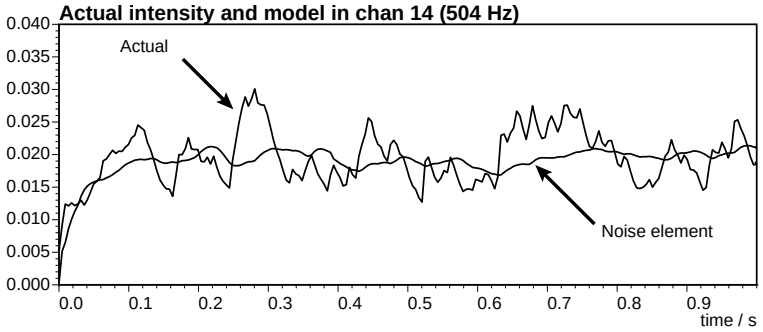

Figure 5.2: A slice through the intensity envelope for the 500 Hz channel showing the time variation of the noise signal envelope, and the corresponding estimate of the underlying noise expectation from the noise cloud element explaining the signal.

Handling overlapped noise elements

Although the different noise elements occupy largely separate spectral regions (0-1 kHz for the continuous band and 1-2 kHz for the additional bursts), there is a degree of overlap at the edge, raising the issue of the reconciliation of predictions based on more than one element to a single signal value. Although the long, background noise element has its energy concentrated below 1 kHz, there is always some energy in the higher channels (due to 'bleeding' in the peripheral filterbank if nothing else); when the low noise band is present alone, this element must account for the entire spectrum, so it must have a nonzero spectral profile in every bin. Figure 5.3 shows a slice across frequency through the intensity envelopes in the middle of the first broad noise burst at t = 0.54 s. The profile of the background element can be seen to decay towards the high frequency indicating the average signal intensity envelope during the low-band noise episodes.

The profile of the second noise band is obtained by calculating the energy needed in addition to the low band to bring the total prediction in line with the observations. The noise model has noted that there is no need for additional energy below 1 kHz, and thus the second element is exactly zero over most of these channels. Note in figure 5.1 that immediately after onset there is in fact some energy allocated to the lower frequency channels of the second noise element, since the random fluctuation in these channels cannot immediately be distinguished from a systematic increase in level that should correctly be associated with the new element. However, these spurious channels are soon deleted when it becomes apparent that, over time, the useful contribution to these channels by the new element is negligible.

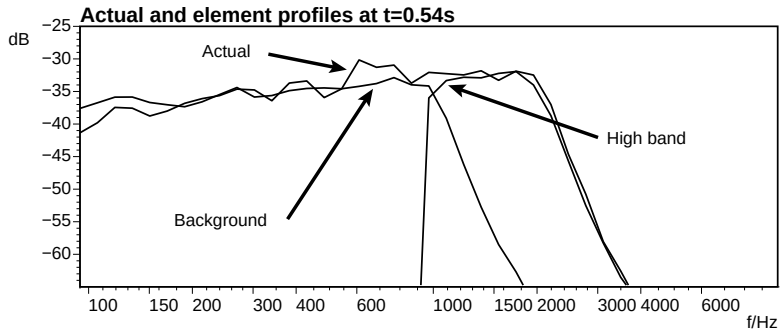

Figure 5.3: A vertical slice through the intensity envelopes of signal and model elements at t=0.54 s. The broad band of noise is explained as the overlap of the 'background noise' element, with its energy concentrated below 1 kHz, and an additional high band of noise between 1 and 2 kHz.

When developing this pair of overlapping noise elements, the system is faced with the problem of apportioning any discrepancy between prediction and actual amongst the different contributing elements. Because the background, low-band noise element has proved to be a stable and accurate model of the signal during the first few hundred milliseconds, its error weights have already become quite small by the time the broader band of noise begins. Consequently, the much larger error weights associated with the new element cause it to absorb the majority of the excess energy. Figure 5.4 shows the variation with time of the intensity envelopes for a single channel, showing how the second element comes to account for the excess energy. The error weight parameters in that channel for both elements are also illustrated, showing that the error weight of the original background noise has decayed to be very small by the time the second element is created.

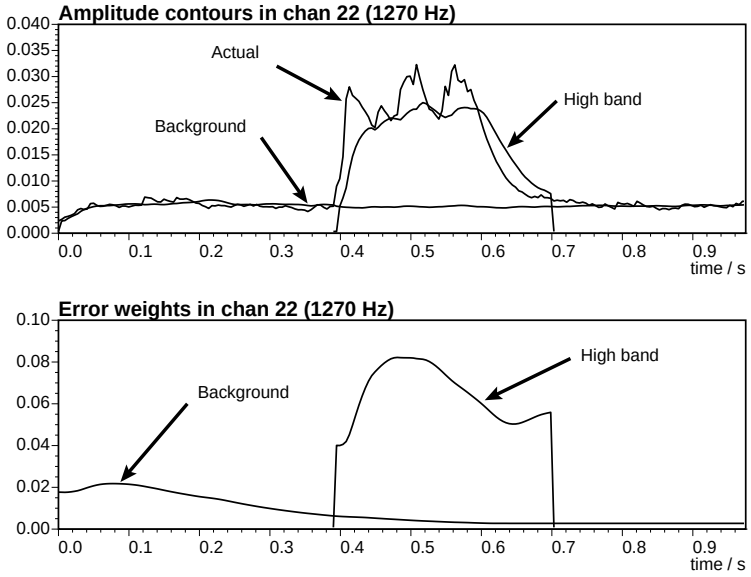

Figure 5.4: Time-intensity profiles for a single frequency channel showing the addition of the 'high-band' element to account for the sudden energy increase at t=0.4 s. The error weights decay after onset while the prediction succeeds.

Creation and termination of noise elements

The operation at the center of the prediction-driven approach is the reconciliation between the predictions based on the internal model and the observed signal features. As described in the last chapter, this comparison between predicted and actual signal characteristics can result in the system flagging, through 'source-of-uncertainty' objects, situations where extra elements are required to enable a subscene to account for an observation, or alternatively that the subscene needs to terminate some of its current components to allow the observed energy to encompass the whole prediction. These situations are detected through the norms of the normalized positive and negative difference vectors – that is, the vector difference between predicted and actual spectra at a given time slice, separated into positive (more energy in observation) and negative (more energy in prediction) components, then normalized by the deviation bounds for each channel. The norms of these two vectors are compared against a simple threshold, and when the positive deviation norm becomes too large, an 'inadequate-explanation-sou' is raised to set into motion the creation of a new element; a large negative deviation norm gives rise to an 'inconsistent-explanation-sou'.

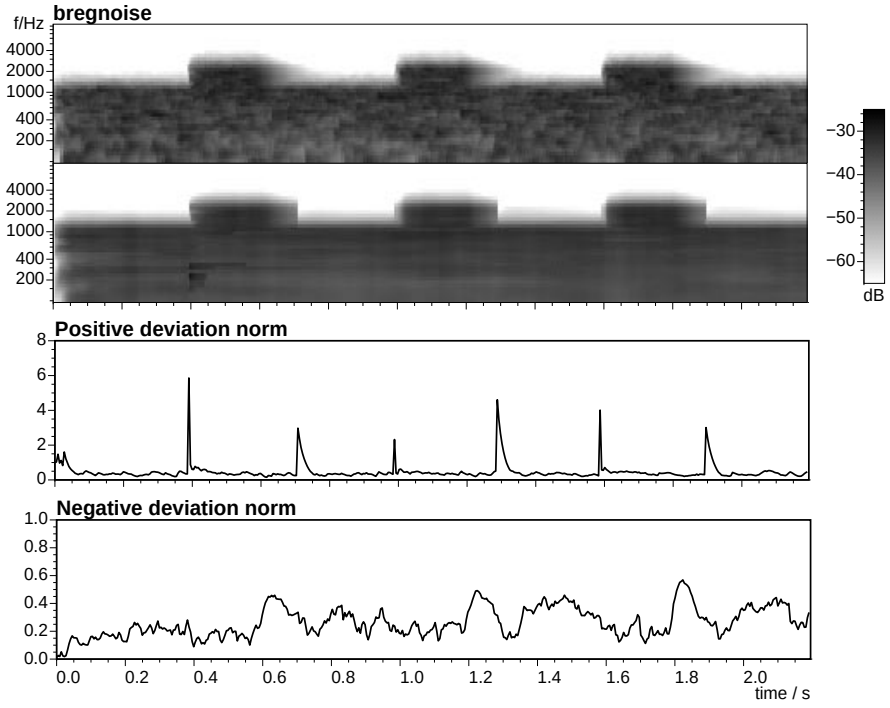

Figure 5.5: The positive and negative deviation norms (norms of the vector difference between predicted and observed spectral slices, normalized by deviation bounds and split into positive and negative components) for the 'winning' explanation of the Bregman noise example, whose intensity envelope is shown at the top of the figure for comparison. The second pane shows the combined envelopes of all the explanatory elements – the complete explanation.

The time variation of these two norms is shown along with the overall signal envelope in figure 5.5. Each subscene (comprising a set of elements that explain some portion of the input) has its own deviation norms; those displayed relate to the 'winning' explanation illustrated in figure 5.1. The

results are just as might be expected; after a startup period, both norms are rather small during the period of stable low-band noise, although there is a fair amount of variation between timesteps as the random fluctuation of the noise signal is more or less similar to the stabilized explanation. When the wider band of noise begins, the positive deviation grows very quickly; the signal is exhibiting proportionally massive increases in signal level in the upper channels which have rather small deviation bounds by virtue both of having been stable for many timesteps and being of low energy to begin with. As soon at the positive deviation norm exceeds the threshold, the engine creates a new noise element, and the norm immediately drops to a very small value – again, as a result of the dual effects of the extra energy provided by the new element eliminating the shortfall between prediction and observation, and because the very large deviation bounds of a newly-created element cause any residual prediction error to be magnified far less in normalization.

The increase in deviation bounds also effects a temporary damping of the negative deviation norm. However, within a few tens of timesteps, the prediction of the broader band of noise as the combination of the two elements has more or less stabilized, and the norms resemble their behavior before the onset. Then, at t = 0.6 s, the upper band of noise disappears from the input. Note that the smoothing applied in the calculation of the intensity envelope means that there is a somewhat softened decay on the trailing edge of the noise bursts; however, the predictions are now consistently in excess of the actual signal, and the negative deviation norm begins to grow. For a while, the system cannot find a way to resolve the 'inconsistent-explanation' condition, since the high-band noise element is still required to account for the decay tail of the burst (i.e. removing the element makes the prediction still worse), but eventually the signal decays to a small enough level that the prediction is better without the second element; the large negative deviation norm means that the system is still on the lookout for an element to remove, and thus the high-band noise element is terminated, and the negative deviation drops down again (at t=0.7 s), mirrored by a temporary jump up in the positive deviation norm, which is ignored for lack of evidence of an onset.

The creation and termination of the elements for the remaining noise bursts are similarly marked by peaks in the positive and negative deviation norms.

Competition between explanations

Thus far, the discussion of the system's analysis of the Bregman noise example has only considered the single hypothesis that was returned as the overall 'best' explanation. However, the system was not able to construct this explanation without any false moves; rather, at various decision points it had to construct several alternative hypotheses and develop them all for a little while until it became clear which ones held out the best promise for a complete solution.

As described in chapter 4, the current implementation only forks hypotheses into competing versions at the times when new elements are being created; if there is ambiguity concerning which type of element to add (such as frequently occurs between the initially similar noise and click elements), the system will create both alternatives, and leave it to the rating process in subsequent timesteps to sift out the better solution. In this short example, there are four instants at which new elements are created (including the initial startup); if each creation led to two alternatives (with the added

element being either a click or a noise), there would be sixteen different subscene hypotheses that could be developed through to the end of the sound.

In fact, the number is different from that for several reasons. The 'evolutionary tree' of subscene hypothesis versions is illustrated in figure 5.6. The element creation at startup leads only to a single hypothesis, since a special-case prevents a sound explanation consisting of nothing but a click element (which would decay away to leave absolutely nothing). However, when the first broad-band noise enters at time-step 86, this hypothesis splits into alternate versions that explain the new energy with either a click or a noise. The sustained level of the noise burst is poorly matched by the click element, which will always try to decay, with the result that its associated hypothesis is only developed for a few timesteps before the rating of the alternative explanation, consisting of two noise elements, is sufficiently attractive that the hypothesis including the click element is permanently abandoned.

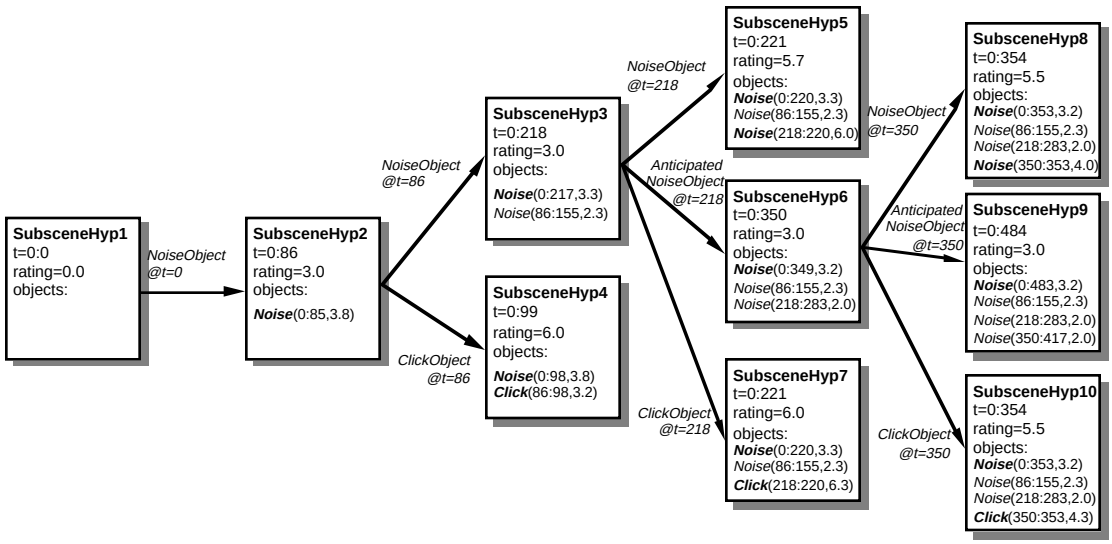

Figure 5.6: Diagram of the evolution of the various alternative explanations for the Bregman noise example. Most of the unsuccessful explanations were identified as unpromising rather soon after their creation, and have thus not been developed for very many timesteps. The time range (in 220.5 Hz timesteps) as well as the terminal rating (expressed in average bits of description per time-frequency cell) are noted alongside each hypothesis. The 'winning' explanation is SubsceneHyp9, which covers the full 484 time steps of the original.

Anticipatory noise bursts

At the next onset, the main hypothesis actually splits into three different versions – explaining the additional energy as either a new click element, a new noise element, or an *anticipated* noise element. In the last section of chapter 4, I explained that although the implementation of abstractions above the level of sound elements had been largely neglected, the one exception was the 'noise-burst-source', which inclined the system to recognize a recurrence of a noise burst that had existed in the past by posting an 'anticipated' noise element with the same characteristics but initially without any support from the subscene hypothesis. When handling an 'inadequate-explanation' condition, the system will inspect any anticipated elements, and, assuming there is a passable similarity between the prediction shortfall and the characteristics of the anticipated element, the anticipation will be

realized as part of a new explanation. Rather than trying to make a difficult choice between the preset anticipated element and a new, unmodified element, the system just creates them both, relying again on the competitive mechanism to sort out the best hypothesis.

Of the three branches arising from the onset of the second broad noise burst at time-step 218, the one based on the anticipated noise element rapidly wins out, since the new noise burst follows the profile of the previous burst very closely, and thus the anticipation is a close match. While the unanticipated noise element would certainly fit itself to the noise burst in due course, its initial unnecessary fumblings in lower frequency channels result in a considerable short-term rating disadvantage, so it is only the hypothesis including the realized anticipation that is developed through to the third noise burst at timestep 350. Here again, it branches into three versions, and, once again, the branch based on the anticipation created when the previous burst was terminated is the one that wins out, going on to form the complete explanation that is the one returned by the analysis procedure.

5.1.2 A speech example

The artificial Bregman noise example was useful in highlighting many aspects of the system performance, but it is not sufficient to build a system that can handle the contrived tests of psychoacoustics if it cannot also handle the far more common examples of the everyday world. The noise example also failed to contain periodic sounds that would involve the 'weft' element. We will now consider a brief fragment of speech to redress these omissions.

Figure 5.7 summarizes the system's analysis of the example – a male voice saying "bad dog" against a background of office noise. In addition to the time-frequency intensity envelope, the periodogram (the summary of the three-dimensional correlogram volume projected onto a time-frequency plane) is also displayed for the entire sound. The analysis consists of all three kinds of element – noise, click and wefts; the weft elements are displayed as both their energy envelope and their pitch track, on the same axes as the periodogram of the input sound.

The analysis of the sound consists of three components, each represented by a different type of sound element. Firstly there is the background noise (the recording was made in an office using a distant microphone). A fairly steady background has been captured as a static profile in the element Noise1. Note that having over 300 ms of 'run-in' noise in the example before the voice starts was very helpful in giving the noise element a chance to stabilize.

The second component is the voiced speech, represented by the two weft elements which are displayed together as Wefts1,2. Looking at the periodogram of the original sound, some prominent features appear a little before time = 0.4 seconds as the voice begins. These features indicate the periodicity of the wide-band voice energy visible in the intensity envelope, and occur at the fundamental frequency of around 100 Hz, as well as an octave below that at about 50 Hz. The periodogram features are explained by the period-track of the two wefts which follow the higher pitch; the upwards search in lag time (corresponding to a downwards search in fundamental frequency) followed by the cancellation of higher-order autocorrelation aliases ensures that the 50 Hz subharmonic is properly interpreted as part of the 100 Hz feature (in the manner illustrated in figure 4.16). The intensity spectrum extracted for the wefts (by sampling the 3-D correlogram volumes at the appropriate time, frequency channel and lag co-ordinates) appear to follow

the energy visible in the full signal quite closely; note, however, the holes in the weft envelopes, particularly around 300 Hz in the second weft; at these points, the total signal energy is fully explained by the background noise element, and, in the absence of strong evidence for periodic energy from the individual correlogram channels, the weft intensity for these channels has been backed off to zero.

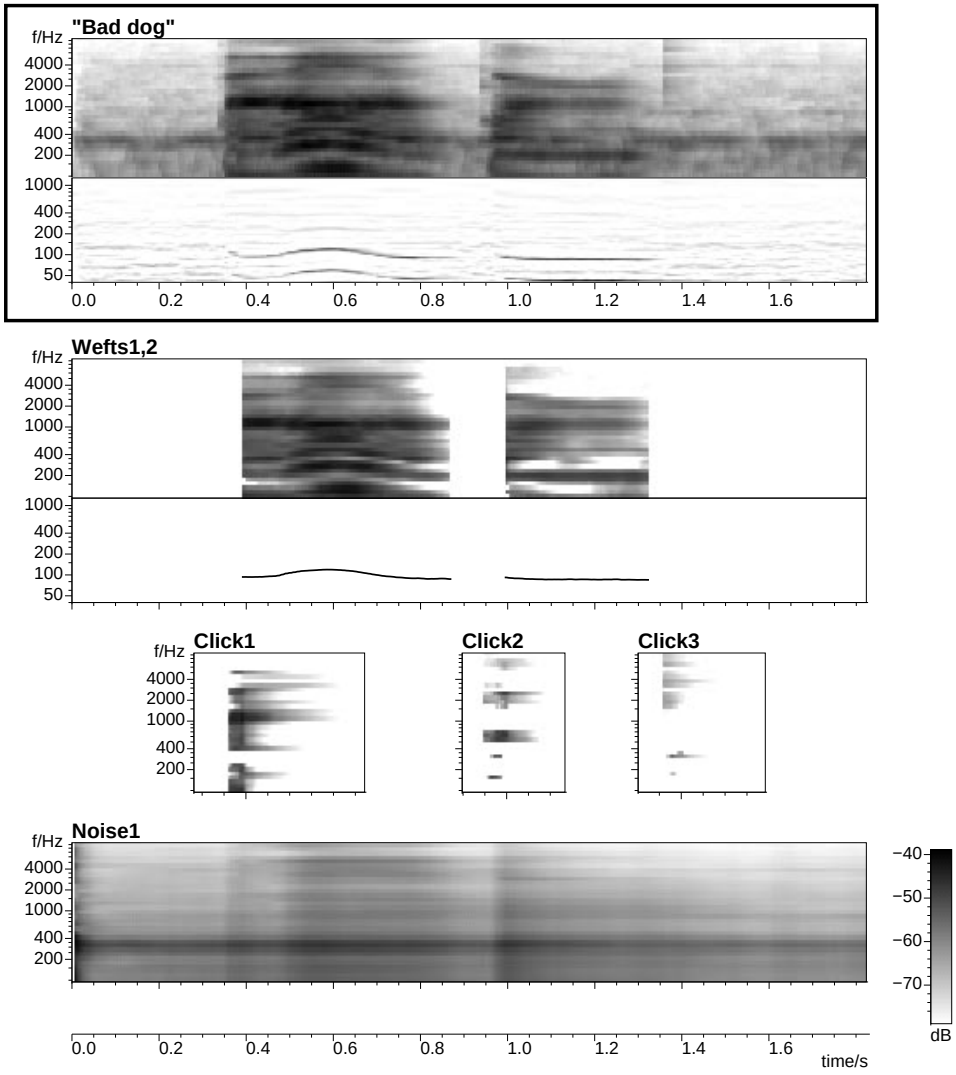

Figure 5.7: The "bad dog" sound example, represented in the top panes by its time-frequency intensity envelope and its periodogram (summary autocorrelations for every time step). The noise and click elements explaining the example are displayed as their intensity envelopes; weft elements additionally display their period-track on axes matching the periodogram.

The third component of the sound comprises the click elements which capture the aperiodic transients of the speech – notionally, the stop releases for the /b/, /d/ and /g/ phonemes in "bad dog" (the two syllables are elided). The click onsets have been well located at the beginnings of the syllables; however, the rapid increase in energy from the voicing, along with a small amount of delay in creating the weft elements, mean that the click onsets have become blurred over about 50 ms before settling into their final decay. Click3, encoding the release of the /g/, has a clean attack, but has also picked up some energy around 350 Hz which probably doesn't belong to it, arising instead from a chance alignment of a fluctuation in the background noise. Although the noise elements are all where they ought to be, in a resynthesis consisting of just wefts and clicks with the background noise removed the clicks do not fuse well with the speech, sounding more like distinct transients that stream apart from the vowel-like periodic signal. The factors governing integration of vowels and consonants are rather complex [Klatt83], and they are not well preserved in the resynthesis of this example; I suspect that most of the fault lies in click elements that are too blurred and incorporating too much low- and middle-frequency energy that does not belong with them.

In the figure we see only the 'winning' explanation. In particular, each of the click elements was chosen in preference to an alternative hypothesis which used a noise element to account for the same energy onset. Click elements provided a preferable explanation in each case because of their intrinsic decay; the tendency of noise elements to be sustained was not a successful match to these examples where the first two transients are rapidly followed by voicing that can mask sustained noise for a while, but eventually disappear, leaving a noise explanation for the onset with predictions too large for the observations; in the click-based hypotheses, the onset-explaining click has already disappeared by the time the voicing fades away, so there is no inconsistent prediction at the syllable ends. Also, the bulk of the syllable has been explained by one element (weft) rather than the overlap of two (weft and noise), leading to an additional rating advantage.

The only other event of interest in the analysis occurred at around t = 0.5 s during the first syllable. The rapid pitch and formant shifts of the vowel led to a situation where the weft prediction was lagging a little behind the actual data, and the system found itself with a prediction inadequate to explain the observed energy in some frequency channels. However, the attempt to construct a new noise or click element to account for this excess was suppressed because the onset map gave insufficient evidence for a genuine onset (discussed under 'hypothesis creation' in section 4.4.2); within a couple of time-steps, the weft had caught up, and the inadequacy resolved itself.

5.1.3 Mixtures of voices

As mentioned in chapter 2, one of the perennial problems considered in speech processing is the separation of overlapping, periodic signals. While not an explicit focus of this project, it is without doubt an important task in the analysis of real sound mixtures, and moreover one that the within-channel periodicity detection of the weft analysis scheme should be able to handle. Here we consider a single example of this kind, illustrated in figure 5.8. The sound is a mixture of male and female voices, one of the test cases used in [Cooke91] and [Brown92] and made available by the latter; Brown's designation is v3n7 (the third continuously-voiced sample mixed with the seventh interfering noise example).

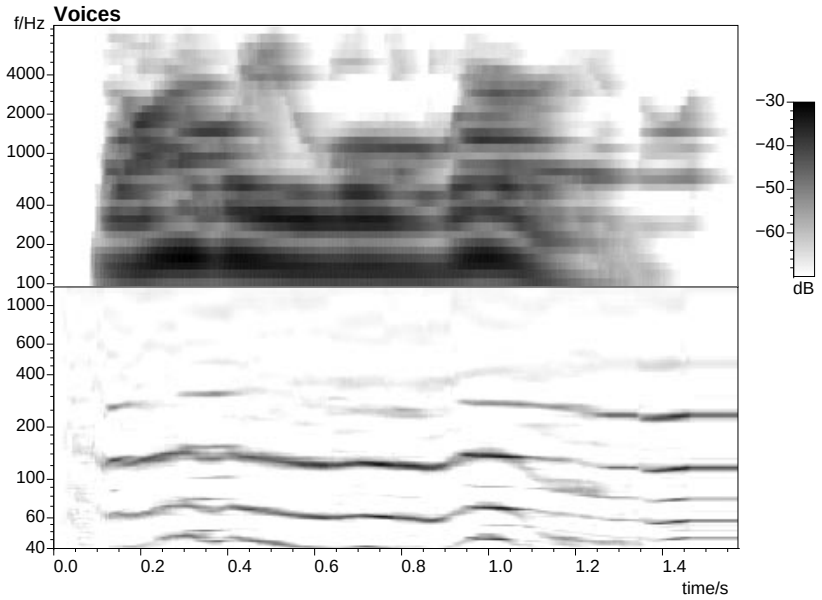

Figure 5.8: The Voices example, a mixture of male and female speech. The upper pane shows the time-frequency intensity envelope, and the lower pane shows the periodogram (summary autocorrelations displayed on a time-frequency plane).

Looking at the intensity envelope, there is little evidence to suggest that two voices are present, and indeed the pronounced harmonics in the lower channels suggest a single voice. In fact, the male-voice is rather louder than the female in the lower spectrum, leading to visible features. Looking at the periodogram, however, reveals that there are indeed several signals present. The well-defined period ridge starting at t = 0.1 s, f = 260 Hz, is clearly distinct in origin from the longer ridge starting at t = 0.1 s, f = 120 Hz. Interestingly, this is mainly apparent based on their different developments – the higher ridge stops abruptly at t = 0.25 s, whereas the lower one continues. Indeed, both ridges start at almost the same time and are close to an octave apart when they do, potentially a difficult separation task.

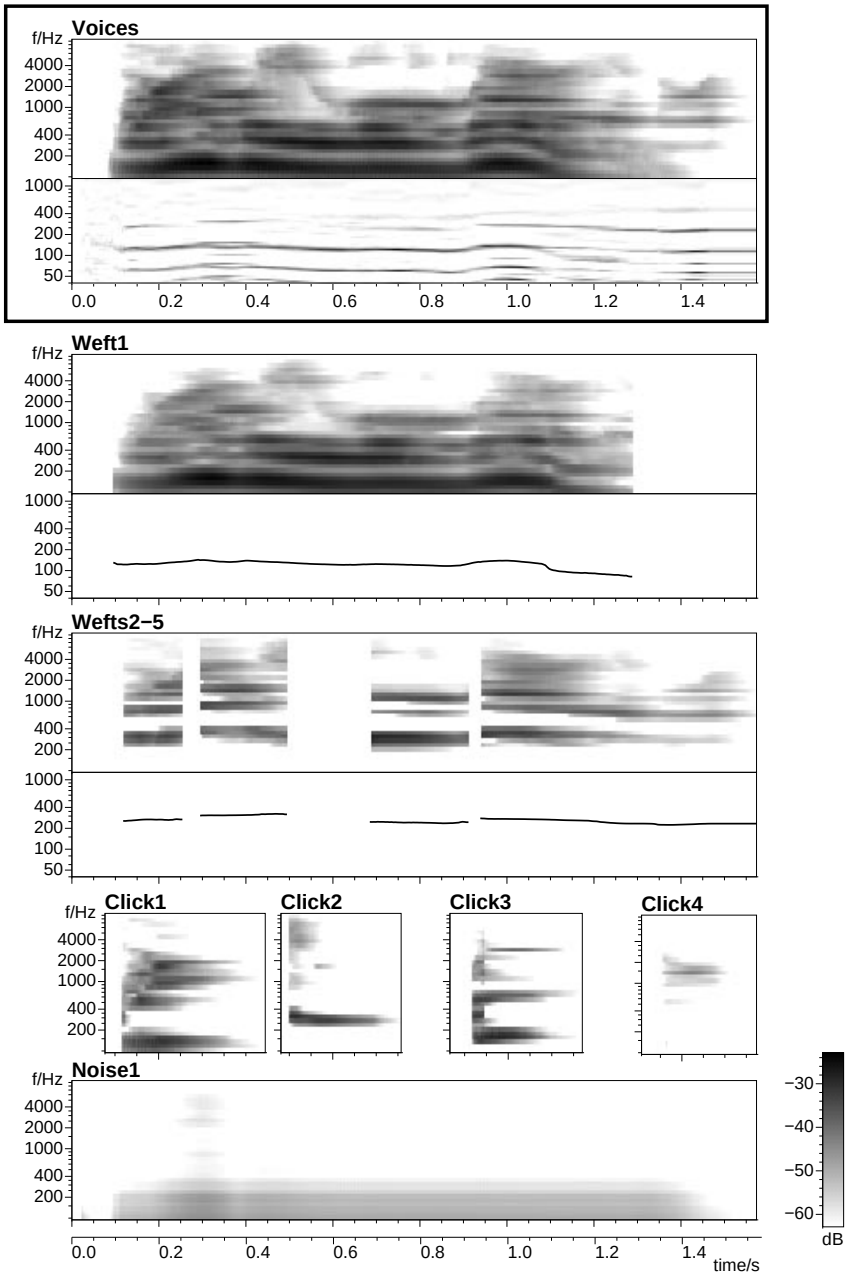

Figure 5.9: The system's analysis of the Voices example as wefts, clicks and background noise.

The system's analysis is shown in figure 5.9. The voices have been analyzed into separate wefts; the male voice was uttering a continuously-voiced phrase (a deliberate choice by the creator of the example to avoid the issue of reconnecting separately-extracted phonemes). This is represented by the long Weft1 whose fundamental frequency is mainly a little above 100 Hz. The female voice, uttering more typical speech punctuated by gaps and unvoiced portions, is represented at least over its periodic stretches by the remaining Wefts2-5 with fundamental frequencies between 250 and 300 Hz.

As in the previous example, the system has generated several additional elements to account for the extra energy not explained by the wefts. These are shown in Click1 through Click4 and Noise1, which has managed to construe the presence of a small amount of background noise even in this short example without 'gaps' (in which the background would be most visible). Click1 and Click3 occur at syllable starts, plausibly attempting to model the /d/ and /c/ of the female's phrase "don't ask me to carry...". Click2 would normally have been suppressed, since the impetus for its creation was the disappearance of Weft3 as its periodogram track fades from view. Normally, a sudden prediction deficit caused by the removal of an explanatory element is ignored because there is too little support from the onset map for the creation of a new element; for this analysis, the threshold for onset map support was made very small in an effort to permit something like Click2 to be created in the hope that it would capture the sibilant energy of the /s/ in "ask", which otherwise leads to a large gap in the wefts representing the female voice during t = 0.5 – 0.7 s.

Unfortunately, the element that was created fails to provide a very helpful addition to a resynthesis of the female voice based on Wefts2-5. Partly this may be due to the same weaknesses that made the additional click elements stream separately in resynthesis of the single-voice "bad dog" example. But for mixtures of voice, the situation is even more difficult. While overlapping periodic signals provide at least a hope of separation on the basis of their distinct periodicity cues, there is no such bottom-up basis to separate overlapping unvoiced signals. (Weintraub's state-labeling system was capable of attributing this situation to its input, but was unable to do anything more sophisticated than splitting observed energy equally between the two voices [Wein85]). In theory, it is a system capable of imposing top-down constraints that will be able to handle such situations successfully. However, the necessary constraints are phonetic and possibly semantic in nature, and the current system did not contain any knowledge at this level of sophistication. One can imagine the behavior of a prediction-driven system which did know enough about speech to recognize the hints to the /a/-/s/ transition such as the ones employed by human listeners to hear out the female speaker.

Although the current system was unable to extract the unvoiced portions of the speech with much success, its ability to separate both voices is worth examining in more detail, particularly in view of the potential confounding of pitch and onset cues mentioned above. (Cooke and Brown only reported separating the male voice that dominates in the lower frequencies and do not mention any extraction of the female voice, which, being only intermittently voiced, lay beyond the scopes they had established for their algorithms). The periodogram is shown on a larger scale in figure 5.10 with the period tracks of the five extracted wefts drawn over the features they are tracking. The lower frequency period ridge starts a little earlier than the higher voice, and this accounts for the initial creation of the lower weft. But we might expect that the arrival of the higher voice would cause the destruction of the lower weft, which is now confounded with the subharmonics of the upper voice. Fortunately, the somewhat linear properties of the summary autocorrelation, where autocorrelations from channels dominated by the higher voice are added to those containing the lower voice, permit the successful disentangling of the alias of the higher weft's period from the base ridge of the lower weft: The alias-removal stage of the weft analysis, where shifted versions of detected peaks are subtracted from the overall summary to suppress the tracking of subharmonics, only removes part of the total summary

autocorrelation peak at the period of the lower voice, leaving enough of a peak behind to cause the system to recognize and extract the lower voice. This is illustrated in the lower panel of figure 5.10, which shows the residual periodogram after the removal of the features tracked by wefts 2-5 and their aliases. These wefts are tracked and removed first (at each time step) because they have the shortest periods, and the analysis searches upwards in period. This modified periodogram is the basis upon which weft 1 is extracted. Note how a second octave collision around t = 1.0 s in the original sound has been successfully removed, allowing the lower voice to follow its declining pitch track, even though the low-frequency based periodicity evidence is quite weak at this point. This points to an interesting potential for autocorrelation-based representations to be able to accommodate octave collisions even on a bottom up basis, effectively exploiting local smoothness in harmonic amplitudes, something that traditionally eludes narrowband Fourier systems (e.g. [QuatD90]). However, the situation of voices whose pitch tracks actually cross (as distinct from colliding at octaves) can still only be solved by the addition of top-down context constraints.

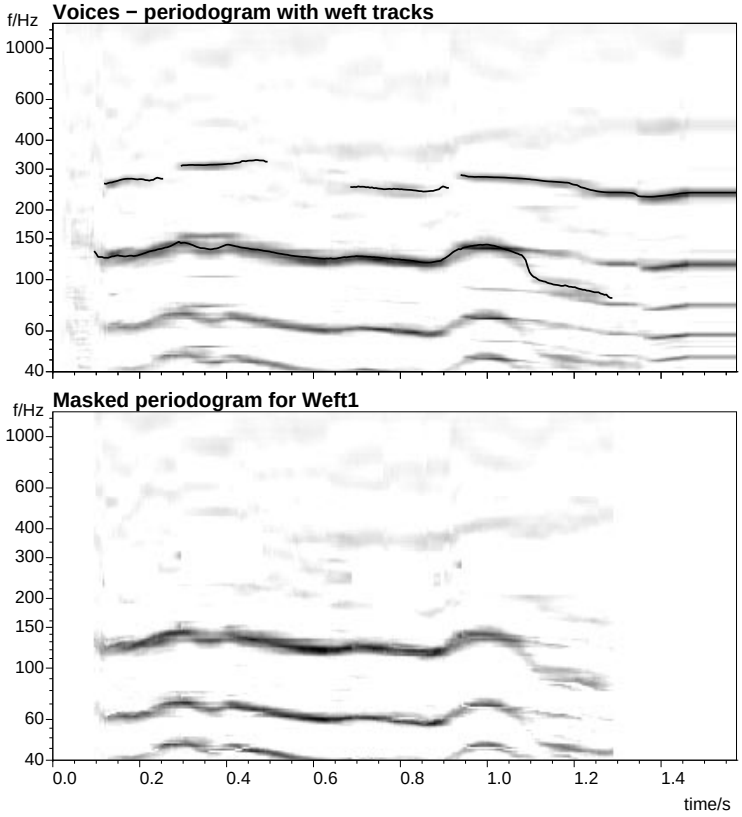

Figure 5.10: Upper panel: expanded view of the periodogram for the voices example with the period tracks of the extracted wefts overlaid. Lower panel: residual periodogram after the features associated with Wefts2-5 have been subtracted. This residual permits the correct tracing of Weft1, without it being pulled off course by a strong colliding alias such as at t = 1.0 s.

5.1.4 Complex sound scenes: the “city-street ambience”

The final example in this section represents the class of dense sound scenes that originally motivated the project; more examples of this kind will be used in the discussion of the subjective listening tests. Figure 5.11 shows the intensity envelope and periodogram of the city-street ambience sound, along with the fifteen elements created by the system to account for it. Although it is rather difficult to see many features in the original sound, the most prominent components to my ears are the honking of car horns, particularly right at the beginning and close to the end, and a ‘crash’ sound as if of a large metal door being slammed, all overlaid on a steady roar of traffic and pedestrians.

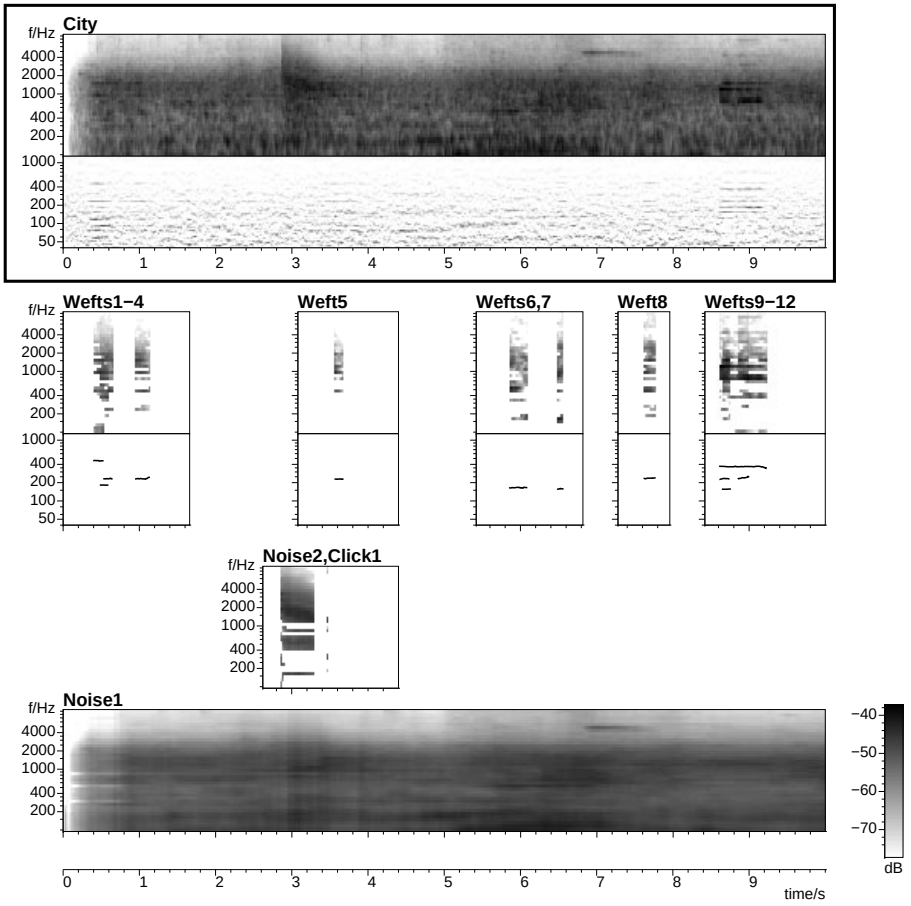

Figure 5.11: Analysis of the city sound. Note that the time scale for this ten-second sound example is compressed in comparison to the shorter examples considered so far.

The system has extracted these major components (not as positive as a result as it might appear, since extracting these features was to quite a large extent the driving force behind, and the test case used in the development of the system!). The first horn, whose various periodicities are just about visible in the periodogram, has been modeled as a cluster of four wefts during the first 1.0 s of the sound. (The grouping of these wefts into a single element was done by hand, but ultimately would be accomplished at the level of source hypotheses). The original car horn clearly has at least two notes, so its

representation as overlapping wefts with differing pitch is appropriate. Similarly, the final horn sound, whose spectral energy is somewhat visible at t = 8.6 s, f = 700–1500 Hz, has resulted in Wefts9-12. The perceptual similarity of these weft clusters to the sound mixture is addressed in the subjective tests; here, we merely note the intended correspondence.

The 'crash' sound actually has a very slow decay in the original sound and has been modeled by Noise2, a noise element; the click element we might expect hit the upper limit of its permissible decay times, and simply couldn't accommodate the sound. The unfortunate result of using a noise element is that the eventual decay of the crash is not handled very gracefully; by about t = 3.4 s, the energy of the crash has decayed sufficiently that the existing background element, Noise1, provides an adequate explanation. Since Noise2 will not decay of its own accord, it is constantly over-predicting, and being caused to decay only by its allocation of negative prediction error from the reconciliation process. When the actual sound transient has decayed sufficiently, the inconsistent-explanation events generated by this overprediction cause the perfunctory termination of Noise2; a better fit might have been obtained by a click element with relaxed decay-time constraints, which would presumably have avoided over-prediction, and would have thus been left to decay more gracefully to zero even when its contribution to the overall signal model was small enough to be eliminated without impact. This is the intended operation of the prediction-driven architecture, where the absence of direct evidence for an element does not prevent that element from forming part of the explanation, so long as the explanation including that element is still consistent.

The analysis also generated a single click element, the briefest of bursts shown just after Noise2. I'm not exactly sure what this is meant to explain, but it is a fairly harmless addition. A more serious problem is the failure of the system to separate the rather obvious narrowband high-frequency feature visible in the intensity envelope at t = 6.8 s, f = 5000 Hz. Instead, this has been spuriously incorporated into the background Noise1. This energy is clearly audible as squealing brakes in the example; the reasons for it being overlooked by the system are considered below in the discussion of the subjective responses to this sound.

5.2 Testing sound organization systems

A qualitative description of the system's behavior helps to convey an understanding of its intended operation, but fails to give any quantitative references for its success at analyzing complex sound mixtures into their components. This is a difficult thing to supply, since we do not have a neat, objective definition of the process being modeled. However, some kind of assessment scheme, by which success and progress can be measured, is very important, both to satisfy the philosophical demands of scientific research, and also for the more practical considerations of obtaining and sustaining support. Comparisons may be odious, but it is clear that the adoption of common assessment standards is usually beneficial to the fundability of a field – as shown by the focus of speech recognition community on well-defined metrics and test sets during their rapid progress in the 1980s. As a field, computational auditory scene analysis has yet to reach a stage of consensus concerning its goals that would permit agreement on a common set of tests and metrics; perhaps this should be a priority if we wish to see continued development in this area. In this section, I will make some observations on

the difficulties in assessing models of sound organization, then describe the subjective tests used with the current model. The test results are presented in section 5.3.

5.2.1 General considerations for assessment methods

Of the previous models of auditory scene analysis, none were assessed by conducting any kinds of formal listening tests, i.e. tests to correspond the systems' analyses with those of actual listeners whose auditory systems were supposedly being modeled. The closest approach to this kind of assessment may be found in hearing-aid research, where novel processing algorithms are assessed for their ability to improve the intelligibility performance of hearing-aid users. This is not a particularly appropriate test for the systems we are considering, since although perfect source reconstruction should enhance intelligibility, scene organization is a task quite different from speech understanding. The technologies involved are very different too: Even the most sophisticated algorithms proposed for hearing-aid applications have been limited to nonlinear spatial filters rather than modeling the kind of feature-based organization being considered in this work ([KollK94], although see [Woods95]).

In chapter 2 the various objective measures used by previous modelers of auditory scene analysis were discussed. While these metrics were useful in illustrating the performance of the systems in question, they were very specific to these projects, and, as noted in relation to [Brown92] (who suffers for his generosity in making his sound examples available!), often gave scores that failed to correspond to subjective impressions of the systems' resyntheses.

The assessment problem comes down to this: A subject listens to a complex mixture of sounds and has an internal experience of that sound as being composed of a collection of distinct sound events. I am trying to build a computer model to reproduce this analysis, but the only way to measure the original is to ask questions of the subject via some kind of psychoacoustic task. Rather little is known about this internal experience, indeed some researchers might take issue with this characterization of the behavior of the auditory system, arguing that there is no genuinely distinct representation of components of a mixture, only various properties associated with the mixture as a whole. It would be interesting to have experimental results addressing this question (such as discrimination tests for a complex sound in a particular mixture context and alone) but that is beyond the scope of the current work.

Assume that we had a prototype computational auditory scene analyzer producing isolated resyntheses of very high quality. One way to test it would be to construct artificial mixtures from recordings of 'single' sound events, and then to compare the system's output with the unmixed originals. Expecting them to match would be to assume that the auditory process we are trying to model is able to successfully and exactly undo the mixture involved in creating the sound. Our subjective experience is that perceptual segregation of components in a sound mixture is usually successful (we are rarely mistaken about the properties of any single source we can distinguish in a mixture), but the assumption that we can *exactly* perform this separation is more suspect. A criteria of exactness that measures the root-mean square error between pre-mixture and post-resynthesis signals is clearly nonsense, since the listener cannot detect that level of exactitude under ideal, isolated listening conditions, let alone in the context of a mixture. Yet if we admit that a certain amount of waveform distortion is tolerable, we are immediately

obliged to involve a human in the assessment loop, since there do not exist good objective measures of perceived similarity between sounds (although such a measure would be very valuable, not least in the development of high-quality audio compression schemes [Colom95]).

This subjective component could take the form of a listener comparing the pre-mixture original sounds and the system's output. If the system were performing extremely well, these two sounds might be difficult or impossible to distinguish, or the listener might have no preference between the two. Unfortunately it may be some time before we have separation systems capable of performing at this level. Even if this were feasible, it still might not be an entirely fair test, since by comparing pre-mixture originals to post-mixture reconstructions we are still making an assumption that the auditory separation process can identify every aspect of a sound in a mixture. While we may be rarely mistaken in our perceptual sound analysis, there may still be many inaccuracies in our impression which are simply of no consequence. A more appropriate subjective test would be to play the *mixture* to the listener (rather than the pre-mixture components), then play the resyntheses extracted by the system to see if they matched perceived objects in the mixture. This would be testing our assumed internal representation in the most direct fashion. A way to construct this experiment along the lines of a two-alternative forced-choice task would be to play the mixture, then both the original component and the resynthesized component, and have the subject choose which one sounded more as if it was present in the mixture. The weakness of this test is that there may be systematic differences between originals and resyntheses whose effect on the preference scores overwhelm the finer details being manipulated. It also carries the underlying assumption that the perfect analysis corresponds to the pre-mixture originals, as do all these hypothetical tests based on synthetic mixtures.

Another way to get back towards a forced-choice paradigm would be to construct pairs of mixtures, one from the original sounds, and the other with one or more of the components replaced with their resynthesized counterparts. If the two mixtures were difficult to distinguish, that would be a reasonable success criterion for the analysis model. This might be a more reasonable test, since details of the *original* that were simply not perceptible in the mixture would be similarly masked for the resynthesis in the same context, thereby making a task that was less likely to be overwhelmingly skewed in favor of the originals. However, we are again faced with being some way away from even this level of performance, but wishing to assess our current-day imperfect systems none-the-less. In the following section, I will present the subjective tests eventually devised to assess the current system which are more appropriate for current, imperfect scene-analysis models. This approach dispenses with pre-mixture 'ideal' isolated signals, so that any recorded sound mixture may be used, and only the listener's internal experience of the separate objects in that mixture is being queried.

5.2.2 Design of the subjective listening tests

Consider the case where we wish to test a subject's perception of a real-world complex sound mixture for which we do not have access to 'reference' isolated components that add together to make the mixture. By dispensing with artificial mixtures, we expand the scope of our stimuli to all kinds of real sounds which may have characteristics significantly different from synthetic mixtures (in terms of reverberation, coloration, density etc.). Further, the question of whether the auditory system achieves an 'ideal' separation is

sidestepped, since no ideal pre-mixture components are involved. But how can we quantify what a subject actually hears in such mixtures? The experiments that were conducted sought to make some kind of direct measurement of that experience, as well as evaluating how closely the automatic analysis matched the listeners' internal representation.

The experiment consisted of three main parts. The first part involved only the original sound-mixture, and was intended to gather information on the listener's perception of a dense sound mixture before being exposed to any of the system's efforts at separation. The second part then asked the listeners to identify and rate the resyntheses from the model, and the third part involved ranking slightly different versions of the resyntheses to verify the local optimality of the model's outputs. Each part is now described in detail.

Part A: Labeling a real sound

The task was for the subject to listen to the sound example, and to supply names for the most prominent distinct events they could hear in the sound. The subjects also indicated the approximate time support of each object they named, more as an aid to disambiguation than as a direct test of perceived event time. The data was collected through a computer terminal with a display, mouse and keyboard. The screen display for this part of the experiment is reproduced in figure 5.12.

The subjects can play the original sound mixture as often as they wish by clicking the 'Play Original' button. (In the experiment, the sound examples were mostly about ten seconds in length, although one was shorter). When a particular sound event has been identified in the sound, a name for it is typed into one of the boxes on the left of the screen. The subjects also have a key, and they are instructed to concentrate on the particular component in the mixture, and hold down the key when they hear that event during the playback of the sound. While the sound is playing, an animated time cursor moves across the display; pressing the key leaves a 'mark' behind this cursor, indicating the reported time support along a left-to-right time axis. The subjects are able to alter this mark (for instance, shifting it to remove a reaction-time lag, or to eliminate spurious responses) by dragging it with the mouse in the manner of a typical computer 'draw' program. By adjusting the marks and replaying the sound to check the correspondence of the auditory event of the perceived object with the visual event of the animated cursor passing over the existing marks, subjects have the opportunity to record time-supports with a reasonable degree of accuracy (they were however instructed not to spend more than a couple of minutes on this task). This process is performed successively for each perceived object or event, until the subject is satisfied that all the 'important' components in the mixture have been labeled, and the indicated timings are acceptable.

This test obtains some kind of 'ground-truth' for the objects that listeners perceive in a dense mixture. Despite inter-subject variation, the results presented in the next section show a surprising amount of consensus concerning the four or five most prominent components. (This experimental setup is somewhat related to that in [Kaern92] [Kaern93], although his repeated noise stimuli evoked highly variable responses). The text labels for the different events have two functions. They can confirm the correspondence, initially suggested by similarity in time-support, between the events recorded by different subjects. The labels are also used in the next parts of the experiment to allow the subject to refer to their own experience of the sound when assessing the model's analysis.

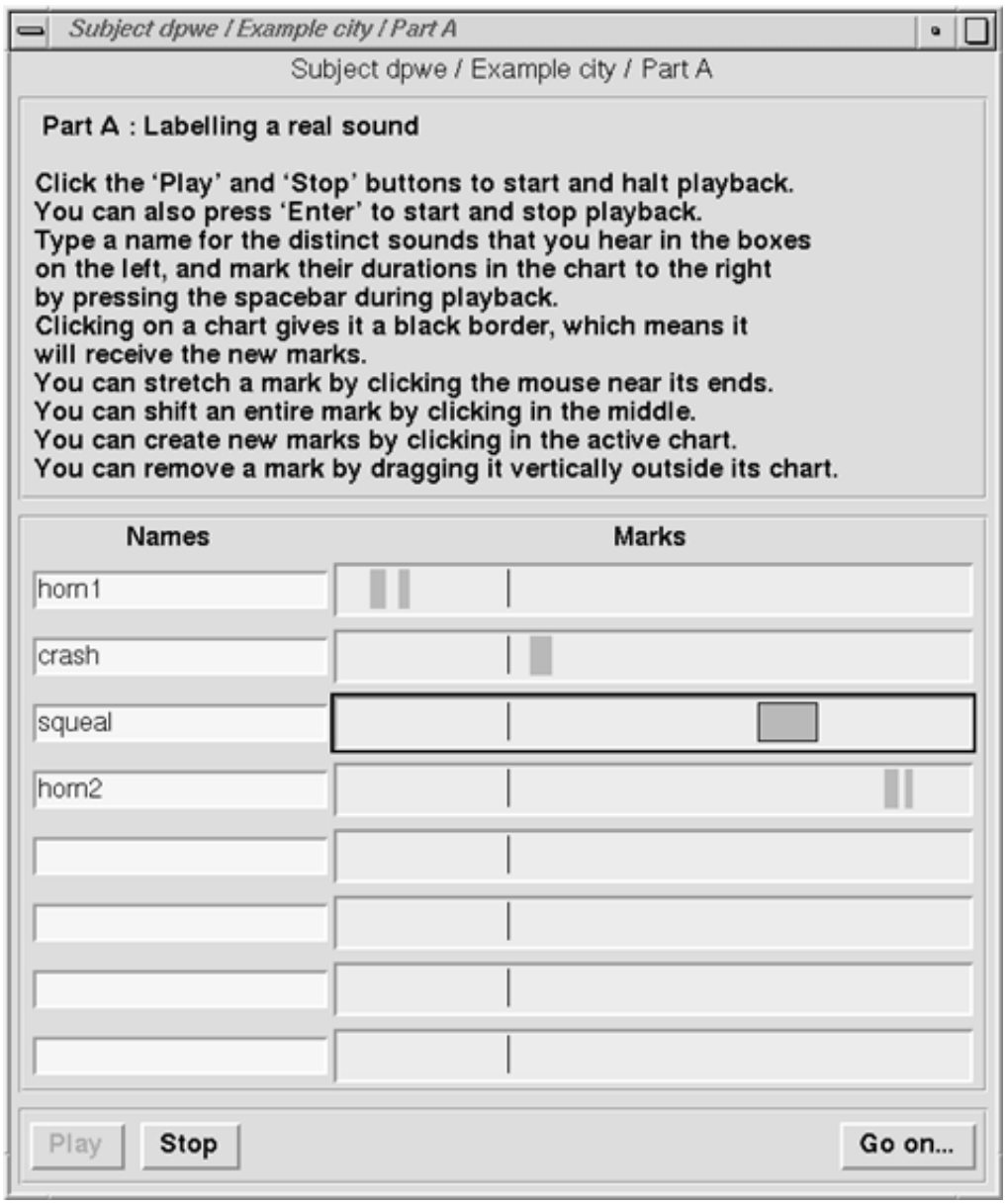

Figure 5.12: Example response screen for part A of the experiment. Subjects type in names for the events they hear in the boxes on the left; clicking the "Play" button replays the original sound at the same time as moving a cursor over the boxes on the right (shown). Pressing the space-bar when the event occurs leaves a gray mark under the cursor which can be 'trimmed' or revised.

Part B: Rating individual resyntheses

The second part of the experiment attempts to link the events reported by the subject with the events generated by the automatic scene analysis system under test. The subject is played a resynthesis extracted by the system from the original sound mixture, and is asked to indicate which, if any, of the previously-indicated events it most resembles. This experiment can be repeated for several objects extracted by the model. Thus, provided the resyntheses bear a resemblance to the events as perceived by the subject, this

test provides evidence that the events detected by the model match the listener's internal experience. An example of display for part B is reproduced in figure 5.13.

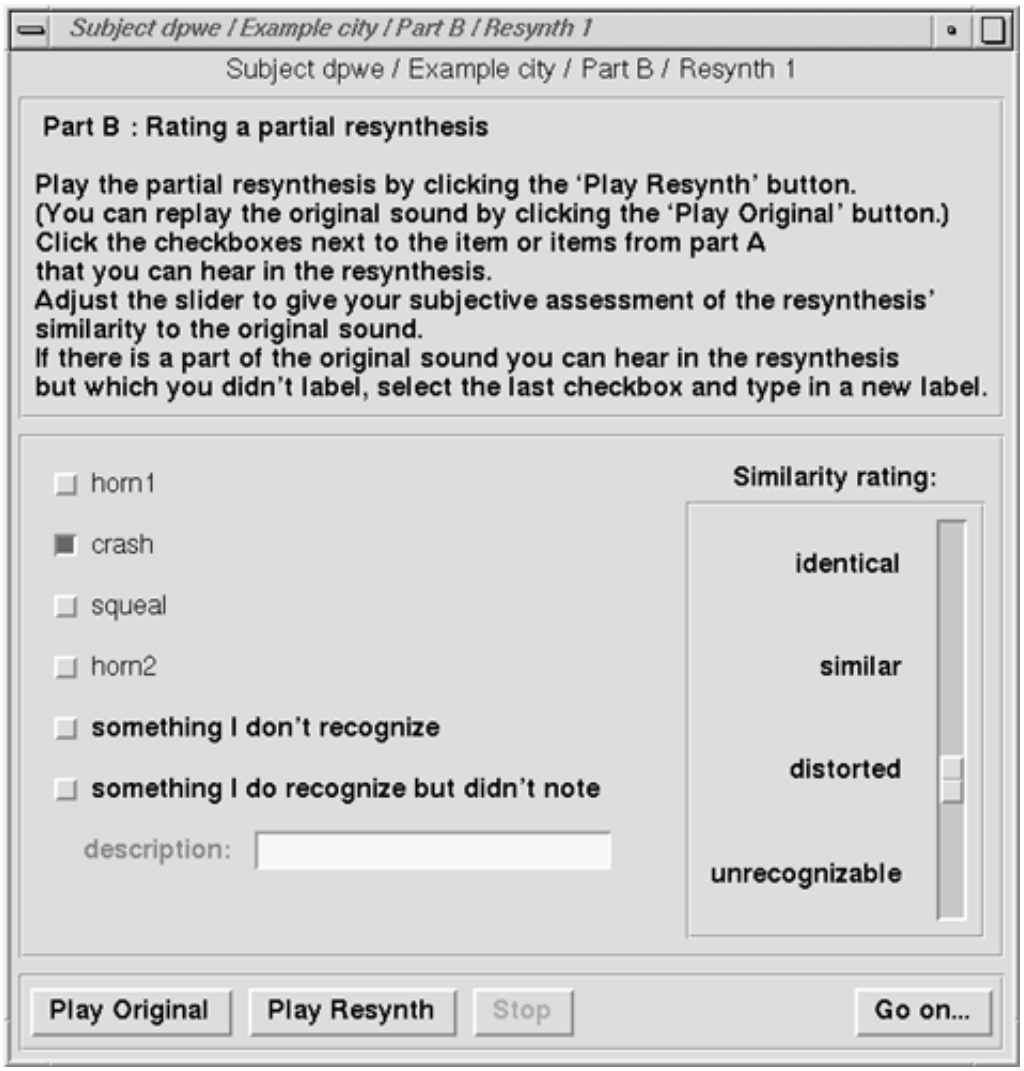

Figure 5.13: The response screen for part B of the experiment. The labels on the left-hand side are copied from the names typed in by the subject in part A of the experiment. Clicking the 'Play Resynth' button plays the resynthesis example under assessment for this particular screen, which may be compared to the original sound mixture by clicking the 'Play Original' button.

The subjects are presented with the list of events they named in the previous part of the experiment, and they can click checkboxes next to each one to indicate a perceived resemblance between the presented resynthesis and the percept they have named. There are additional checkboxes to indicate failure to recognize the resynthesis as any part of the original sound, and also a place to enter a new label, in case the subject finds themselves wishing to indicate a correspondence to a percept which they did not name in part A, either inadvertently or because they judged it too insignificant. (Once they have started listening to the resyntheses, the subject cannot go back and change their responses to part A).

Assuming correspondence can be successfully established, the subject then rates the perceived quality of the resynthesis with the slider on the right of the display. It is quite possible that a resynthesis will be identifiable as an imitation of a particular part of the original sound while still differing from it considerably in detail. (Indeed, this is generally the case with current models!). The subjective scale ranges from 'identical' to 'unrecognizable', although it is unlikely that either of these extremes would be used – only the most distracted of listeners could consider the resyntheses used in this experiment as 'identical', and a genuinely unrecognizable resynthesis would, arguably, be unratable since it is not known to what it should be compared. (The default, initial position of the slider is at the bottom, i.e. "unrecognizable"). Clearly, the results of such a subjective rating scale will vary according to the interpretations and standards of individual listeners. Aggregate responses can still present some kind of consensus, and useful comparative results can be obtained by normalizing the results of each subject through rank ordering or some other ensemble scaling. Subjective quality ratings are usefully employed in the assessment of hearing aids [KollPH93].

Part C: Ranking of resynthesis versions

The third part of the experiment is intended to assess how well the system extracts events from the sound mixture within the space of objects that it can possibly represent, as determined by its parameterizations and resynthesis procedures. The idea is to produce a set of different versions of a particular resynthesized event, including the 'best effort' of the system along with several 'distorted' versions whose parameters have been altered in some systematic way. The subject is asked to sort these different versions according to how closely they resemble the perceived object to which they apparently correspond. Notionally, a given system output occupies a point in a high-dimensional space of model object parameters; by sampling a few nearby locations, it might be possible to show that this point is a 'local minimum' in terms of its perceptual distance from the original. The difficulty with constructing this part of the test lies in choosing the size and 'direction' of the step to the neighbors i.e. coming up with distortion methods that are just discriminable, and that can provide a useful indication of relative resynthesis quality. If a certain 'distorted' version is preferred systematically over the 'best' version, this can provide useful guidance for improving the system. If there is a random spread of preferences over the different versions, the space spanned by the versions does not exhibit much variation in perceptual similarity, possibly because the steps are too small, more probably because the resemblance is so poor.

The response screen for part C is illustrated in figure 5.14. Subjects can listen to each different version of the resynthesis by clicking one of the buttons in the central panel, and can repeat the original mixture with the 'Play Original' button at the bottom. The buttons for the individual versions are movable with the mouse, and the subject can arrange them from best to worst as indicated on the display.

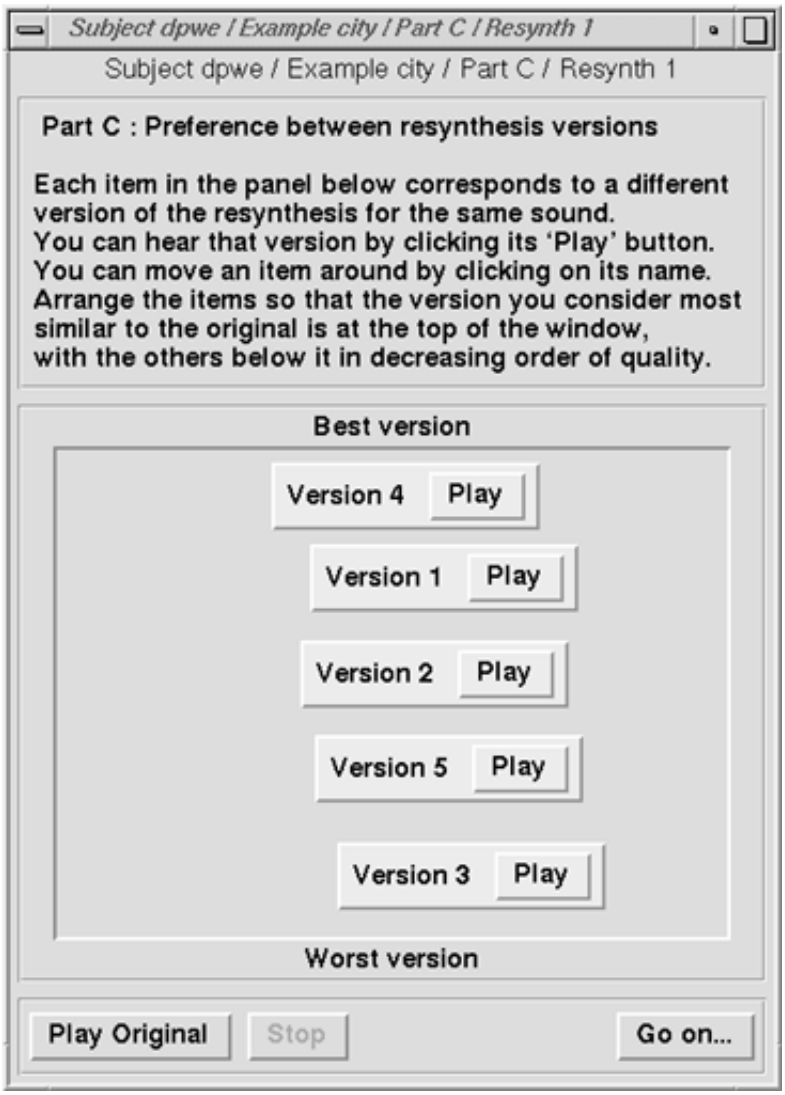

Figure 5.14: Example display for part C of the experiment. The central panel contains five moveable buttons; clicking the 'play' label plays the particular resynthesis version associated with that button; clicking on the name portion allows the subject to move the button around the screen. The subject's task is to arrange the buttons so the 'best' version (in terms of resembling a part of the original mixture) is at the top of the panel, with successively worse versions below it. The initial ordering of the buttons is randomized.

The particular form of the distortions used in the current experiment are described in section 5.3.6 along with the presentation of the results.

Other experimental procedure

The three parts to the experiment were arranged with part A first, to collect the subjects' labeling for the original sound mixture, followed by a series of part B responses for several resyntheses derived from the sound mixture, then a series of part C trials presenting versions of the resyntheses that had been rated in the part Bs. Separating the B and C parts for a given resynthesis made it less likely that the subject would recognize the 'ideal' version in part C as matching the one presented in part B. The entire

sequence (A, Bs and Cs) was conducted for one sound mixture before moving on to the next test set of original sound and resyntheses.

The subjects were given a sheet of written instructions before starting which explained the general form and goals of the experiment. In addition, the test started with a training trial, consisting of a single instance of each part of the experiment, which the subject completed under the guidance of the investigator. After the training, the investigator withdrew, and the subject completed the remaining four sound examples. Three of these were ten-second dense sound mixtures, and the fourth was a brief mixture of two voices; each is described in detail in the next section.

The test took on average half an hour to complete, with the fastest subject completing in twenty minutes, and some subjects taking closer to an hour.

General comments

The great advantage of these tests compared to objective measures is that they provide information that genuinely relates model behavior to listeners' subjective experience of real sound mixtures. There is a considerable temptation to work with artificial mixtures where the ideal, pre-mixture separate sounds may be used in assessment, but this ducks the consideration of *real* perceptual organization by assuming that it is 'perfect'. By contrast, the experimental procedure described here does not involve any 'ideal' separations; the listener only hears the full original sound mixture and the resyntheses created by the model.

I offer that these experiments are useful beyond the scope of the current project. The results of part A do not even involve a particular model, but provide a practical, if rather obvious, way to obtain base data on the auditory organization of complex scenes against which different computer models may be compared. The ratings and rankings of parts B and C may be applied to any model whose output is a resynthesis of the components it has identified. The importance of resynthesis is a matter of some contention: Certainly, under these tests a model can only be as good as its resynthesis path, and good resynthesis requires a great deal of care. (The results below indicate that the current system could be improved in this regard). Human listeners do not, in general, resynthesize sound, so why should we include this in models of the auditory system? The problem is that without resynthesis and the kinds of listening tests presented here that it enables, it is extremely difficult to interpret whatever other output the model might provide. Typically, one is reduced to tenuously-grounded interpretations of graphical displays or other non-acoustic media. By contrast, systems that provide resyntheses may actually be compared against each other, directly and quantitatively, using the ratings collected in part B. A proof-of-concept illustration of this was provided by the final, double-voice example used in the current test, where the resyntheses of the current model were compared with the output of [Brown92], from whom the mixture was originally obtained. The results of this comparison are presented in section 5.3.6.

5.3 Results of the listening tests

The experiments described above were conducted for ten subjects (volunteers from among friends and colleagues) for a series of sound examples. This section presents the results of these tests of both the subjective experience of these sound examples and the ability of the current model to duplicate the auditory scene analysis performed by the listeners.

5.3.1 The training trial

By way of introduction to the presentation of the experimental results, let us look briefly at the training trial, in which the subjects completed examples of each part of the experiment under the supervision of the investigator in order to become familiar with the various tasks and modes of interaction. The sound example for this trial was an artificial mixture constructed to have a reasonably unambiguous interpretation. To ten seconds of generic 'crowd babble' were added a single toll of a bell, and, a couple of seconds later, the sound of an aluminum can being dropped onto a concrete floor; both sounds were clearly audible above the crowd and had well-defined beginnings to make the indication of time support as easy as possible, at least for onset.

The responses to part A of this trial are illustrated in figure 5.15, which shows every one of the 25 events recorded by the ten subjects. The responses have been grouped (manually) into sets judged to correspond to the same source, something that was not difficult owing to the considerable consensus in labels and agreement in time supports. The two major groups, 'Bell' and 'Can', include one response from every subject; these groups are summarized in a line at the top of each group by bars connecting the average onset and offset times, with shaded extensions to the bar connecting the maximum time limits indicated by any subject for that event. These summaries of the subject responses are constructed in the subsequent examples for any event named by three or more of the subjects, and are repeated on the figures illustrating the system's output to allow direct comparison between the system's and the subjects' analyses of a given example.

A few points to note: The instructions given to the subjects were not explicit concerning whether they should indicate offset times (i.e. the full extent of the objects they perceived); some subjects evidently 'tapped' the response key at the start of the sound rather than trying to indicate the length of the sound. There is a spread of several hundred milliseconds in the reported onset times, even for the very well-defined events in the trial example (whose actual energy onsets are clearly visible in the time-frequency display). Part of this may have arisen from differences in interpreting the experimental instructions; while subjects had the opportunity to 'trim' the marks that they had made in order to remove reaction-time delays, they were not explicitly told to do so, and some subjects may not have felt this was required. Finally, two subjects indicated the separate bounces of the can sound, again clearly visible in the energy display, whereas the others indicated just the overall duration or the first onset. These multiple responses are incorporated into the summary bars only by their outer time limits; the inner structure, while valuable and interesting, was discarded from the summary.

The training trial included a single resynthesis for one example each of experiment parts B and C. However, this resynthesis was not in fact generated by the current model (it was derived from an earlier, unrelated

model, and from the isolated can sound rather than the noisy mixture) and thus those results are not relevant here.

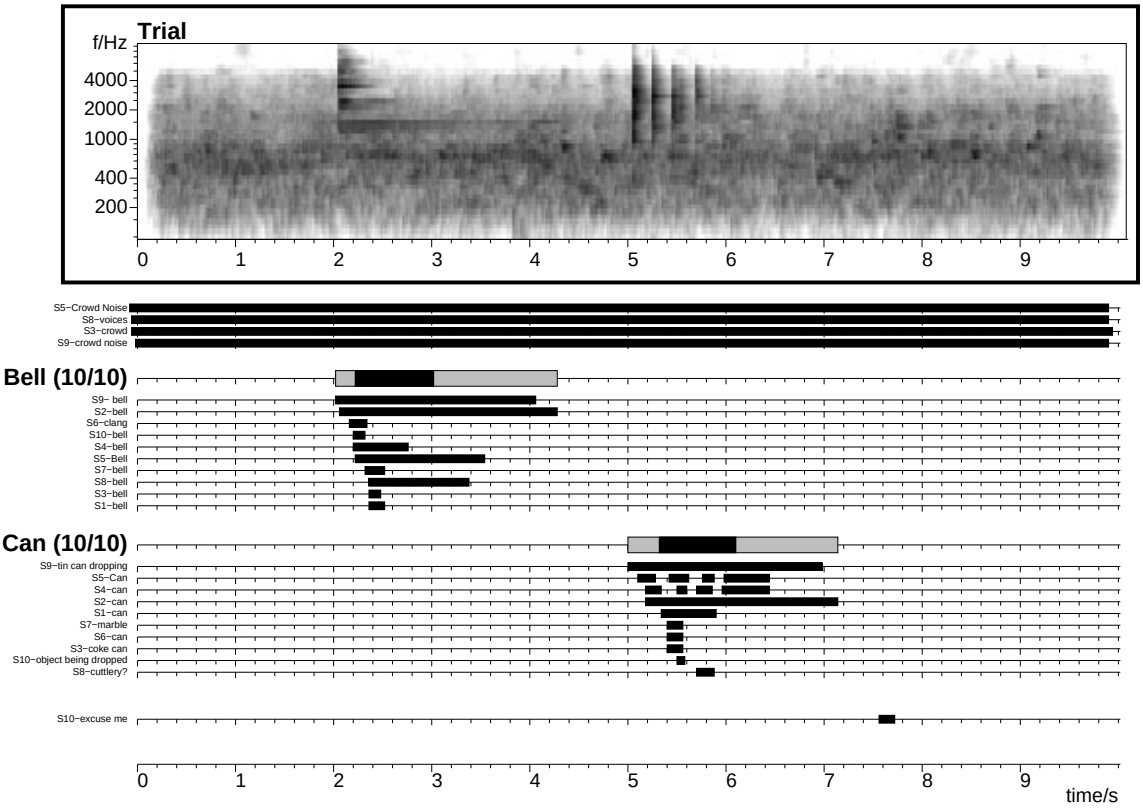

Figure 5.15: Responses to part A of the training trial. The time-frequency intensity envelope for the example is displayed in the top panel. Each individual subject's response line is presented underneath as a time line overlaid with a thick bar showing the marks recorded by the subjects in part A of the experiment. The responses have been sorted into groups on the basis of their similar time-supports and labels. The aggregate properties of each group are summarized in the bar at the top of the group, whose onset and offset times are the mean averages of those in the group, with shaded extensions to the maximum extents reported by any subject. At the left of each response line is the actual label typed by the subject, prepended with the subject number. The labels for the summary lines were chosen somewhat arbitrarily to capture the consensus concerning the event's identity. The figures after these labels indicate the number of subjects reporting that event (i.e. the number of response lines in that group).

5.3.2 The city-sound

The first real sound example was the city-street-ambience, introduced in chapter 1 as part of the inspiration for this project, whose analysis by the system was already discussed in section 5.1. The complete set of subjects' responses from part A of the experiment is illustrated in figure 5.16.

The responses show a surprising amount of consensus concerning the events present in this sound. Prior to the experiment, I was a little anxious that the task would be too difficult; this was based on the informal experience of playing such sound examples to colleagues and visitors by way of illustration for some technical point, only to be met with blank stares; what *I* had heard in the sound, having listened to it over and over again, was quite different from the impression of someone hearing it for the first time. However, it seems that the experimental conditions, in which the subject could attend to the sound without distraction, and was able to replay the example as often as desired, led quickly to a situation where the subjects all experienced very similar impressions of the content of the sound.

Three events, which I have labeled "Horn1", "Crash" and "Horn5", were reported by all ten subjects; happily, these were the events whose resyntheses were chosen for assessment in parts B and C. These events had the smallest spread of onset times, with the separation between earliest and average onset times (i.e. the gray lead-in to the summary bar) a little over 200 ms for the "Crash" event. Of the remaining events, ("Horn2", "Truck", "Horn3", "Squeal" and "Horn4"), all were reported by at least half of the subjects, and for the most part onset times agreed to within a few hundred milliseconds. This would seem to be a plausibly reliably picture of the subjective content of this sound example.

In figure 5.17, this subjective description is compared to the objects produced by the system's analysis which we saw previously in section 5.1. On the whole, the system has been rather successful at producing elements that correspond to the separate events recorded by the subjects, without generating extra elements without subjective correspondence. The weft elements all line up with the various subjective "Horn" events, with the possible exception of weft 6. Note that the grouping of separate wefts into aggregate objects (such as wefts 1-4, matching Horn1) was done by hand (before gathering the experimental results) to construct resynthesized events of greater coherence. This manual grouping was performed in lieu of automatic organization of these elements into 'source' objects; the presumption is that such automatic grouping would be relatively simple to create, at least for this example, on the basis of similar onset, time support and modulation period (for the sequential grouping of wefts 3 and 4, and wefts 6 and 7). The subjective "Crash" event corresponds well to the Noise2 element, although it is questionable whether the Click1 element truly forms a part of this, or if it is in fact a spurious artifact of the analysis.

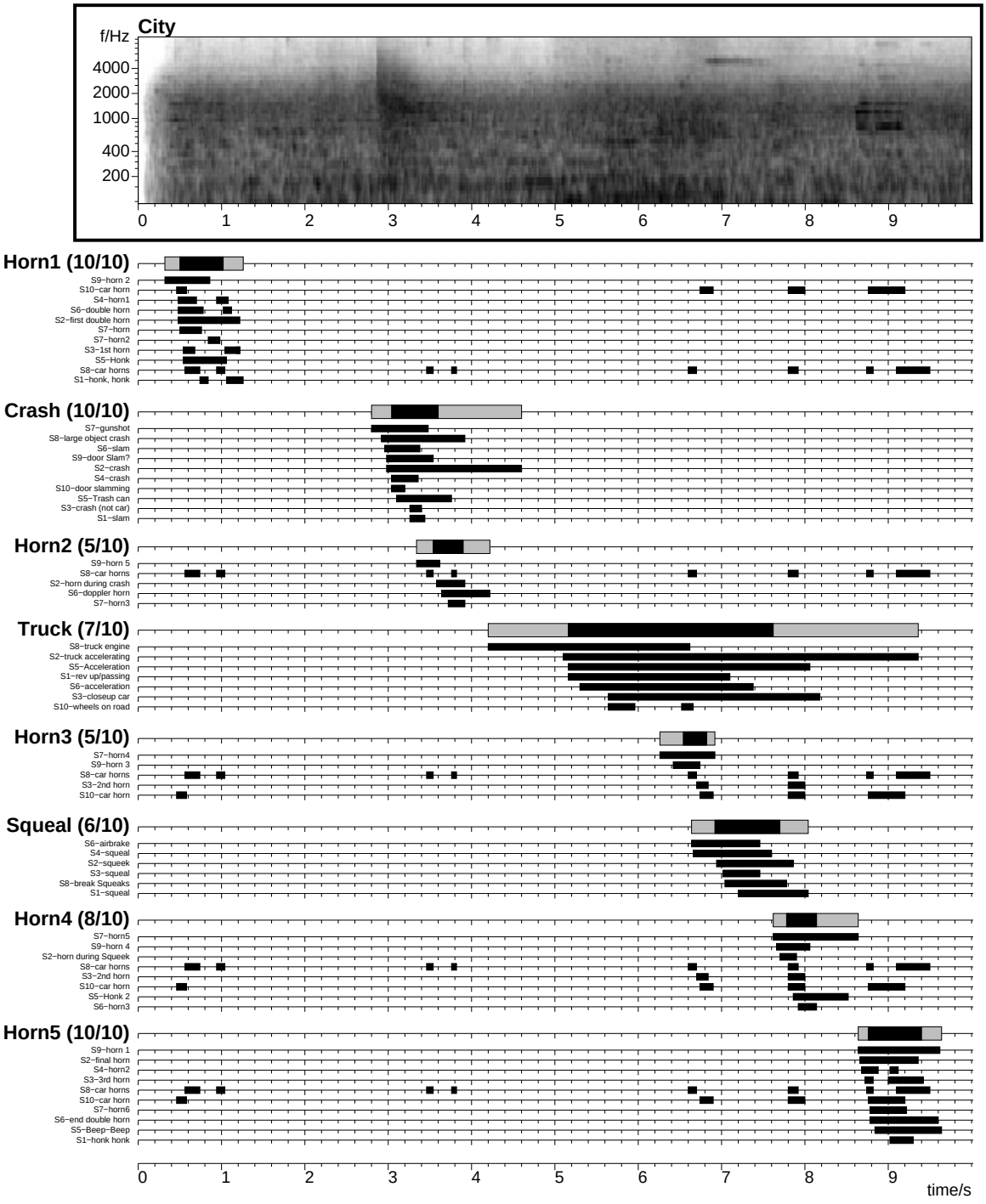

Figure 5.16: Perceived sound events for the city street ambience sound example, as recorded by listeners. The top panel shows the time-frequency energy envelope of the signal. The remaining lines are the individual response bars from part A of the experiment, manually arranged into eight groups and summarized in the bars at the top of each group. Lines related to several groups are repeated.

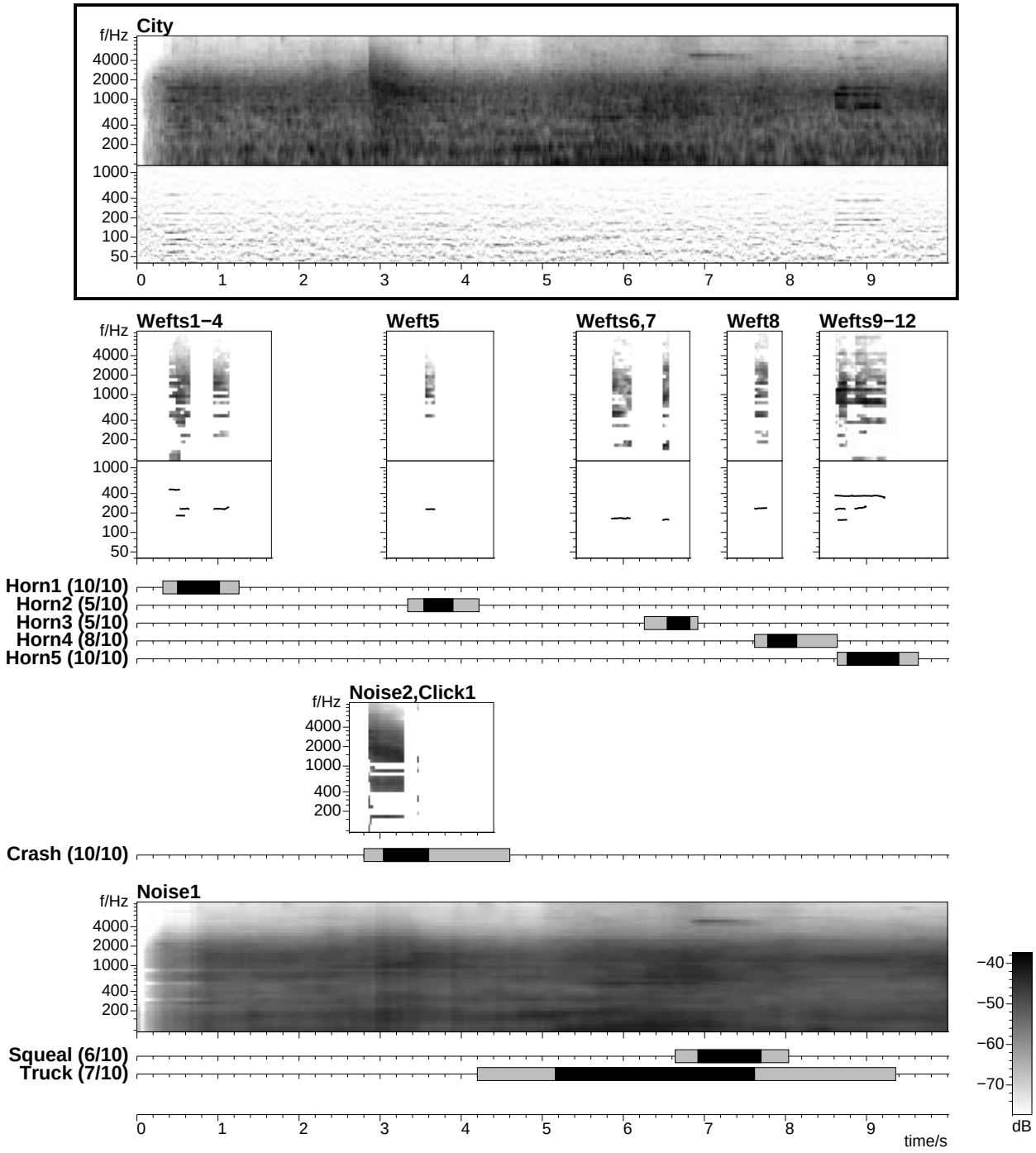

Figure 5.17: Comparison of system's and subjects' responses to the city-sound. Every object in the figure is rendered on a single time-scale labeled at the base. The intensity envelope and the periodogram of the original sound mixture are at the top. Each of the fifteen elements returned by the system is illustrated by its time-frequency envelope; wefts also show their pitch tracks. The summary response bars from figure 5.16 are displayed below the most similar element.

This example contains two glaring failures: Neither the "Truck" nor the "Squeal" subjective events were extracted as separate objects by the system; rather, their energy was merged into the background-noise element, Noise1, along with the residual traffic and street noise. These failures provide interesting insights into the system's behavior, and suggest possible solutions. The energy leading to the "Squeal" response is quite visible in the time-frequency envelope as a narrow streak at around 5 kHz starting a little before t=7 seconds. This feature was large enough and sufficiently sustained to be incorporated quite visibly into the Noise1 element, even though that element was by this time very stable and reacting only sluggishly to changes in the observed signal. Why did this rather obvious addition of energy not lead to the creation of a new element? The problem lies in the way that the 'positive-deviation norm', the parameter used to monitor the excess observed energy, is calculated as the norm of a vector of normalized energy increases in each frequency channel. Because the squeal energy is mostly concentrated in a single frequency channel, the energy increase in this channel was diluted by the absence of an increase in all the other channels to result in a norm that did not exceed the threshold for the creation of new elements. In the light of this mistake, I tried a modified positive-deviation norm which magnified the norm by a factor related to the total number of channels in which positive energy was recorded, emphasizing narrowband energy contributions over broader energy increases. Dividing by the root of the number of channels showing a positive deviation made the system recognize the squeal onset; unfortunately, this modification came too late to be included in these experimental results. (Despite the temptation of 'tweaking' the system to match the test results, all the analyses shown here are the ones created as part of the preparation for the experiments i.e. before the subjects' labeling had been collected. Every element from these analyses is illustrated; the only post-hoc interpretation has been the manual grouping of some elements to correspond to single subjective events).

A second problem highlighted by the squeal is the lack of a narrowband, sinusoidal representational element – ironic, since such elements were the representational staple of previous systems including [Ellis94] (which in fact extracted the squeal sound as one of the few things it could make sense of in this example). In the current system, periodic energy which in other systems might have been represented as sinusoids is supposed to be captured by weft elements. However, the frequency of the squeal sound at about 5 kHz was outside the 'perceptible pitch range' of 40-1280 Hz implicit in the delays sampled by the periodogram, so there was no way for this squeal to lead to a weft. Indeed, the current envelope smoothing prior to autocorrelation would hide this periodicity even if the periodogram axis were extended. (After modifying the positive deviation norm, the system created a noise element with energy in this single channel to explain the energy – a reasonable, but perceptually unsatisfactory, compromise). Clearly some kind of high-frequency narrowband element, differentiated from noise on the basis of its unfluctuating energy profile, is necessary to account for this and similar sounds, something to be considered for future system enhancements.

The second perceptual event not separated by the system is the "Truck" sound, a low-frequency rumble appearing to be the acceleration of a large automotive engine. This not obvious in the time-frequency intensity display, probably because the majority of its energy is below the 100 Hz lower limit of the frequency axis. Unlike the limited spectral view afforded the system by its initial filterbank, the sound played to the subjects had no low-frequency cutoff. Recall the 100 Hz limit was employed to avoid numerical problems

with very narrow, low-frequency filters, justified by the argument that there would be few situations where energy below this limit would be critical for the recognition of separate objects; unfortunately, the city-sound appears to contain such a situation! Extending the frequency axis down to 50 Hz or beyond might reveal enough excess energy to trigger the creation of an element for the truck sound. The correct element for such a sound is not immediately obvious. It should probably be a weft, since the periodicity of the rumble is an important part of its character. (Vague suggestions of a rising periodicity are discernible at the lower edge of the periodogram display around t = 5 s). However, wefts have not been tested with noisy, low-frequency signals of this kind. The combination of a low-frequency weft and a noise cloud in the higher frequencies might give the best perceptual match.

The results of parts B and C of the experiment, concerning subjective ratings of resyntheses for the "Horn1", "Crash" and "Horn5" events, are discussed along with the resyntheses from the other examples in section 5.3.6 below.

5.3.3 "Construction" sound example

The second sound example was a similar dense, ambient sound, taken from the same professional sound-effects collection [Aware93]. The sound is described as 'construction site ambience', and contains a diverse mixture of machinery, knocking, hammering and voices. The subjective responses to this example are illustrated below its intensity envelope in figure 5.18.

Again, the level of consensus between the subjects' responses is good. The first group, "Saw", corresponds to the rather prominent noise of a motorized saw cutting through a plank (or so it sounds). This energy is visible covering a wide frequency band in the upper half of the intensity envelope. The remaining events correspond mainly to transients; interestingly, both "Wood hit" and "Metal hit" have been reported with multiple onset times by the majority of subjects who noted them. In the spectrogram, successive onsets are visible; evidently the resemblance between the nearby onsets led the subjects to report them in groups rather than as separate events.

The one remaining non-transient event is the "Voice" object, visible as smudges of energy at around f = 800 Hz at times t = 8.1 s and 8.5 s. This is a clear percept, although the voice sounds a little "weird", to use the description given by one subject. Interestingly, the pitch of the voice is perceived as much lower than the 800 Hz band in which its visible energy is concentrated.

The perceptual event summary bars are repeated on figure 5.19 alongside the elements extracted by the system for this example. The analysis is largely as expected: Noise1 models the background machine ambience, with a second sustained noise element, Noise2, providing the extra energy from the "Saw" event, whose time support it nicely matches. Of the eight click events extracted by the system, there is an almost perfect correspondence with the subjective responses – unexpected, since the example contains a large number of audible transients of varying prominence that required the subjects to make a somewhat arbitrary choice of the ones to report. The most suspect element is Click6, which probably wasn't being indicated by the "Clink1" response. Note also that the "Metal hit" subjective event, which was indicated as two distinct onsets in five of its eight component responses, corresponds to a pair of elements, Clicks2,3, which align credibly with the distinct onsets reported. However, the subsidiary subjective onsets indicated for "Wood hit" have not been extracted; the entire event corresponds to the single Click1 element.

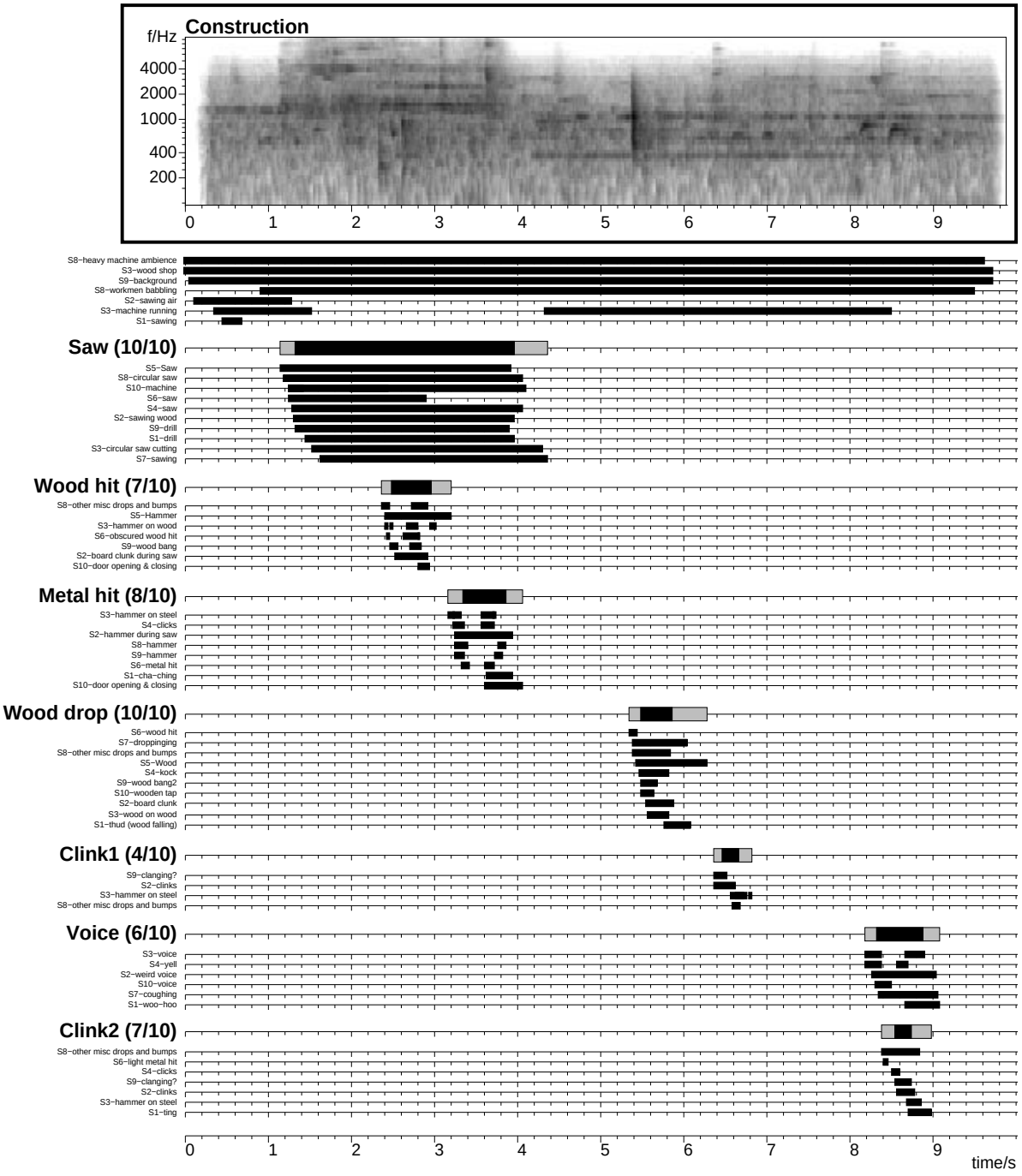

Figure 5.18: Responses from part A of the experiment for the "Construction" sound. The six response lines at the top, which appear to refer to background noise, have not been formed into a specific group.

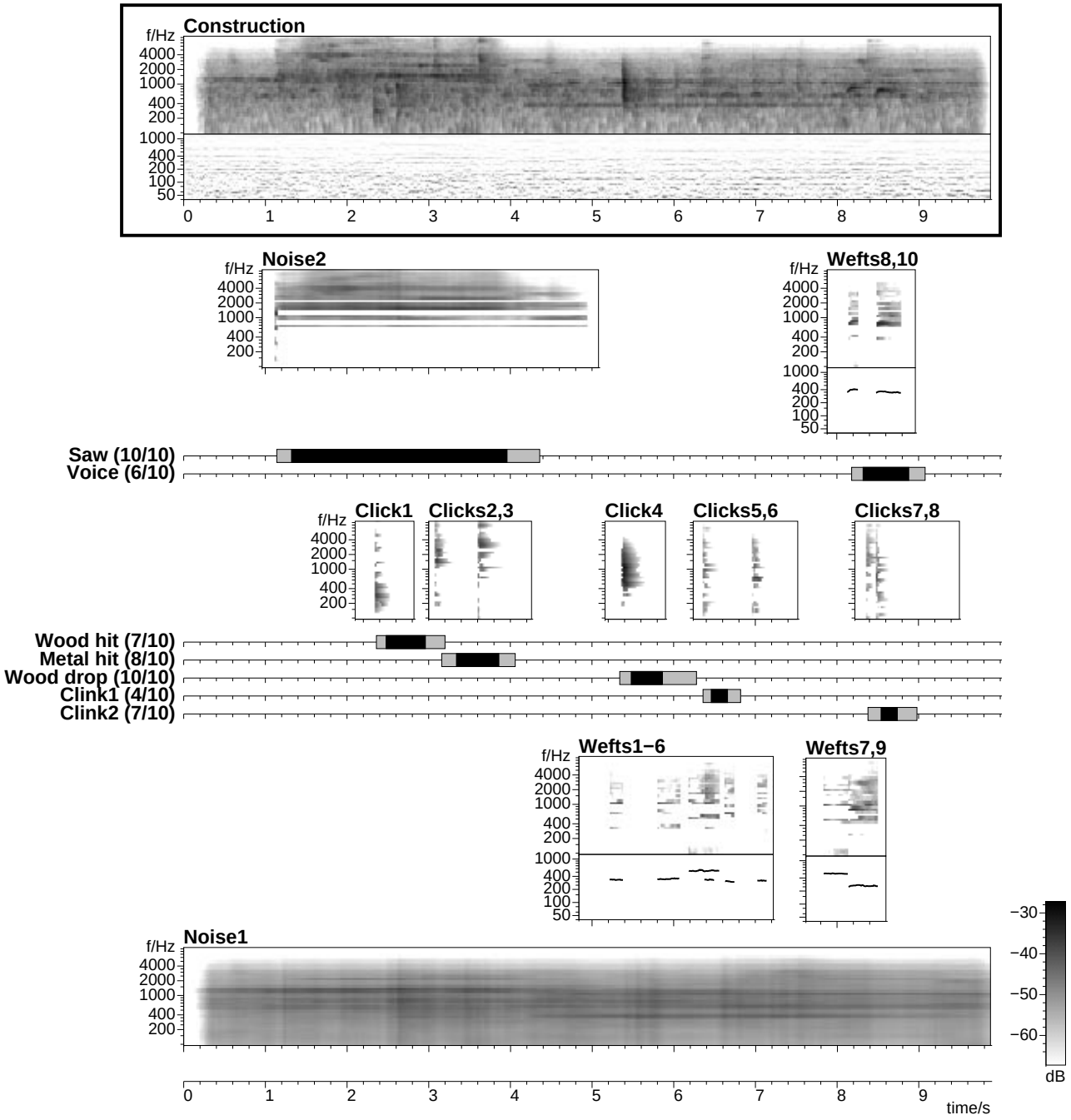

Figure 5.19: The system's analysis of the "Construction" example along with the summary subjective events. The time-frequency intensity envelope and periodogram of the original sound are illustrated at the top of the figure. The remaining objects are the elements generated by the system and the summary response lines, all drawn on a common left-to-right time scale given by the scale at the bottom.

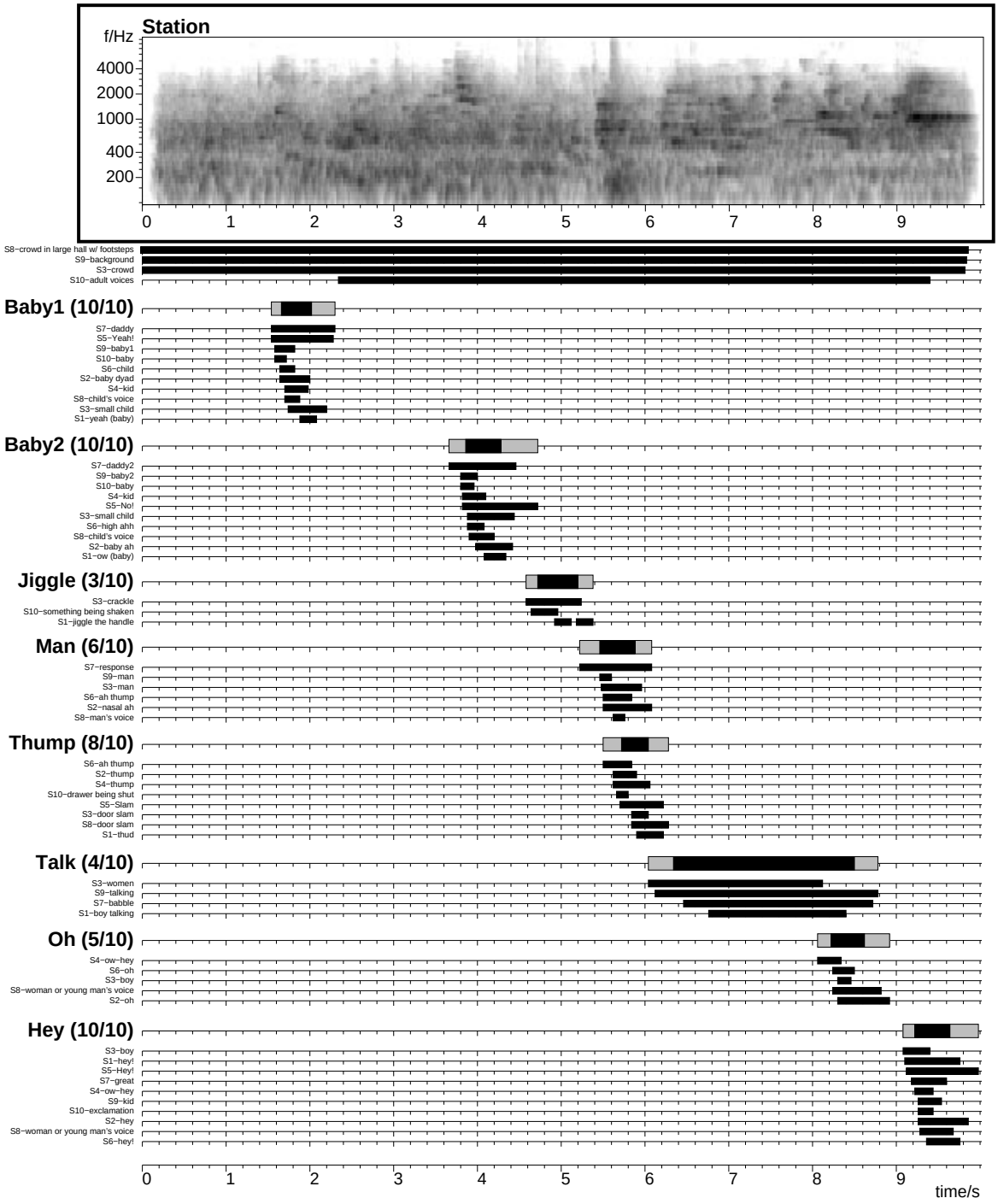

Figure 5.20: Responses from experiment part A to the "Station" sound example. The four responses at the top indicating almost the entire sound have been excluded from a summary group.

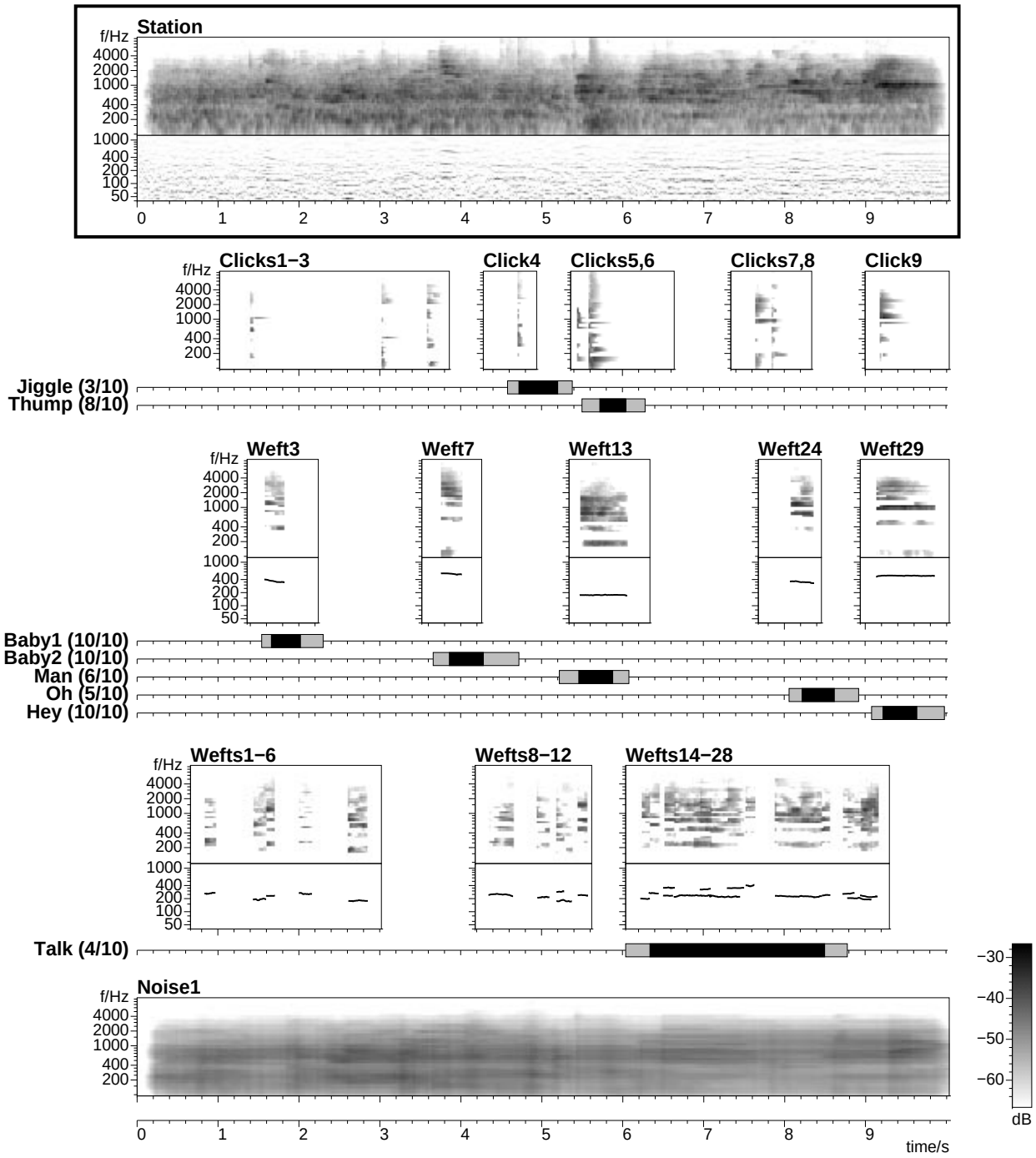

Figure 5.21: System's analysis of the "Station" sound example compared to the summary subjective events from 5.20.

The system created quite a number of weft elements in the region after the initial saw noise had ceased. Of these, Wefts8,10 align well with the "Voice" event, which was indeed indicated as having two parts by some subjects (and appears in two pieces in our inspection of the time-frequency intensity envelope). The remaining wefts do not have any subject-reported correspondence, although it is likely that they are intermittently detecting the periodic components of the machine indicated by some subjects in their description of the background noise. It should be noted that there is temporal overlap, as well as pitch-track proximity, between the 'background' Wefts7,9, and Wefts8,10 which were picked out as "Voice" event. This piece of manual selection is one of the hardest to defend; presumably, a more robust weft extraction might have distinguished between them by making Wefts7,9 part of an ongoing background tonal structure that excludes Wefts8,10

The system's resyntheses tested for this example were from Noise2 (corresponding to the subjective event "Saw"), Click4 (corresponding to "Wood drop") and Wefts8,10 (corresponding to "Voice"). The first two of these match events reported by all subjects; "Voice" was reported by just six of the ten subjects, although in part B of the experiment it was rejected as unrecognizable by only one subject; the other three rated it as "something I recognize but did not previously note". These are discussed in section 5.3.6.

5.3.4 "Station" sound example

The third sound example was another fragment from the sound-effects collection. It is nominally "train station ambience", consisting of babble, footsteps etc. in a highly reverberant environment. The most prominent events in the ten-second excerpt are a few loud vocal exclamations and one or two transient 'thumps', as of door slams or dropped bags. The intensity envelope is shown in figure 5.20 along with the subject responses.

There is perhaps a little more variation in the pattern of responses to this example, again reflecting the choice each subject had to make between reported events and events that were perceived but not adequately salient to report. Three events, "Baby1", "Baby2" and "Hey", were noted by all subjects. The remainder were noted by at least half of the subjects, except for "Talk", for which the subjects' labels suggests that it is an aggregation of individually-indistinct events whose combination was deemed notable, and "Jiggle", whose identity is not clear, although it probably results from the energy visible above 4 kHz at around t = 4.6 s.

The elements generated by the system are illustrated in figure 5.21 in comparison to the summary subjective events. As with the previous two ambient examples, Noise1 has been created to represent the steady background noise with a fairly static noise envelope. Of the nine click elements extracted, Clicks5,6 align well with the "Thump" event, and Click4 appears to have picked up at least a part of the mysterious "Jiggle" event. The remainder of clicks have no subjective event correspondents.

The example gave rise to 29 weft elements, presumably picking up on brief fragments of periodic modulation from the background babble of mixed voices. Of the wefts, the most energetic correspond well to the distinct voice-events reported by the subjects, "Baby1", "Baby2", "Man", "Oh" and "Hey", as shown. Of the remaining wefts, there is a considerable concentration starting at around t = 6.2 s, which lines up well with the composite "Talk" subjective event. The other wefts have no correspondence to the subjective responses, other than forming part of the background crowd noted by some subjects.

The resyntheses used in experiment parts B and C for this example were based on Weft7 (“Baby2”), Weft29 (“Hey”) and Click6 (“Thump”) and are discussed in section 5.3.6 along with the subjective evaluation of the resyntheses from the other examples.

5.3.5 The double-voice example

The final sound example was quite different from the others: It was a short fragment of a mixture of male and female speech, taken from the sound examples made available by [Brown92] (who refers to it as v3n7), and originally constructed for [Cooke91]. The purpose of including this example was to provide an illustration of how this kind of experiment can be used to compare between different models of auditory scene analysis, as long as they produce resyntheses. This example was chosen from among the sound examples Guy Brown made available for his thesis, meaning that I could run the current system on his mixture and then gather subjective ratings of the separations generated by both systems. Thus the main point of using this example was to obtain quality ratings for resyntheses derived from it; there was not much room for perceptual ambiguity in the labeling of the sound (at least in terms of the sources present – figuring out what they were saying is harder!), so the results of experiment part A are not particularly interesting. It was important to conduct a part A for this example, however, in order to familiarize the subjects with the sound, and to maintain the protocol with which the subjects had become familiar and which would lead subsequently to rating judgments. For completeness, the results of part A are presented in figure 5.22; however, there is no need to compare these to the system’s analysis (previously shown in figure 5.9) since the correspondence is self-evident.

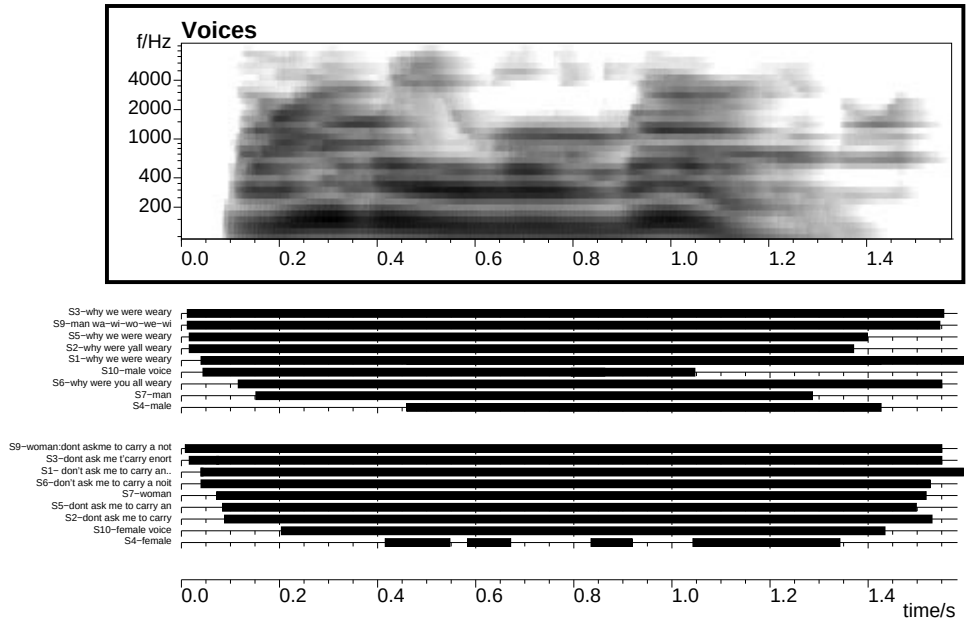

Figure 5.22: Subject responses for the “Voices” sound example as gathered through experiment part A.

5.3.6 Experiment part B: Rating of resyntheses

So far we have only considered the subjective labeling of the different sounds perceived in the mixtures in comparison to the objects extracted by the model in terms of their visual, abstract representations. But parts B and C of the experiment made direct tests of the listeners' judgments of the system's output. These results are now summarized.

For each of the three ambient sound examples ("City", "Construction" and "Station"), three of the objects derived by the system were chosen for presentation to the listeners. (In a couple of cases, several weft objects were combined into a single resynthesis example as a manual simulation of higher-level grouping by proximity or similarity, as noted above). These resyntheses were presented to the listeners in part B of the experiment for each original sound. In addition, a further trial of part B consisted of a combined resynthesis of every object separated by the system. Getting the subjects' ratings of this 'complete resynthesis' obtained a kind of upper-bound on the resynthesis ratings, since mutual masking, contextual cues to object identity, and the neutralization of errors in dividing or merging object energy, should make the perceived quality of a resynthesis improve as more and more elements are added.

The subjective ratings assigned to the resynthesized objects are presented in the table 5.1, which shows the complete summary results for the part B trials. There were altogether five different original sounds in the experiment, comprising a training trial, the three ambient examples and the final "Voices" example. Each trial of part B presented a single resynthesis; the subjects made a labeling for the resynthesis in terms of the events they had previously reported in part A, and gave a subjective rating of the resynthesis quality. In the table, the first column is the name of the base sound example. The second column specifies the particular resynthesis, either as the name of the analysis object(s) involved (taken from the labels in figures 5.17, 5.19, and 5.21), or as "(all)" to indicate the complete resynthesis composed of all analysis objects. (The training resynthesis, which came from an unrelated system, is called "can2". The "Voices" example includes the "brnfi" resynthesis from the system of [Brown92], as discussed below).

The third column gives the summary subjective event label best matching the resynthesis object, according to the correspondences drawn in figures 5.17, 5.19, and 5.21. The next two columns give the counts of 'right' and 'wrong' responses out of the total number of valid responses collected (some trials were lost through errors). These data come from the check-boxes to the left of the part B response screen shown in figure 5.13. A response is counted as "right" if the subject indicated that the resynthesis contained their contribution to the group of subjective event responses named in column three. If the subject named a new event (indicating that they recognized the resynthesis but had not labeled it in part A), the response counted as "right" if the name supplied suggested the intended correspondence (the analysis of these scores thus include some subjective judgments on the part of the investigator). A response is "wrong" if the subject described the resynthesis as failing to contain the intended sound, or if the sound was indicated to contain sounds other than the intended sound. Thus a single trial could contribute both "right" and "wrong" scores if the subject heard both the intended and unintended events in the resynthesis. For the complete resynthesis "(all)" examples, the trial contributed to "right" if the subject

indicated that it contained most events, but also to "wrong" if certain labeled events were pointedly left out of the set judged to be present.

The remaining columns concern the subjective rating scores. Recall that in experiment part B, the subjects had a slider labeled with "unrecognizable", "distorted", "similar" and "identical" which they adjusted to indicate their assessment of the resynthesis quality. The rating was actually recorded as a integer between 0 and 100, where the labels aligned with 5, 35, 65 and 95 respectively. The default position of the slider was at zero, and thus if the subject refused (or forgot) to rate the resynthesis, a zero was recorded. The "#rat." column shows the number of nonzero responses obtained for each example; responses of zero were excluded from the analysis, but these formed less than 5% of the dataset.

Example	Resynth.	Event	Right	Wrong	#rat.	Rating	Norm. rat.
Training	can2	"Can"	10/10	0/10	10	75.2 (17.4)	1.43 (0.91)
City	Noise2	"Crash"	9/10	1/10	9	40.1 (12.7)	–0.58 (0.81)
	Wefts1-4	"Horn1"	9/10	2/10	10	50.5 (22.6)	0.11 (0.97)
	Wefts9-12	"Horn5"	10/10	0/10	9	46.0 (18.4)	–0.12 (0.63)
	(all)		10/10	4/10	9	56.2 (18.0)	0.51 (0.95)
Construct.	Noise2	"Saw"	9/10	1/10	9	44.7 (14.5)	–0.20 (0.92)
	Click4	"Wood drop"	10/10	0/10	10	60.5 (8.9)	0.60 (0.45)
	Wefts8,10	"Voice"	8/10	3/10	9	42.3 (18.1)	–0.32 (0.90)
	(all)		9/9	2/9	9	59.3 (8.0)	0.63 (0.43)
Station	Weft7	"Baby2"	9/10	1/10	10	42.4 (13.9)	–0.31 (0.79)
	Click6	"Thump"	9/10	1/10	9	48.8 (15.6)	0.08 (0.80)
	Weft29	"Hey"	7/10	3/10	8	32.0 (9.7)	–0.90 (0.56)
	(all)		10/10	4/10	10	48.6 (14.6)	–0.08 (0.71)
Voices	Weft1	"Male"	9/9	1/9	9	43.9 (15.5)	–0.16 (0.76)
	Wefts2-4	"Female"	9/9	1/9	9	29.8 (14.9)	–0.95 (0.68)
	brnfi	"Male"	9/9	1/9	9	36.9 (17.9)	–0.65 (0.79)
	(all)		9/9	0/9	9	60.9 (23.9)	0.63 (1.10)
Totals			155/165	25/165	157	48.5 (19.7)	0.00 (1.00)

Table 5.1: Summary results for subjective ratings of resyntheses from part B of the experiment: right/wrong counts, average rating scores and average normalized ratings.

The numbers in the "Rating" column are the average rating score; the standard deviation for each set is shown in parentheses. The results are rather bunched around the middle of the scale, but biased towards the lower end, with only the fake training example exceeding "similar" (75.2 compared to 65), and a couple of the ropier weft examples falling below "distorted" (29.8 and 32.0 compared to 35). The final column, "Norm. rat.", gives the means and standard deviations of the ratings after normalization for individual subjects, to provide a more precise comparative score between examples: Since ratings generated by a single subject presumably are more internally consistent than the ratings reported by different subjects, all the ratings reported by a single subject were pooled to find the mean and variance of that subject's rating responses. The normalized ratings are obtained from the raw

ratings for a particular example by subtracting the mean rating for that subject and dividing by the subject's overall rating standard deviation. These numbers are thus distributed around zero, with a normalized rating of 1.0 indicating that subjects rated this example one standard deviation better than their average responses; negative normalized ratings indicate examples judged to be worse than average. The unnormalized average ratings address the overall absolute subjective rating of the example, whereas the normalized ratings give a more precise aggregation of each subject's relative judgment of the quality of a particular example compared to the others presented in the experiment.

Right/Wrong results for ambient examples

Looking first at the "Right" scores, we see that on the whole the subjects were able to identify the resyntheses as corresponding to the intended portion of the original sound. Of the nine isolated-event resyntheses from the ambient examples (i.e. the first three rows for City, Construction and Station), only two were unidentified by more than one subject – the "Voice" event in the Construction example, and the Station's final "Hey" event. Both of these weft elements suffered spectral distortion owing to their low level relative to the background noise; listening to the "Hey" resynthesis reveals that it has been extended in comparison to what I perceive in the mixture, apparently because the weft analysis has tracked the voice into its reverberant tail. (The Station example was highly reverberant, something that listeners accommodate very effectively; the system made no particular provision for such situations).

The pattern of "wrong" responses is similar, with "Voice" and "Hey" doing the worst (one subject labeled "Hey" as part of their all-encompassing background noise label, which was counted as both "right" and "wrong"). The "(all)" resyntheses for the ambient examples also pick up a number of "wrong" scores, mainly from labeled events that the subject judged absent from the resynthesis. These omissions tended to be for smaller, less prominent objects.

Rating score results for ambient examples

As noted above, the rating scores are bunched fairly closely around the middle of the scale. As anticipated, the ratings of the full resyntheses are on the higher side, although no better than the best individual element for the Construction and Station examples. The "Hey" resynthesis, with its reverberation-related problems, stands out as the lowest-rated among the ambient examples, undercut only by the Voices' "Female" discussed below. The best-rated resynthesis is the "Wood drop" from the Construction example; for some reason, this achieved a notably better match to the perceived transient sound than the Station's "Thump" event, also modeled by a click element with energy concentrated in the mid-to-low frequencies. The "Crash" event from the City example has some perceptual similarity to these transients, but, as discussed, its relatively slow decay caused it to be modeled as a noise element rather than a click; the unfortunate consequence of this was that the resynthesis had an abrupt termination, visible in figure 5.11, rather than dying away smoothly to silence which would have been more perceptually agreeable. Also, both the "Crash" and "Thump" resyntheses appear to have provided too little low-frequency energy since their shifted-down (in frequency) versions were preferred in part C, discussed below.

The Voices example

The final Voices example was included mainly to demonstrate the way in which different analysis systems may be compared with this kind of

experiment. The original mixture of male and female voices was used in the system of [Brown92], and his separation of the male voice was one of the resyntheses presented to the listeners, "brnfi". The current system's analysis, illustrated in figure 5.9, produced resyntheses for both the male and female voices; however, the difficulties in extracting non-vowel energy discussed in section 5.1.3 meant that the "Female" resynthesis was interrupted by gaps where the sibilants and consonants should have been, leading to a highly distorted resynthesis which was accordingly rated with the lowest overall score. The real interest in this example comes from comparing the two versions of the "Male" source – Weft1 from the current system, and "brnfi" from Brown's system. In the event, the difference is quite small, with the normalized ratings showing Weft1 managing about 0.5 sd's better than "brnfi". No meaningful conclusions can be drawn from a comparison based on a single example; the real significance of this result is that the comparison can be made at all. Here are two different models with utterly unrelated resynthesis approaches being subjected to a quantitative comparison, something that the system-specific assessment methods of previous projects could not provide, yet a class of measurement that must become increasingly significant in the field.

Ratings pooled by element type

Since the resyntheses fall into several distinct classes according to the types of element from which they are derived, it is of interest to pool ratings across resyntheses in the same class. These results are presented in table 5.2, which arrange the data of table 5.1 by resynthesis class rather than by base example.

Resynthesis class	**#rat.**	**Rating**		**Norm. rat.**	
Noise elements ("Crash", "Saw")	18	42.4	(13.8)	–0.39	(0.89)
Click elements ("Wood drop", "Thump")	19	54.9	(13.8)	0.35	(0.69)
Weft elements (7 examples)	64	41.3	(18.1)	–0.36	(0.86)
Total for individual elements	101	44.1	(17.5)	–0.23	(0.88)
Full resyntheses (4 examples)	37	56.1	(17.7)	0.41	(0.89)
Total for all system resyntheses	138	47.3	(18.3)	–0.06	(0.93)

Table 5.2: Rating results pooled by resynthesis type. Only the resyntheses produced by the current system are included (i.e. excluding "can2" and "brnfi").

Although this pooling appears to show that the click elements achieve the best resynthesis ratings, with noise and weft elements roughly on a par, these results cannot be given too much authority since there are only two examples each of the noise and click elements. What is more clearly shown is that the pooled results of the eleven isolated resyntheses (three for each ambient example plus two from Voices) rate noticeably worse than the aggregate rating for the full resyntheses (of which they are a part). This could be interpreted as confirmation that presenting the resynthesized elements in a dense context rather than in complete isolation hides their flaws. This is to be expected; the full resyntheses include the background noise element, which effectively provides a broadband masking noise typically only a few dB below the other resynthesized elements. In the full resyntheses, the experiment can take advantage of the prediction and

inference mechanisms at work in the listener to improve the perceived quality of the elements!

5.3.7 Experiment part C: Ranking of resynthesis versions

Each of the nine isolated-event resyntheses from the ambient examples (but not the complete resyntheses) was subjected to four systematic distortions, making a total of five 'versions' to be ranked in experiment part C. The intention of this part of the experiment was to provide evidence that the resyntheses generated by the system were better than other, similar sounds, or, failing that, to give pointers towards how the resyntheses might be improved. The subjects were presented all five versions as movable 'buttons' on screen which they had to arrange in declining order of similarity to an event perceived in the original sound; the data were recorded as rank orders from zero to four by sorting the final y-ordinates of the buttons on the screen. (In a couple of trials, subjects managed to move two buttons to exactly the same height, presumably indicating no preference between them; this was recorded as the average of the two rank scores e.g. if two lowest objects were at the same height, they were both ranked as 3.5). Subjects were not able to skip this screen, meaning that in the small number of trials where a subject had failed to identify the correspondence of a resynthesis in part B, they were left to rank the versions according to more abstract preferences. The initial order of the buttons was randomized, so failing to exclude trials that might have been ignored by the subject does not present a systematic bias. In retrospect, it would have been better to provide a button that allowed the subject to indicate "I don't know what this sound is and I can't rank its versions."

Generation of distorted versions

In order to conduct part C of the experiment it was necessary to define several 'distortion operators' to apply to the objects. The specific goal of this experiment was to investigate if the system's resyntheses lay in a local quality optimum in the parameter space constituted by each sound element model. Consequently, the distortions were defined to operate upon the parameters of an element prior to resynthesis, and the versions were generated by feeding the distorted parameters to the same resynthesis procedures used for the 'best' examples.

The parameter space is unthinkably large, and the choice of distortion operators similarly huge; for want of better alternatives, four distortions were defined representing one step either side of the undistorted parameterization on two dimensions, nominally 'frequency' and 'smoothness'. The precise definition of these distortions varied somewhat depending on the element type to which they were being applied, but in general applied to the time-frequency energy envelope that the element was intended to reproduce:

- A 'smoothed' version applied raised-cosine smoothing windows along both axes of time-frequency envelope generated by the element. The time-window had a width of 11 timesteps (approximately 50 ms) and five frequency bins (meaning that each channel was mixed with its spectral neighbors to two bins on each side). As with all the distortion operations, these windows were chosen to provide small perceptual modifications that were none-the-less reliably distinguishable from the original.
- The complement of the smoothed version (stepping the other way on the notional smoothness axis) was a 'sharpened' version: The smoothing window was normalized to have unit gain for a constant signal, so

subtracting the smoothed time-frequency envelope from the undistorted original separated an 'unsmooth' component (in the sense of the 'unsmooth masking' operation of image processing). This component was increased by 50% and added back to the smoothed envelope to generate a new envelope with the same local-average properties as the original profile, but with exaggerated deviations about the local average.

- The frequency dimension was more simply defined: A 'shifted-up' version simply displaced the entire time-frequency envelope up by one frequency channel; the 'shifted-down' complement was displaced down one bin from the original.
- Weft elements include the additional parameter of their period contour, used to construct a variable-period pulse train to excite their spectral envelope. This contour was similarly smoothed and sharpened (to reduce or enhance deviations about the locally-average pitch), and shifted up and down in frequency. For the frequency shifts, a factor of 3% (half a semitone) was used; although the one-channel shift of the envelope corresponded to a 12% frequency (i.e. six channel per octave, where the sixth root of two is approximately 1.12), a pitch shift of this magnitude was very obvious, so a smaller factor was used to make the distorted versions less obviously different from the original.

Figure 5.23 shows examples of an element envelope (in this case a weft) and its four distorted derivative versions.

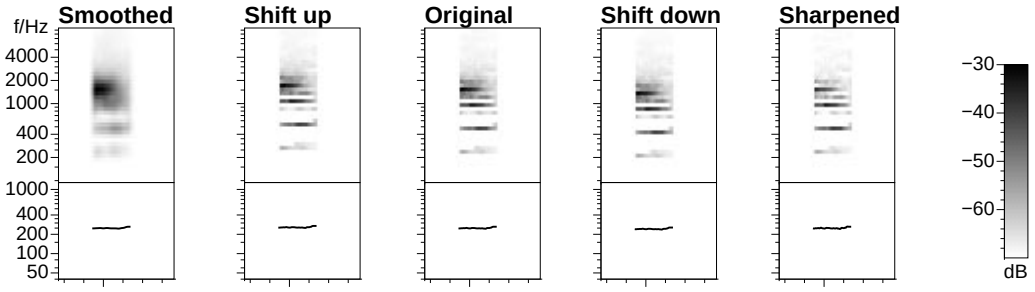

Figure 5.23: Illustration of 'distorted' time-frequency envelopes and period tracks. From left to right, the versions are: 'smoothed', 'shifted-up', original, 'shifted-down' and 'sharpened'.

Ranking of distorted versions

The results of ranking the distorted versions of the nine resynthesized objects used for part C are summarized in table 5.3: for each example, the versions are listed from left to right in declining order along with their average rank (where the 'best' version has a rank of zero, and the bottom version scored 4). Each ranking score is the average over ten results, one from each subject.

Example	Resynth.	Event	Version preference (avg. rank)				
City	Noise2	"Crash"	sd (1.0)	sh (1.2)	**or** (1.6)	sm (2.5)	su (3.7)
	Wefts1-4	"Horn1"	**or** (0.6)	sh (0.8)	sm (2.1)	su (3.1)	sd (3.4)
	Wefts9-12	"Horn5"	sh (1.0)	**or** (1.2)	sm (1.8)	sd (2.8)	su (3.2)
Construct.	Noise2	"Saw"	**or** (1.0)	sh (1.1)	sm (2.1)	su (2.2)	sd (3.6)
	Click4	"Wood drop"	sh (0.8)	**or** (1.6)	sd (2.0)	sm (2.1)	su (3.5)
	Wefts8,10	"Voice"	**or** (1.0)	sh (1.6)	sm (1.8)	su (2.6)	sd (2.8)
Station	Weft7	"Baby2"	**or** (0.5)	sh (1.9)	sh (1.9)	su (2.5)	sd (3.2)
	Click6	"Thump"	sd (1.1)	sh (1.3)	**or** (1.5)	sm (2.6)	su (3.5)
	Weft29	"Hey"	sh (0.6)	**or** (0.7)	sm (1.8)	sd (3.1)	su (3.7)

Table 5.3: Ranking of resynthesis versions for the nine isolated-event resyntheses from the ambient examples. Each row shows the average ranking of the five versions, sorted in descending order; thus, the first version column contains the most preferred, and the last column has the version judged least similar. Versions are indicated by two-letter codes; 'or' = original (system's parameters, shown in bold), "sm" = smoothed time-frequency envelope, "sh" = 'sharpened' envelope, "su" = shifted up in frequency, "sd" = shifted down.

The first thing to note is that the versions have been ranked with some consistency; if subjects had ordered them at random, we would expect the average ranks to be about the same for all versions; instead, the ordering is sometimes quite pronounced. The second thing to note is that, unfortunately, the initial hypothesis that the system's original versions would be the 'most similar', is not supported. The "or" version only comes out on top in four out of nine trials. In three cases it is in second place, and in each of the remaining two resyntheses there are two versions whose average rank is better than that of the original. When the original is beaten out of first place, it is always ranked below the 'sharpened' version ("sh"), but in the two cases where the original is third, the 'shifted-down' version has been ranked first, ahead of both "or" and "sh". These two resyntheses, "Crash" from the City sound, and "Thump" from the Station example, both correspond to transients in the original sound that might be expected to contain significant low-frequency energy. The inference from the success of the 'shift-down' version in these two cases is that the system failed to assign enough low-frequency energy to these resyntheses, possibly because the energy was hidden in a wide-variance background noise element, or perhaps because of the slower onsets in the narrower, low channels preventing the element creation logic assigning enough energy to the new element in this region. In any case, this part of the experiment has succeeded in its secondary goal, to indicate specific areas for system improvement.

Rankings pooled by element type

As with the ratings, it is informative to look at the average ranking scores for the different versions averaged across each element type (noise, clicks or wefts), especially since the distortion operator varied according to the element upon which it was operating. These results are presented in table 5.4; rather than sorting the versions by rank, the average ranking of each version is shown in a separate column:

Resynthesis class	# resyn.	Avg. rank: or	sh	sm	sd	su
Noise elements	2	1.3	1.1	2.3	2.3	3.0
Click elements	2	1.6	1.1	2.4	1.6	3.5
Weft elements	5	0.8	1.2	1.9	3.1	3.0
Total	9	1.1	1.1	2.1	2.6	3.1

Table 5.4: Ranking of resynthesis versions pooled by the type of the element upon which the original resynthesis is based. Of the nine resyntheses for which rankings were gathered, there were two each of noise and click elements, and the remaining five were wefts. Each resynthesis contributes ten ranking scores to each version.

Pooling the pairs of noise and click element resyntheses leaves the 'sharpened' element ranked ahead of the original in both cases; for the click elements, even the 'shifted-down' version matches the ranking of the original. One explanation of the comparative success of 'sharpened' versions of transient elements is that the sharpening distortion will presumably have emphasized the initial onset, which may have been blurred by the analysis, again indicating a specific direction for system improvement.

In all cases, the 'shifted-up' version has been ranked very low, although, by contrast with the click elements, 'shifted-down' is ranked even worse for weft elements. The additional effect on the pitch contour has presumably added to the disinclination towards frequency-shifted versions for the weft elements. Smoothed versions are rated approximately in the middle in all cases. When all elements are pooled together, the original is at least ranked first, although it has to share that position with the sharpened version, emphasizing the salutary results of this part of the experiment, that there is clearly room for improvement in parameter extraction and resynthesis of the elements.

5.4 Summary of results

In this chapter we have seen how the system handles real sounds, and used the subjective responses of listeners to assess its output. The listening tests turned out to be very successful, insofar as they satisfied several different objectives. From the most general perspective, the tests gathered some 'ground truth' regarding the perception of distinct events in dense ambient sound mixtures; while these results were hardly startling, it was interesting to see the level of agreement reached between the subjects in the number and identity of objects in what were rather contextless and ambiguous sound fragments. The perception of complex, real sound scenes of this type is necessarily a difficult thing to measure, yet audition in these circumstances is sufficiently different from the reductionist scenarios of conventional psychoacoustics to warrant separate investigation.

For the ambient sound examples the system's thresholds were manually tuned to produce a reasonable output according to my judgments. Even so, the extent to which the objects generated by the system corresponded to the subjective events reported by the listeners was very encouraging, with rather few cases in which the system identified spurious events or the subjects named events not recorded by the system. Where discrepancies did occur, they often pointed to specific remedies through system enhancements.

Resynthesized objects were correctly identified with their corresponding event in over 90% of trials; however, the listener's responses to resyntheses of particular system-identified objects showed quite emphatically that the overall quality of analysis and resynthesis has much room for improvement. The overall average rating (including the forgiving 'complete' resyntheses) of 47.3 lay closer to 'distorted' than to 'similar'. The experiments themselves were a success in terms of generating a reasonably consistent pattern of responses to lend some confidence to the overall ratings and the judgments pertaining to the different types of resynthesis.

A similar picture emerges from the investigation of 'distorted' versions of particular resyntheses; the experiments succeeded in collecting subjective responses that gave consistent and quite strong results. These results were that resyntheses in various classes had definite and specific weaknesses, but, armed with these results, the more egregious weaknesses can hopefully be cured in future developments of this work.

Chapter 6 Summary and conclusions

6.1 Summary

The previous three chapters formed the complete presentation of the prediction-driven architecture and its preliminary implementation. In this final chapter I will take a slightly broader perspective to look at the things I might have added to the system (in an ideal world without time constraints), and some of the aspects of audition I still don't know how to incorporate into this approach. The whole question of how well this may be considered a model of human hearing will be reconsidered, and finally I will present my vision of the development of the field of computational auditory scene analysis.

6.1.1 What has been presented

Before drawing conclusions, let us briefly review what this thesis has contained. I started by presenting a view of the phenomenon of auditory scene analysis and a perspective on the important problems facing computer modelers. The review of previous and related work characterized most of the larger projects in the area as 'data-driven' – forming abstract descriptions of observed sound based purely on the detected low-level features, without allowing for the influence of expectations or other context. In motivating a hypothesis-led, prediction-driven approach, I highlighted various ubiquitous auditory phenomena, such as robust interpretation of masked signals and resolution of ambiguity, which cannot be accomplished without a top-down component to the architecture. I then proposed a solution to this shortcoming, a system consisting of an internal world model, comprising hierarchic explanations of generic sound elements, which is continuously reconciled to the observed sound cues.

The implementation did not manage to encompass all aspects of the proposed approach, and in particular the relative paucity of explanatory abstractions meant that some of the more interesting possibilities of reconstruction of corrupt sound could not be pursued. The foundation of the system that was built consisted of a vocabulary of three types of generic sound elements, suitable for the representation of noisy, transient and tonal sounds. These were generated by a blackboard-based incremental reconciliation engine whose goal was to account for the 'indispensable' features extracted from the observed sound by the front-end. These features were overall time-frequency intensity envelope, indicating the presence and spectral location of acoustic energy, and the periodogram, a summary of the short-time autocorrelation of the envelope in each frequency channel, which reveals the presence of signal periodicity generally experienced as pitch.

This system exhibited a number of interesting behaviors, including the ability to model overlapping sounds and pursue alternative hypotheses in the face of ambiguity, and proved successful in the analysis of examples of the complex 'ambient' sound scenes that originally motivated this project. The success was confirmed in the most direct manner possible, by comparison with the

analyses of real listeners, as revealed by psychoacoustic testing. While listeners' ratings of the quality of the system's resyntheses indicated that there is much room for improvement, the underlying successful correspondence between objects identified by the listeners and objects extracted by the system presents strong evidence for the soundness of the approach, at least within this domain of dense, noisy sound mixtures.

Some care was taken to ensure that the subjective testing protocol developed for this project would be applicable to other automatic sound organization systems, since it relies neither on special properties of the system (other than its ability to generate resyntheses, arguably a necessary feature of such systems) nor on specially-prepared sound examples. As a demonstration, one part of the experiment made a direct comparison between the current model and the resynthesis from a previous computational auditory scene analysis system, as a proof-of-concept for how such comparisons could be made.

6.1.2 Future developments of the model

It goes without saying that when starting the project I expected to construct a rather more complete system. Indeed some of the aspects that were not constructed were abandoned only quite recently. Although these omissions have mostly been noted already, the following list consists of the system features that I would most like to add in the future:

- **Higher-level explanations:** Much of the theoretical attraction of the prediction-driven approach stems from its capacity to incorporate high-level abstractions for observed signals, conferring the ability to predict and hence 'perceive' complex sound patterns even when they are masked or only ambiguously present. Several of the more involved aspects of the implementation, such as the RESUN blackboard engine, appear over-engineered as they stand since they are there to permit a deeper explanation hierarchy that is not currently used. Adding a wider range of hypotheses at the 'source' level, and layering higher levels of explanation above these sources, is for me the most exciting (as well as the most challenging) open issue for the system that has been presented.
- **Improvements in cues, elements and resynthesis:** The relatively desultory ratings awarded to the system's resyntheses in the listening tests reflect the difficulties arising from trying to spread finite development resources between the many aspects of the system. More care in the resynthesis algorithms would probably pay considerable dividends in eliminating damaging but avoidable artifacts. However, some problems are more profound; the signal models underlying the generic sound elements should probably be more sophisticated (such as permitting shifts in frequency for noise objects), and the analysis procedures used to recover signal parameters are certainly far from optimal. The detection and extraction of weft elements is a particularly glaring example of this, where the current algorithm for recovering the intensity of periodic components mixed with noise or other interference is based on a rather heuristic, and inadequately verified, analysis. Hopefully, a more careful derivation of this relationship would greatly reduce the spectral distortion apparent in the current weft resyntheses.

  It should be acknowledged also that the system is intrinsically monophonic; an obvious avenue to pursue would be the inclusion of the binaural spatialization cues that are known to help real listeners.

However, little consideration has been given to how such information might be incorporated (as discussed below).

- **Interpolation of corrupt data:** One of the unexplored benefits of a prediction-based system with high-level signal data is the ability to 'infer' the presence of signal continuations despite temporary masking by louder sounds. Although this is primarily an aspect of the higher-level explanations mentioned already, such reconstruction could probably be implemented quite successfully even at the level of the bottom-level elements. The missing piece here is an algorithm to go back and interpolate best-guesses over the missing data if and when the target signal reappears. A preliminary algorithm of this kind was presented in [Ellis95c], but never made it into the current system.
- **Revision of previous analyses:** As touched upon in the theoretical motivation of the prediction-driven architecture in chapter 3, there are certain situations in which the correct interpretation of a sound cannot be made until some time into the future, when the subsequently-revealed context dictates the correct analysis. Where ambiguity is detected in advance, the system as presented can incorporate such retroactive context via the preference between multiple hypotheses. However, in some cases, even the ambiguity does not emerge until later (such as the case of the alternating noise stimuli starting with a wide band of noise, illustrated in fig. 3.2 (b)); to handle such situations, a system must be able to go back in time and revise its representation for information that has already been 'finished'. Although the architectural complications are considerable, there is no intrinsic reason why the current system could not include a mechanism of this kind, that 'backed up' the evolution of hypotheses to revise or divide them at an earlier time-step. The analysis system described was strictly incremental, making decisions only at each hypotheses' idea of the current time ; however, the underlying blackboard engine places no restrictions on switching between current and past hypotheses.

These are the aspects of the system that appear most worthy of development in terms of the original intentions of the project and my hopes for the model's ability to reproduce human auditory behavior. However, the real motivation for developments may depend on particular goals, for instance, speech enhancement, towards which such a system might be aimed. Goals create their own agenda of development priorities to handle the immediate obstacles; it seems unlikely that the special case illustrated in figure 3.2 (b) will soon be one of them.

6.2 Conclusions

Any presentation of a new computational system will inevitably consist of a mass of details, but what are the broader lessons to be drawn from this work? In the introduction I mentioned various specific questions and assumptions that the project would address; we can now revisit these points. There are also a number of unanticipated issues that arose during the project that deserve mentioning here. The ultimate goal of this work, to create a functional model of human audition can now be re-evaluated in terms of the kinds of perceptual phenomena that this approach at least begins to explain; other aspects of real hearing whose emulation still presents theoretical challenges will also be considered.

6.2.1 Reviewing the initial design choices

The project embodied a somewhat specific point of view on the mechanisms of auditory perception. Despite the indifferent quality ratings resulting from the subjective tests, the mere fact that such tests could be conducted, that the system could identify and extract sufficient structure from complex real-world sounds to permit a comparison with the impressions of real listeners, serves as strong support for the assumptions upon which the model was based. Specifically, although the following ideas cannot be said to have been proven in this work, my original commitment to them has certainly been reinforced by the experiences of the project:

- **Full explanation:** Deeper than the technical distinctions between this work and its predecessors lies the philosophical difference between seeking to extract a target sound from unwanted interference, and the approach espoused here to find an explanation for *everything* in a sound. That this forms an essentially correct interpretation of human audition does not seem controversial, however, this work has helped to uncover the more particular implications of complete explanation in accommodating the difficulties that arise from the interaction between different sounds in a mixture.
- **Models for aperiodic sound:** The goal of modeling a broad range of sounds, rather than focusing on a specific target class such as voiced speech, required a comprehensive representational basis. While identification of a background 'noise floor' has been attempted in several speech processing systems, the dynamic modification of these elements, and the addition of explicitly transient click elements were particular to this approach.
- **A new representational vocabulary:** Although the basis set of noise clouds, transient clicks and tonal wefts was highlighted as one of the more speculative parts of the system, they turned out to be an adequate foundation for the analysis and resynthesis of the sound ambiences addressed in the experiments. Moreover, these elements were able to represent a given sound on a much coarser scale, i.e. with many fewer elements, than the representational primitives such as sinusoids or contiguous time-frequency patches used in previous systems. To achieve a coarser-grained representation without inappropriately grouping energy from distinct sources is of great benefit in simplifying and reducing subsequent processing.
- **The world model:** The specific domain of dense ambient sound mixtures addressed made intermittent masking of sound sources by one another an inevitable feature of the environment. Consequently, the analysis needed to have an abstract internal representation capable of maintaining the existence of inferred sound-producing processes even when the evidence of their presence was temporarily obscured. The approach of a top-down internal world model, reconciled with the observed input, rather than a bottom-up analysis derived directly from the input, made persistence in spite of hidden evidence the default behavior of the system, as appears to be the case in the auditory system.
- **Multiple hypotheses:** Top-down explanation, which essentially tries to guess and then confirm an abstract explanation, can raise specific problems of ambiguity in identifying correct solutions. By taking the approach of developing multiple hypotheses until the best alternative

emerged, many such ambiguities could be resolved without complicated backtracking and revision.

- **Blackboard engine:** The blackboard architecture was a natural choice for a sensor-interpretation problem involving both bottom-up and top-down reasoning in which the precise sequence of processing steps could not be specified in advance. While certain features of blackboard systems, such as their ability to apportion effort between processing at widely differing levels of abstraction, were not exploited in the current implementation, the basic capacity of the blackboard engine to manage multiple hypothesis hierarchies proved to be a very effective framework for the system.

6.2.2 Insights gained during the project

One of the strongest arguments in favor of building computer simulations as a way to investigate phenomena of perception and cognition is that we have a unfortunate tendency to overlook many of the most critical issues in the brain's operation, perhaps arising from the recursive nature of considering our own minds' function. An attempt at a concrete implementation of the principles and processes deduced from physiological and psychological evidence has the benefit of making any gaps in the original reasoning painfully obvious, and I think it is fair to say that the difficulties encountered by modelers of auditory organization have contributed to the theoretical understanding of the phenomenon. Although the idea of a prediction-reconciliation model for auditory organization seemed reasonably straightforward, some of the pieces required for the current implementation were not anticipated but turned out to be very interesting none-the-less. The particular problem of allocating prediction error between overlapping objects, and its solution through the concept of decaying 'error weights' reflecting the stability of each element, are an example of a practical solution that could have more general implications for the structure of real audition. While the minimum-description length principle is mainly an aid to the developer which placed a specific interpretation on the otherwise vague concept of a hypothesis quality rating, the associated questions of the parameters to use in assessing the quality of hypotheses, answered as a combination of prediction accuracy and prediction specificity, should form an important part of any auditory scene analysis system.

The system as presented is *incremental*, in that it observes its input in strictly advancing small time steps; most previous systems (with the exception of [Mell91]) were not. Rather, those system would consider a chunk of sound on the order of a second in length before deciding, as a whole, what it contained. The distinction between an incremental system whose 'current beliefs' can be meaningfully described for every instant of the input, and systems whose analysis is updated only on a coarser timescale, turned out to be more significant that I had anticipated. The human auditory system operates incrementally, in that it must provide a provisional analysis of current sounds whether or not they have evolved to full characterization (although the exact delay between exposure to a complex sound and its correct organization is difficult to define). Compared to a batch system, many of the issues of multiple hypotheses and handling ambiguity take on much greater significance under the incremental constraint, presumably reflecting their importance in real audition too.

Resynthesis is an obvious goal for certain kinds of applications, such as the restoration of degraded recordings, but it is not immediately clear that the

pure intellectual goal of understanding and reproducing the organization of sound in the brain need be concerned with resynthesis. However, in considering the assessment of this work, it became clear that, because there is no reference for the task we are trying to model other than human listeners, the only option currently available for the validation of computational auditory scene analysis systems is through listening tests. Administering such tests necessitates the realization of system's output in sonic form. Resynthesis does involve rather stern requirements compared to a system whose output might be left in more abstract and undetailed terms, but probably these are requirements not to be avoided if we are serious in our goal of modeling real audition.

The final lesson of this work comes from what wasn't accomplished. It is easy to agree with the statement that more sophisticated models of hearing need to address more abstract levels of auditory processing and knowledge; the fact that this project never really reached the heights of abstraction to which it initially aspired attests to the difficulty of building such models. Partly this is a question of scale, in that it is very difficult to experiment with models of abstract function unless you have a suitable infrastructure to ground those models in real sound examples. If the infrastructure needs to be built, it must take precedence over, and may eventually entirely eclipse, the original abstract goal. Also, the increasing separation between the inferred processing of abstract representations occurring in the brain and the measurable phenomena of psychoacoustics make it that much harder to think clearly about what should be happening at these levels. One of my hopes, to achieve some clarification concerning the practical details of managing and dealing in abstract representations of sound as a byproduct of constructing computational models, must wait still longer for realization.

6.2.3 A final comparison to real audition

This work was introduced unambiguously as an effort towards the understanding of auditory function; a particular sound processing system has been presented and examined, but there are still many legitimate questions over the relevance of this model to the sound organization performed in the brain. What aspects of the system are most significant to its goal of providing insight into real audition, and what are the phenomena of hearing that the model does not address? How should we regard the very obvious differences between this system and its putative prototype, most significantly that one is implemented on a microprocessor and one with neurons?

The strength of this system's claim to be an interesting model of hearing stem from its potential to reproduce and explain various hearing phenomena that gave difficulties to preceding systems. These were presented in chapter three: Essentially, there are a wide range of circumstances in which the ear must find it difficult or impossible to gather information from a particular sound source (owing to the corrupting effect of other sources), yet for the most part we are unaware of these temporary processing obstacles because of preconscious inference and restoration in the auditory path. Such processes are highlighted by psychoacoustic demonstrations such as phonemic restoration and the continuity illusion, but we may speculate that they are constantly at work in day-to-day acoustic scenarios. A prediction driven model naturally incorporates the ability to exhibit these phenomena, since rather than relying on direct evidence for every sound-source believed to be present, the default mode of operation is to assume the continuity of the internally-represented sounds until there is a specific reason (either from

external data or internal knowledge) to change them. It is a weak argument, but the fact that an architecture based on multiple competing hypotheses strikes me as a very promising model for real audition must say something about its compatibility with the introspective experience of hearing, which, while obviously unreliable, is still some of the most detailed evidence we have to apply to the problem.

At the other extreme, the results of the experiments do provide some very concrete support for the relevance of the model to real audition, albeit not without equivocation. Although there is a natural tendency to focus on the specific flaws in the system's organizations and resyntheses, I would prefer to emphasize the unprecedented success of the model in producing anything resembling the complete analysis of a dense scene by a human listener. This success is tempered by the fact that, to my knowledge, no-one has previously attempted to model this particular aspect of perception (i.e. the complete explanation of dense ambient scenes), but I believe that these are appropriate issues to be addressing given the current state of the field.

There are of course a great many aspects of audition for which the system and the approach it embodies do not offer any obvious solution. Some of the more intriguing of these include:

- **The integration of different cues:** Although this has long been identified as an important aspect of auditory processing (for instance, in Bregman's 'trading' experiments [Breg90]), and despite the fact that abstract representations offer a suitable integrated location at which to combine information from different sources, it turned out that the system built for this project never really had to address the general problems of arbitrating between alternative interpretations mediated by orthogonal cues such as harmonicity and spatial location. (Perhaps using binaural inputs to the model and the listening tests would have made it unavoidable to address this issue, although it is possible that serious *conflicts* between cues, the only times that the manner of their integration takes on significance, are rather rare in practice). There are some very interesting but difficult questions here arising from the paradox of needing an identified object before its parameters can be reliably calculated at the same time as needing the information from those parameters to construct the object robustly; regrettably, the current work has made little contribution to this question.
- **Duplex perception:** This is another favorite example of the complexity of hearing, in which a simple sound element such as a modulated sinusoid is seen to contribute to more than one perceived sound source, perhaps being heard both as a whistle and contributing to the quality of a simultaneous vowel. Constructing such demonstrations is relatively difficult, and I would prefer to be able to discount such oddities as irrelevant to real, everyday hearing. Certainly, if they could be explained in terms of the masking properties of the whistle permitting the reconciliation without contradiction of input components to the a-priori likely interpretation of a vowel, then the prediction-driven approach would fit right in. However, no direct effort was made to consider this general question in the current work.
- **The perception of modulation:** The detection of patterns of energy fluctuation spread across different peripheral frequency channels, as in comodulation masking release (CMR, [HallG90]) has been the subject of some very suggestive recent psychoacoustic experimentation. The model

as presented does not have the capability to represent and compare details of aperiodic modulation between frequency channels that such phenomena appear to imply. That said, the detection of periodicity in the correlogram/weft analysis consists primarily of the recognition of common modulation across frequency bands; it may be that some modification of this approach would be able to encompass not only pitched sounds but noisier comodulated examples too, perhaps via the addition of extra layers or dimensions to the periodogram output.

- **Abstract sound explanations:** At the opposite extreme to the low-level mechanisms of modulation detection, the structure of the higher levels of abstraction were also not a part of the current system, as has already been discussed. There are many questions concerning such abstractions that need to be addressed; if sound events are to be organized as instances of known patterns, what is the correct balance between the number and generality of such patterns, and how do specific abstract entities (such as "a clarinet note") inherit properties from the more general classes in which they presumably reside ("wind instruments", "musical sounds", "sounds with a pitch" as well as perhaps "instruments used in jazz music" or other overlapping classifications).
- **Developmental and learning issues:** Even establishing a suitable hierarchy of abstractions does not solve this question completely, since real listeners evidently acquire their own interpretative predispositions through some combination of innate tendency and learning from their environment; it may be that learning over a range of timescales is an important prerequisite to the ability to organize the full range of sounds we can handle. We have yet to reach a stage of sophistication in our systems where such questions can be effectively investigated.

Some would consider the fact that the current system is based upon the 'traditional AI' framework of rules, representation and symbolic processing as a significant obstacle to its consideration as a plausible model of what goes on in people. The points raised in relation to artificial neural network models in chapter two still seem valid: While there are without doubt numerous extremely interesting discoveries to be made about how the auditory system implements its processing using the structures of the brain, these are ultimately issues only of implementation. In the same way that looking at highly-optimized procedures or the output of a compiler may disguise aspects of the underlying algorithm being employed, it might be convenient or even important to attempt to understand the function of audition independently of how it is actually achieved by neurons. At the same time, it may be that the algorithms employed are so overwhelmingly influenced by the capabilities of their implementational substrate that they will only be comprehensible from that perspective; certainly, both approaches are warranted at this stage. However, I do not consider it a disadvantage that the serial control scheduling of the blackboard core of this system is on its face an extremely unbiological form of computation. From just a slightly broader perspective, its structures of parallel development of competing hypotheses may turn out to be a very serviceable parallel to the behavior of the brain viewed in similar terms. That their lowest levels of implementation are almost unrelated is no more important than the difference between solving an integral on paper or using an automated tool like MACSYMA.

As we have already observed, work in this field is complicated considerably by the impracticality of direct comparison between model and original except at the very highest and lowest levels, though psychoacoustic tests and

physiological examination respectively. Spanning this gap requires many pieces, and there is a problem of apportioning the blame for weaknesses in any particular model, even at the coarsest level, between front end, intermediate representation, resynthesis procedure or any of the other major components. This makes progress towards a successful model a slow and confused path, never knowing if the compromises in a particular aspect of the system are truly unimportant, or the one crucial flaw obstructing some interesting behavior. Given these multiple possible points of failure, it is sometimes surprising that the models we do have work as well as they do. Such a realization is the comfortless truth facing those of us who persevere with such modeling.

6.3 The future of Computational Auditory Scene Analysis

The initial goals of this project were too ambitious, and while it's fun to think in grandiose terms about the 'whole' of audition, it is symptomatic of the nascent state of the discipline that I consider my area of study to be computational auditory scene analysis without being able specify an obvious specialization within it. Like Newell & Simon's "General Problem Solver" [NewS72], and even Marr's "Vision" [Marr82], today's most sophisticated theories will appear naive and almost willfully simplistic in the quite near future. This is inevitable; the challenge is to make the discoveries that will permit a more realistically complex model of auditory perception.

Progress will be made when individual researchers are able to focus their attention on individual parts of the problem, rather than being distracted by the simultaneous and interdependent development of disparate model components. This is becoming easier with a trend for modelers to package their work and make it available as a substrate for other researchers – for instance, the cochlea model of [Slaney93] used here, and the integrated modular physiological models made available by the group at Loughborough [OMaHM93]. These pieces reflect some kind of consensus concerning acceptable models of the auditory periphery and low-level processing, but before we can see the same kinds of tools becoming available for, say, intermediate auditory representations, we will need much better evidence for what the appropriate representations should be. Before we can really attack this problem in depth, we need to find a common framework for the whole of audition to act as a unifying focus for this work. Such a framework could emerge from a model that was peculiarly successful at accounting for a range of phenomena, or a breakthrough in physiological results. It's hard to anticipate, but in all likelihood we are some time away from being in such a position.

In the meantime, progress will continue in the areas already identified as the main components of models of auditory organization. Increased physiological knowledge needs to be matched by improved interpretations of the significance and purpose of the inferred operations. There is a tantalizing parallel between the autocorrelation structures that account so well for pitch detection [MeddH91] and cross-correlation models of interaural cue detection [Gaik93]. Perhaps we will be able to identify a class of delay-and-combine networks as the general building-blocks at least for the lower levels of auditory signal processing.

However, the really exciting advances from my perspective lie in the application of higher-level knowledge. Automatic speech recognition presents an interesting parallel: Many recent improvements in the performance of

speech recognizers arise from the incorporation of increasingly sophisticated language models which simplify the identification problems of the front-end by imposing powerful constraints on the word-sequences that are permitted to exist. An ability to express and manipulate these kinds of constraints in the domain of auditory scene analysis would produce similar gains, and in the same way that language models approximate principles at work in people listening to speech, powerful high-level constraints probably constitute the largest portion of the difference between computational and human auditory scene analysis. Unfortunately, the concept of a 'grammar' of real-world sounds, a way to describe the kinds of limits on what we will believe or notice that we hear, is still a baffling idea.

The comparison of machine vision and machine listening is sobering. There are similarities between our field and the computer models of vision of ten or fifteen years ago, and vision is still very far from being a 'solved problem'. However, there are reasons to be optimistic: Firstly, many of the lessons gained rather painfully in vision research (the illusion of 'direct perception'; the importance of representations and abstractions; the plurality of cues) have been incorporated directly into theories of hearing rather than having to be rediscovered. Secondly, hearing is simpler than vision, in terms of sensory bandwidth or brain real-estate, and at the same time the inexorable advance of computational power makes possible models that would previously have been unthinkable. Perhaps hearing embodies a balance between sensory richness and processing complexity at the right level to permit our current theoretical and practical capabilities to uncover basic principles of perception. Finally, there some specific differences between the two domains that might at least help auditory modelers avoid some degenerate traps. Dynamic features are central to our perceptual experience, yet there is a temptation in visual research to 'simplify' problems to the analysis of static images. To the extent that 'static sounds' can be defined at all, they constitute a wholly uninteresting class of stimuli, thus all researchers who deal in sound must treat time as a 'first class' dimension. Perhaps we can ultimately repay our conceptual debt to the machine vision community with general principles of dynamic perception of benefit to both modalities.

Perception is the right biological mystery to be studying at the moment, given our experimental and computational tools. Its solution will lead naturally into the deeper cognitive secrets of the brain. I look forward to the advances of the coming years with excitement and impatience .

Appendix A Derivation of the weft update equation

The weft element was introduced in chapter 4 as a way of representing, in a single object, energy exhibiting a particular rate of periodic modulation across the whole spectrum. A weft is defined in two parts: its period track, indicating the common modulation rate as it varies in time, and its 'smooth spectrum', describing the amount of energy that shows the modulation rate at each time-frequency cell involved. Sound is resynthesized from weft elements by a simple source-filter algorithm, with the period track generating an impulse train which is shaped to match the 'smooth spectrum' by slowly-varying gains applied to each channel of a filterbank.

The more difficult problem is to extract the weft parameters from a real sound. The period track is derived from the periodogram, a summary of the short-time autocorrelation in every frequency channel, as described in chapter 4. If the sound consists only of energy modulated at a particular rate, then measuring the spectrum is simply a question of recording the energy in each frequency channel of the original sound. However, in the majority of interesting cases, the periodically-modulated energy is mixed with other energy, either unmodulated or showing a different periodicity. This appendix presents an analysis of such a situation, leading to the equation which is used to estimate the energy properly belonging to a given modulation period from the autocorrelation function of sound mixture.

The periodicity analysis performed by the front end first breaks the sound up into frequency channels with the cochlea-model filterbank, then rectifies and smoothes this band-pass signal to get an envelope, and finally forms the short-time autocorrelation of this envelope signal by passing it through a tapped delay line, multiplying each tap by the undelayed envelope signal and smoothing the product. This procedure generates the correlogram volume (time versus frequency versus autocorrelation lag) and is illustrated in figure 4.9. Consider first a purely periodic sound such as an impulse train with period τ subjected to this analysis. Assuming the pulse train period is long compared to the impulse response of the filter channel being examined (for instance, the upper spectrum of male voiced speech), the bandpass signal will consist of repetitions of the filter's impulse response spaced τ apart, one for each pulse in the pulse-train. The smoothed envelope of this signal will follow the general shape of the impulse response, subject to a constant gain factor in peak amplitude resulting from the rectification and smoothing. Examples of these signals are shown in figure A.1.

As expected, the short-time autocorrelation of this signal shows a peak at the signal period, τ (9 milliseconds in the figure). Since the signal is exactly periodic, multiplying it by itself delayed by τ is the same as multiplying it by itself with no delay, and thus the short-time autocorrelations at zero lag and at a lag of τ are the same, both equaling the average power of the envelope signal (since the smoothing of the product has calculated the average). The envelope level is proportional to the excitation level in this band, and thus after factoring out the fixed effects of the filter impulse-response length and the envelope extraction process, the intensity of energy in that frequency channel which is periodic at τ is obtained directly from the level of the

autocorrelation peak at that period. This is the basic argument given in subsection 4.3.3.

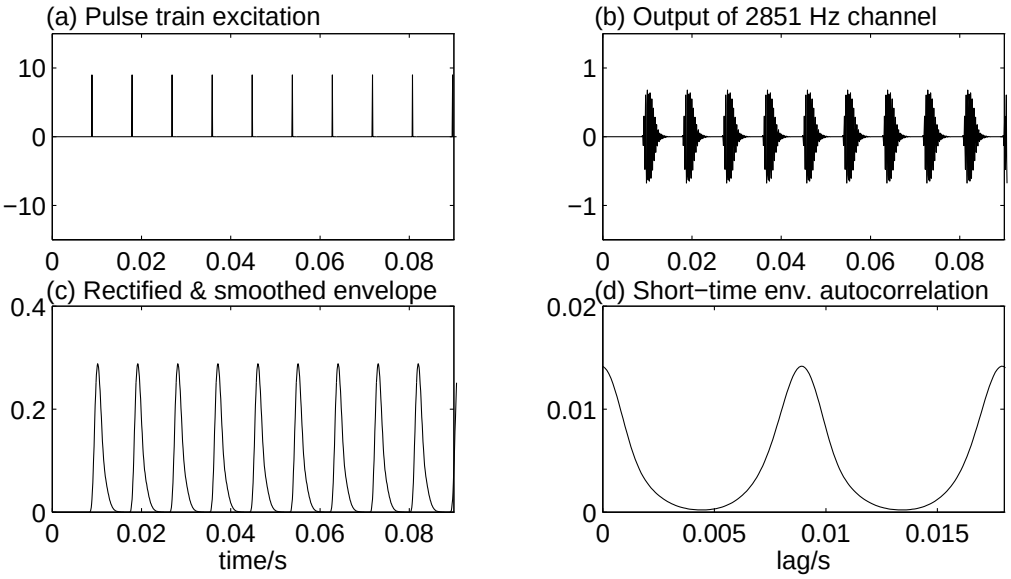

Figure A.1: Periodicity analysis of a purely periodic signal. Panel (a) shows the excitation, a pulse train. (b) shows the output of a filter channel far above the fundamental frequency of the pulse train, which is approximately the impulse response of the filter repeated for each pulse in the excitation. The envelope of this signal is shown in (c), obtained by rectification and smoothing. (d) shows the short-time autocorrelation of this signal, obtained by multiplying the envelope with versions of itself delayed by each lag value. The peak at 0.009 s corresponds to the period of the original excitation.

If, however, the signal contains additional energy not modulated at τ, this relationship no longer holds. Consider the case where the excitation consists of the purely-periodic pulse train plus some stationary noise. The bandpass signal will be the linear sum of the filter response to the pulse-train alone and to the noise . In the spaces between the pulses of fig A.1(b), the sum will be dominated by the noise energy, and the envelope will vary around the average level of the rectified noise. Where the impulse excitation has made a significant contribution to the filter output, the two components will have a random phase alignment, and the smoothed envelope will record the average level of their incoherent addition – such that the expected value of the square of the envelope equals the square of the noise-free envelope plus the expected value of the square of the noise signal's envelope. This situation is illustrated for real instances in figure A.2.

The problem we would like to solve is to extract the energy of the periodic excitation from the autocorrelation of the signal to which noise has been added. Comparing figure A.1(d) with A.2(d), the peak at the periodic lag has been boosted by the additional noise; however, the amount of periodic energy present in the excitation is essentially the same since only aperiodic noise has been added, and we would like to be able to use additional measurements from the autocorrelation to factor-out this overestimation of periodic energy level resulting from the noise.

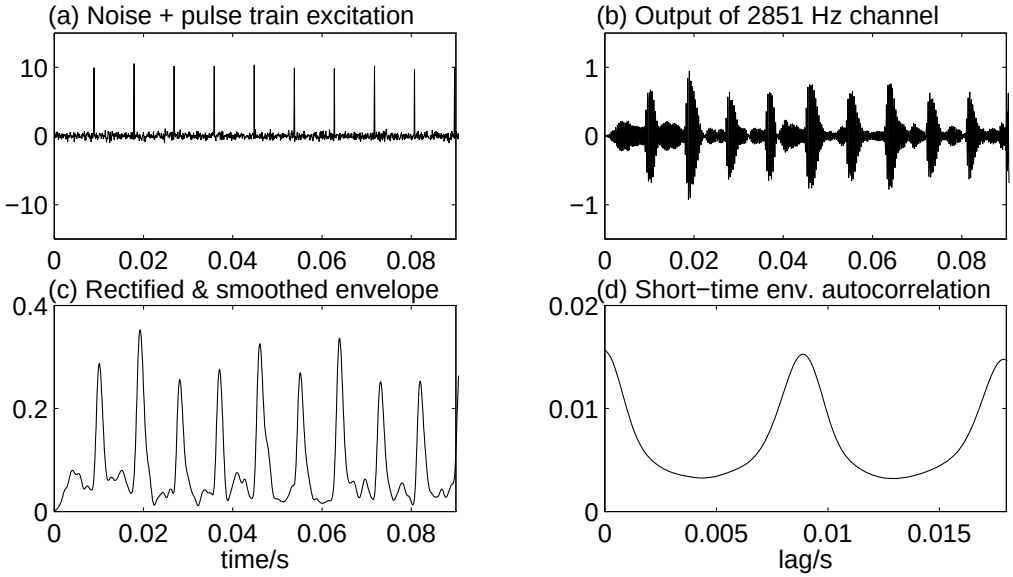

Figure A.2: Excitation, band-pass signal, envelope and short-time autocorrelation for the impulse train of fig. A.1 after the addition of some Gaussian white noise. Note that compared to fig. A.1(d), the autocorrelation has slightly raised peaks and considerably raised troughs.

Consider the following approximation to this situation. In figure A.1, the envelope of the impulse-train-excited filter output is a series of disjoint bumps associated with each pulse, since the filter impulse response has most of its energy concentrated over a time which is shorter than the excitation period. Let us approximate this envelope as rectangular pulses of a certain duration, separated by stretches of zero amplitude. If we then add stationary noise to the excitation, the expected level of the envelope will be that of the noise alone in these gaps; during the rectangular bursts, we assume that the noise and pulse-response signals add incoherently. Thus if the envelope of the noiseless tone-burst signal is:

$$e_T(t) = \begin{cases} T & r \cdot \tau < t - t_0 < (r + \delta) \cdot \tau \quad r = 0, \pm 1, \pm 2 \ldots \\ 0 & \text{otherwise} \end{cases} \tag{A.1}$$

where T is proportional to the impulse train energy in the filter channel, τ is the impulse train period, $\delta \cdot \tau$ is the duration of our rectangular approximation to the filter impulse response envelope, and t_0 is an arbitrary phase offset, then the expected value for the envelope resulting from excitation by a pulse train with added noise is:

$$E\left[e_M(t)\right] = \begin{cases} U & r \cdot \tau < t - t_0 < (r + \delta) \cdot \tau \quad r = 0, \pm 1, \pm 2 \ldots \\ N & \text{otherwise} \end{cases} \tag{A.2}$$

where N is proportional to the noise energy in the filter channel, and the level of the envelope during the bursts,

$$U = \sqrt{N^2 + T^2} \tag{A.3}$$

This situation is illustrated in figure A.3.

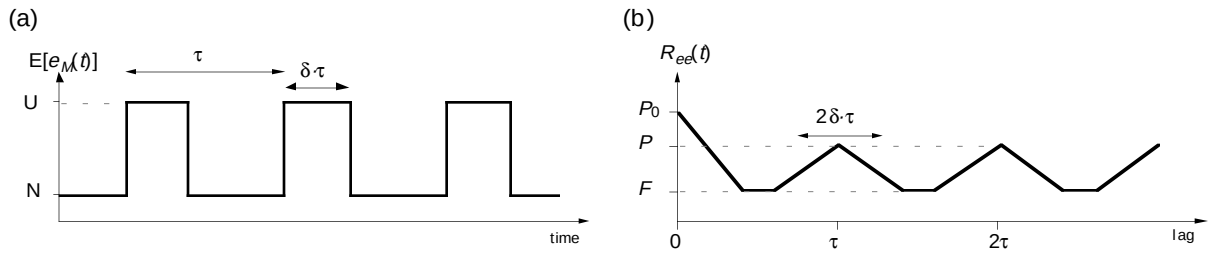

Figure A.3: Simplified approximation to the expected value of the envelope of the filter output excited by the noise+pulse excitation of A.2. Panel (b) shows its theoretical autocorrelation.

In considering the short-time autocorrelation of this signal, we can drop the expected-value operator, since the autocorrelation involves averaging over some time window. Thus the theoretical form of the autocorrelation of this approximation to the bandpass envelope is $R_{ee}(t)$, as illustrated in figure A.3(b), for the case where δ, the proportion of the excitation period occupied by the impulse response, is less than 1/2. In this case, the peak autocorrelation at the lag equal to the excitation period τ is simply the average level of the envelope multiplied by its image at a distance τ away, i.e.:

$$P = R_{ee}(\tau) = \delta \cdot U^2 + \gamma \cdot N^2 \qquad \text{(A.4)}$$

where δ is the proportion of the time that the envelope is at U, and γ is the proportion of the time it is at N, i.e.:

$$\gamma + \delta = 1 \qquad \text{(A.5)}$$

The autocorrelation at lag zero is somewhat larger (shown as P_0 in the figure) since this is the average squared value of the signal, and we can no longer treat the bandpass envelope as its expected value, but must consider the actual variations arising from the noise excitation whose amplitude depends on the amount of smoothing used to extract the envelope. However, it turns out that the value of P_0 is not relevant to this analysis.

Recall that our goal is to find a way of estimating the contribution of periodic excitation from the autocorrelation of the envelope of a channel excited both by the periodic energy and additional corruption. What we would like to do is find some parameters, measurable from the real autocorrelation functions calculated in the implementation, that will lead to T, the parameter in our approximation which is proportional to the amplitude of the periodic excitation. In the noise-free case, we were able to use the square-root of the autocorrelation peak at the appropriate period which would indicate the amplitude of the periodic excitation; in order to solve simultaneously for the noise level, we need to define a second measurement from the autocorrelation function. Looking at figure A.3, one possibility would be the *minimum* value of the autocorrelation, labeled as F in the figure. F is the average product of two shifted versions of the envelope approximation where the pulses do not overlap, and hence:

$$F = 2\delta \cdot U \cdot N + (1 - 2\delta) \cdot N^2 \qquad \delta < \frac{1}{2} \qquad \text{(A.6)}$$

However, this turns out to be a noisy parameter, since the influence of the noise excitation on the autocorrelation can form narrow notches, distorting the intention of the minimum-value measurement. A second problem is that

the flat minima of figure A.3(b) only occur when $\delta < 1/2$, i.e. when the filter impulse response support is less than half of the excitation period, which usually only holds at a few channels at the very top of the spectrum.

A more reliable measure is the average value of the autocorrelation, taken over the width of the period in question to capture the true average of the periodic autocorrelation resulting from the modulated envelope. In order to avoid the boost from the correlated noise at lag zero (i.e. the difference between P and P_0), the average is calculated over the autocorrelation from lag=τ to $2 \cdot \tau$. By considering this average as the expected value of two points chosen at random from the signal envelope, we can see that the average value,

$$A = \left(\delta \cdot U + \gamma \cdot N\right)^2 \tag{A.7}$$

i.e. the square of the average value of the envelope. This is confirmed by solving geometrically for the area under the autocorrelation function of figure A.3(b) using the value of F given in eqn. (A.6).

Both P, the level of the autocorrelation peak at the excitation period, and A, the average level of the autocorrelation over one excitation period's worth of lag, are easily extracted from the system's front-end. If we know the value of δ (and hence γ), we can solve (A.4) and (A.5) for N and U; this gives a quadratic whose two roots correspond to the symmetric solutions when $N<U$ and $N>U$; we consider only the first root, since U must be greater than N:

$$N = \frac{2\gamma\sqrt{A} - \sqrt{4\gamma^2 A - 4\left(\gamma^2 + \gamma\delta\right)\left(A - \delta \cdot P\right)}}{2\left(\gamma^2 + \gamma\delta\right)} \tag{A.8}$$

$$U = \sqrt{\frac{P - \gamma \cdot N^2}{\delta}} \tag{A.9}$$

We can get back to our periodic-excitation-amplitude parameter T as:

$$T = \sqrt{U^2 - N^2} \tag{A.10}$$

More particularly, the periodic excitation energy in that frequency channel is given by the energy of the signal whose envelope was approximated as the sequence of rectangular bursts of height T and 'duty cycle' δ. Hence the actual periodic excitation energy (the value for one time-frequency cell of the weft's smooth-spectrum) is:

$$E_p = \delta \cdot \left(\frac{T}{k}\right)^2 \tag{A.11}$$

where k is a scaling parameter to account for the envelope extraction operation. For simple smoothing of a half-wave rectified sinusoid, the ratio of smoothed level to the rms level of the original sinusoid is $\sqrt{2}/\pi \approx 0.45$, which is the value used here for k.

The analysis has relied on an approximate model of the signal envelope and a simple noise assumption. However, the final equations provide an estimate of the periodic excitation energy that depends on robust parameters from the autocorrelation – the peak and average values. The further assumption used in extracting wefts is that this relationship holds, to a greater or lesser

extent, for other situations (i.e. when $\delta > 1/2$) and for other forms of interference (such as periodic excitation with a different period). The only obstacle that remains to using the results of this analysis is to determine the value of δ for a given frequency channel and a given excitation period τ. Indeed, we would like to be able to use this relationship even for low-frequency channels where the filter impulse responses may be *longer* than the excitation period, and hence the signal envelope will show the results of specific phase cancellation between successive instances of the impulse response, making peak and average autocorrelation values depend on the excitation period in a complicated fashion.

This problem was handled by constructing a table of δ's value for every possible combination of frequency channel and target period. Both these parameters were quantized, so the total table size was only $40 \cdot 240 = 9600$ values. We previously defined δ in terms of the approximation to the filter impulse response envelope, which leaves it undefined for combinations of channel and period for which that approximation cannot be applied. Consider, however, the last term of the numerator of the solution for the quadratic in N of eqn. (A.8):

$$(A - \delta \cdot P) \tag{A.12}$$

If this goes negative, the result of the square-root is larger than the first term in the numerator, giving a negative value for N, and an illegal solution to the problem. When this occurs algorithmically, we assume that the average value A was smaller than expected for the peak value P, thus the noise interference (which contributes to the level over the whole autocorrelation) must be small compared to the periodic excitation (which contributes mainly to the peak at the corresponding lag). Hence, we assume that the noise is effectively zero, and $T = U = \sqrt{P}$.

More importantly, however, considering this situation also gives us the answer to our problem of choosing a suitable value for δ in situations where the original assumptions may not hold. If the noise level is actually zero, eqn. (A.8) will solve to zero implying that the term in (A.12) will evaluate to zero. Thus an equivalent definition for δ is as the ratio of average-to-peak autocorrelation for a noiseless periodic excitation,

$$\delta = \left. \frac{A}{P} \right|_{\text{purely periodic excitation}} \tag{A.13}$$

By measuring the peak and average autocorrelation levels for the analysis filterbank excited by pulse trains of each different period, the entire table of δ(freq. channel, modulation period) was empirically constructed. This was the last piece needed to permit the extraction of weft spectral energy levels using equations (A.8) to (A.11).

Another possible explanation for the relatively advanced state of computer vision in comparison to computer hearing is that paper, historically the medium of choice for academic communication, is far more amenable to carrying representations of visual signals than of acoustic ones. Moreover, an additional argument to support my contention that the time is ripe for research in computational auditory scene analysis is the rapid growth in networked multimedia communications, making it increasingly common to find research results presented via sound examples universally available on the Internet. While the substance of this thesis might be most successfully experienced in its venerable, printed form, one unexpected benefit of finishing it now, rather than, say, three years ago, is that I can make the sound examples immediately available to a wide community through the World-Wide Web.

The web site for this thesis can be found (at the time of writing) at:

`http://sound.media.mit.edu/~dpwe/pdcasa/`

where 'pdcasa' stands for prediction-driven computational auditory scene analysis. The rest of this appendix duplicates the description of the sound examples available at the web site.

The site is structured as a set of pages each describing a particular sound example, following the examples in chapter 5.

Example 1: Alternating noise (from section 5.1.1)

This sound is the artificial alternation of low-band and broad-band noise used as an example of the 'old-plus-new' organizing principle. The sound examples illustrating the system's organization into a continuous low-band of noise with additional high-frequency bursts, are:

1.1 The original sound (2.2 seconds)
1.2 The noise cloud comprising the continuous low-frequency noise
1.3 The remaining noise clouds for the higher-frequency energy, merged into a single soundfile
1.4 All the system-generated noise clouds summed together to reproduce the input.

Example 2: Single speaker (from section 5.1.2)

The second sound example is a male speaker saying "bad dog" against a quiet but audible background of fan noise. The sound examples, illustrated in figure 5.7, are:

2.1 The original sound (1.8 seconds)
2.2 The background noise cloud ("Noise1" in the figure)
2.3 The two wefts comprising the voiced speech ("Wefts1,2")
2.4 The three click elements ("Click1", "Click2" and "Click3")
2.5 The wefts and clicks added together, attempting to reproduce the speech without the noise. The clicks (corresponding to stop releases in the original sound) do not stream well with the voiced speech.
2.6 All the elements to reconstruct the original.

Example 3: Mixture of voices (from section 5.1.3)

This is the mixture of male and female voices used as an example in Brown's thesis. The analysis objects are illustrated in figure 5.9. The sound examples are:

3.1 The original voice mixture (Brown's "v3n7", 1.6 seconds)
3.2 Background noise cloud ("Noise1")
3.3 Weft corresponding to continuously-voiced male speech ('Weft1")
3.4 Four wefts comprising the voiced female speech ("Wefts2-5")
3.5 Four click elements attached to the speech onsets ("Click1"-"Click4")
3.6 Attempted reconstruction of female voice with both vowels and consonants ("Wefts2-5", "Click1"-"Click4", "Noise1").
3.7 Reconstruction with all elements.
3.8 Brown's example resynthesis of the male voice, for comparison.

Example 4: City-street ambience (from section 5.1.4 & 5.3.2)

The sound example that motivated the entire thesis, as illustrated in figure 5.11. The sound examples are as follows, where the subjective event name is used when appropriate:

4.1 The original sound (10 seconds)
4.2 Background noise cloud ("Noise1")
4.3 Crash ("Noise2, Click1")
4.4 Horn1, Horn2, Horn3, Horn4, Horn5 ("Wefts1-4", "Weft5", "Wefts6,7", "Weft8", "Wefts9-12")
4.5 Complete reconstruction with all elements

Example 5: Construction site ambience (from section 5.3.3)

This was the second ambient example used in the subjective listening tests. The elements and the subjective groups are illustrated in figure 5.19.

5.1 Original sound (10 seconds)
5.2 Background ("Noise1")
5.3 Saw ("Noise2")
5.4 Wood hit, Metal hit, Wood drop, Clink1, Clink2 ("Click1", "Clicks2,3", "Click4", "Clicks5,6", "Clicks7,8")
5.5 Voice ("Wefts8,10")
5.6 Extra tonal background ("Wefts1-6", "Wefts7,9")
5.7 Complete resynthesis with all elements

Example 6: Station ambience (from section 5.3.4)

The final ambient example used in the listening tests was this reverberant recording of voices etc. The subjective events and analyzed elements are illustrated in figure 5.21.

6.1 Original sound (10 seconds)
6.2 Background ("Noise1")
6.3 Jiggle, Thump & other click objects ("Click1-3", "Click4", "Clicks5,6", "Clicks7,8", "Click9")
6.4 Baby1, Baby2, Man, Oh, Hey ("Weft3", "Weft7", "Weft13", "Weft24", "Weft29")
6.5 Talk & other background voices ("Wefts1-6", "Wefts8-12", "Wefts14-28")
6.6 Complete resynthesis of all elements.

Appendix C Computational environment

Chapter 4 presented the design and overview of the system implementation, but certain readers might be curious about the nuts and bolts of the constructed system. This section provides the barest sketch of the tools used for the implementation. I would be happy to provide any further details; you can contact me via email as dpwe@media.mit.edu or dpwe@icsi.berkeley.edu.

Hardware

All the computation was performed on the machines of the Media Lab Machine Listening group, primarily Silicon Graphics Indigo and Indigo^2s.

Software tools

The fixed signal processing of the front-end was performed with Matlab scripts. I had to write an interface for a special file format to deal with the three-dimensional data of the correlogram (with the 'DAT' file format used elsewhere at MIT), but apart from that it was relatively straightforward.

Programming environment

At the core of the blackboard system that comprised the prediction-reconciliation engine was the IPUS C++ Platform (ICP) by Joe Winograd [WinN95]. A complete blackboard system is constructed on this platform by implementing hypotheses, actions, etc. in C++ that use special macros defined by the main code. I used SGI's Delta C++ compiler and their debugger, occasionally dropping back into Gnu g++ when I needed better single stepping.

Having a good user interface for the system was important for debugging, since the amount and complexity of the data in the blackboard system would make the use of conventional debuggers very arduous. John Ousterhout's Tcl/Tk provided a wonderful environment for rapid prototyping and trivial user-interface construction. An interface between the C++ analysis program and the Tk user-interface was semi-automatically constructed with a highly-customized version of the ObjectTcl extensions written by Dean Sheenan. I also used Mike McLennan's [incr Tcl] object-oriented extensions for Tk-like mega-widgets.

Listening tests

Tcl/T k also provided an ideal environment for the listening tests, allowing the rapid construction of a sophisticated graphical front-end, and even managing to do most of the data analysis. A couple of small extensions provided sound output control from Tcl to play the sound examples.

Document preparation

This thesis was written on a Macintosh Powerbook Duo 270c in Microsoft Word 5.1, the same software I've been using since I started my Ph.D! Block diagrams were drawn on the Mac with Canvas; some of the earlier plots come straight from Matlab, but the later ones are postscript generated by Tk's Canvas object. In a very short time, I was able to put together a kind of

special-purpose drawing environment and a hierarchy of graphic objects that made the figures of chapter 5 relatively simple to produce. It's frightening to imagine how much of this work would have been impossible for me without the Tcl language and its community of contributing authors.

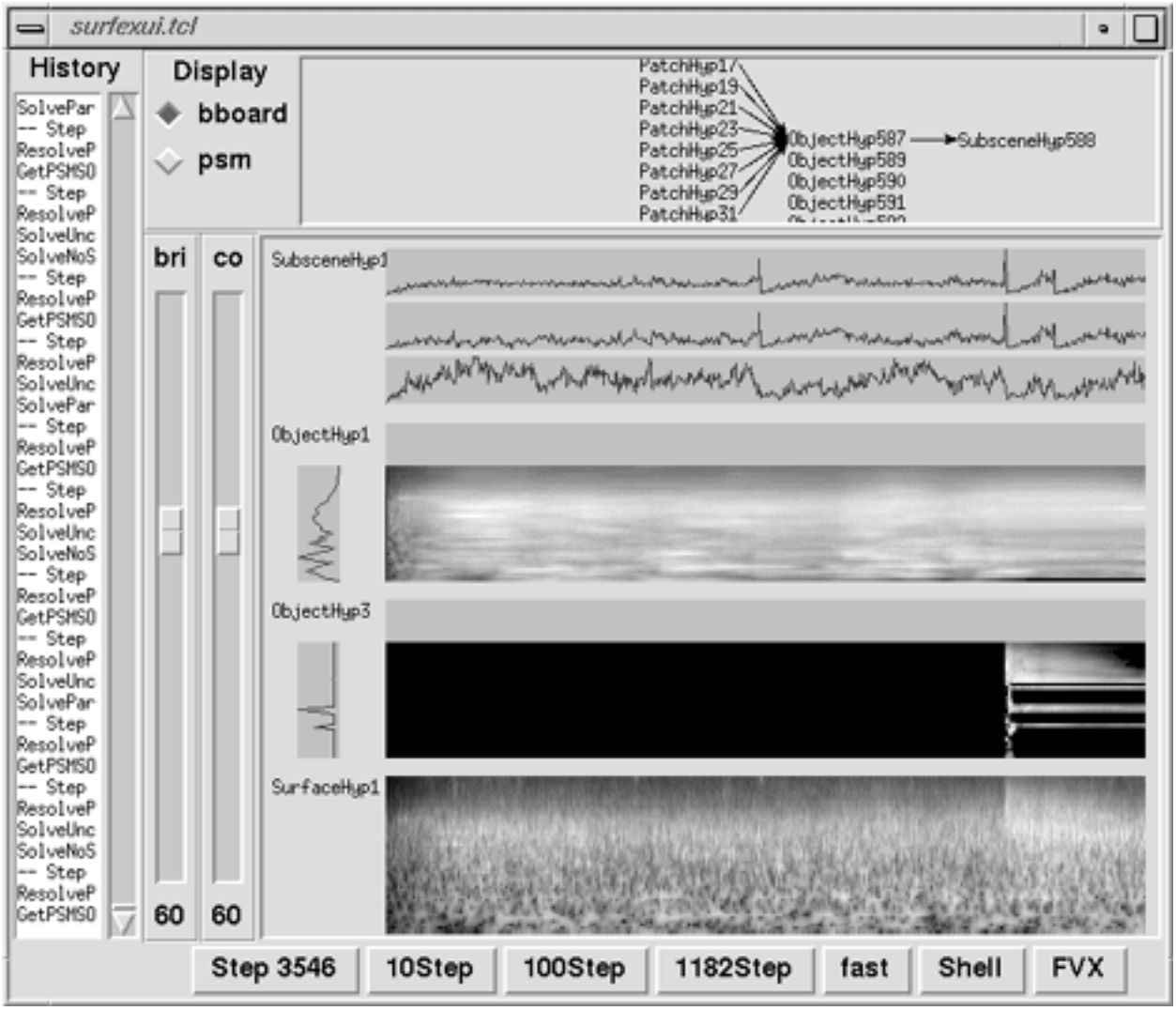

Figure C.1: Screenshot of the Tcl user interface to the C++ blackboard analysis system.

References

[AllenN92] Allen, J. B., Neely, S. T. (1992). “Micromechanical models of the cochlea,” Physics Today 45(7), 40-47.

[AssmS89] Assmann, P. F., Summerfield, Q. (1989). “Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies,” J. Acous. Soc. Am. 85(2), 680-697.

[AssmS94] Assmann, P. F., Summerfield, Q. (1994). “The contribution of waveform interactions to the perception of concurrent vowels,” J. Acous. Soc. Am. 95(1), 471-484.

[Aware93] Aware, Inc. (1993). “Speed of Sound Megadisk CD-ROM #1: Sound Effects,” Computer CD-ROM.

[Baum92] Baumann, U. (1992). “Pitch and onset as cues for segregation of musical voices,” presented to the 2nd Int’l Conf. on Music Perception and Cognition, Los Angeles.

[BeauvM91] Beauvois, M. W., Meddis, R. (1991). “A computer model of auditory stream segregation,” Q. J. Exp. Psych. 43A(3), 517-541.

[Beran92] Beranek, L. L. (1992). “Concert hall acoustics,” J. Acous. Soc. Am., 92(1), 1-39.

[BerthL95] Berthommier, F., Lorenzi, C. (1995). “Implications of physiological mechanisms of amplitude modulation processing for modelling complex sounds analysis and separation,” in working notes of the workshop on Comp. Aud. Scene Analysis at the Intl. Joint Conf. on Artif. Intel., Montréal, 26-31.

[Bilmes93] Bilmes, J. A. (1993). “Timing is of the essence: Perceptual and computational techniques for representing, learning, and reproducing expressive timing in percussive rhythm,” M.S. thesis, Media Laboratory, Massachusetts Institute of Technology.
<ftp://ftp.media.mit.edu/pub/bilmes-thesis/index.html>

[Bodden93] Bodden, M. (1993). “Modeling human sound-source localization and the cocktail-party effect,” Acta Acustica 1, 43-55.

[Breg90] Bregman, A. S. (1990). *Auditory Scene Analysis*, MIT Press

[Breg95] Bregman, A. S. (1995). “Psychological Data and Computational ASA,” in working notes for the workshop on Comp. Aud. Scene Analysis at the Intl. Joint Conf. on Artif. Intel., Montréal, 4-8.

[BregP78] Bregman, A. S., Pinker, S. (1978). “Auditory streaming and the building of timbre,” Can. J. Psych. 32, 19-31 (described in [Breg90]).

[Brooks91] Brooks, R. A. (1991). “Intelligence without reason,” MIT AI Lab memo 1293, presented at the Intl. Joint Conf. on Artif. Intel.
<ftp://publications.ai.mit.edu/ai-publications/1000-1499/AIM-1293.ps.Z>

[Brown92] Brown, G. J. (1992). “Computational auditory scene analysis: A representational approach,” Ph.D. thesis CS-92-22, CS dept., Univ. of Sheffield.

[BrownC95] Brown, G. J., Cooke, M. (1995). “Temporal synchronisation in a neural oscillator model of primitive auditory stream segregation,” in working notes of the workshop on Comp. Aud. Scene Analysis at the Intl. Joint Conf. on Artif. Intel., Montréal, 41-47.

[Carly91] Carlyon, R. P. (1991). “Discriminating between coherent and incoherent frequency modulation of complex tones,” J. Acous. Soc. Am 89(1), 329-340.

[CarvL91] Carver, N., Lesser, V. (1991). “A new framework for sensor interpretation: Planning to resolve sources of uncertainty,” Proc. Nat. Conf. on Artif. Intel., 724-731.
<ftp://ftp.cs.umass.edu/pub/lesser/carver-aaai91-resun.ps>

[CarvL92a] Carver, N., Lesser, V. (1992). “Blackboard systems for knowledge-based signal understanding,” in *Symbolic and Knowledge-Based Signal Processing*, eds. A. Oppenheim and S. Nawab, New York: Prentice Hall.

[CarvL92b] Carver, N., Lesser, V. (1992). “The evolution of blackboard control architectures,” U. Mass. Amherst CMPSCI tech. report #92-71.
<ftp://ftp.cs.umass.edu/pub/lesser/carver-92-71.ps>

[ChurRS94] Churchland, P., Ramachandran, V. S., Sejnowski, T. J. (1994). “A critique of pure vision,” in Large-scale neuronal theories of the brain, ed. C. Koch and J. L. Davis, Bradford Books MIT Press.

[Colb77] Colburn, H. S. (1977). “Theory of binaural interaction based on auditory-nerve data. II. Detection of tones in noise,” J. Acous. Soc. Am. 61(2), 525-533.

[ColbD78] Colburn, H. S., Durlach, N. I. (1978). “Models of binaural interaction,” in Handbook of Perception, vol. IV: Hearing, ed. E. C. Carterette, M. P. Friedman, Academic, New York.

[Colom95] Colomes, C., Lever, M., Rault, J. B., Dehery, Y. F., Faucon, G. (1995). “A perceptual model applied to audio bit-rate reduction,” J. Audio Eng. Soc. 43(4), 233-239.

[CookCG94] Cooke, M., Crawford, M., Green, P. (1994). “Learning to recognize speech from partial descriptions,” Proc. Intl. Conf. on Spoken Lang. Proc., Yokohama.

[Cooke91] Cooke, M. P. (1991). “Modeling auditory processing and organisation,” Ph.D. thesis, CS dept., Univ. of Sheffield

[CullD94] Culling, J. F., Darwin, C. J. (1994). “Perceptual and computational separation of simultaneous vowels: Cues arising from low-frequency beating,” J. Acous. Soc. Am. 95(3), 1559-1569.

[DarwC92] Darwin, C. J., Ciocca, V. (1992). “Grouping in pitch perception: effects of onset asynchrony and ear of presentation of a mistuned component,” J. Acous. Soc. Am. 91(6), 3381-90.

[deChev93] de Cheveigné, A. (1993). “Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing,” J. Acous. Soc. Am. 93(6), 3271-3290.

[DenbZ92] Denbigh, P. N., Zhao, J. (1992). “Pitch extraction and the separation of overlapping speech,” Speech Communication 11, 119-125.

[DudaLS90] Duda, R. O., Lyon, R. F., Slaney, M. (1990). “Correlograms and the separation of sounds,” Proc. IEEE Asilomar conf. on sigs., sys. & computers.

[Ellis92] Ellis, D. P. W. (1992). “A perceptual representation of audio,” MS thesis, EECS dept, Massachusetts Institute of Technology..
<ftp://sound.media.mit.edu/pub/Papers/dpwe-ms-thesis.ps.gz>

[Ellis93a] Ellis, D. P. W. (1993). “A simulation of vowel segregation based on across-channel glottal-pulse synchrony,” MIT Media Lab Perceptual Computing Technical Report #252.
<ftp://sound.media.mit.edu/pub/Papers/dpwe-asa93dnv.ps.gz>

[Ellis93b] Ellis, D. P. W. (1993). “Hierarchic Models of Hearing for Sound Separation and Reconstruction,” Proc. IEEE Workshop on Apps. of Sig. Proc. to Acous. and Audio, Mohonk.

<ftp://sound.media.mit.edu/pub/Papers/dpwe-waspaa93.ps.gz>

[Ellis94] Ellis, D. P. W. (1994). "A computer implementation of psychoacoustic grouping rules," Proc. 12th Intl. Conf. on Pattern Recognition, Jerusalem. <ftp://sound.media.mit.edu/pub/Papers/dpwe-ICPR94.ps.gz>

[Ellis95a] Ellis, D. P. W. (1995). "Hard problems in computational auditory scene analysis," posted to the AUDITORY email list. <http://sound.media.mit.edu/AUDITORY/postings/1995>

[Ellis95b] Ellis, D. P. W. (1995). "Underconstrained noisy representations for top-down models of auditory scene analysis," Proc. IEEE Workshop on Apps. of Sig. Proc. to Audio and Acous., Mohonk. <ftp://sound.media.mit.edu/pub/Papers/dpwe-waspaa95.ps.gz>

[Ellis95c] Ellis, D. P. W. (1995). "Modeling auditory organization to detect and remove interfering sounds," presented at Inst. of Acous. Speech Group meeting on Links between Speech Technology, Speech Science and Hearing, Sheffield. <http://sound.media.mit.edu/~dpwe/research/pres/shefpres-1995jan/>

[EllisR95] Ellis, D. P. W., Rosenthal, D. F. (1995). "Mid-Level Representations for Computational Auditory Scene Analysis," in working notes of the workshop on Comp. Aud. Scene Analysis at the Intl. Joint Conf. on Artif. Intel., Montréal, 111-117. <ftp://sound.media.mit.edu/pub/Papers/dpwe-ijcai95.ps.gz>

[EllisVQ91] Ellis, D. P. W., Vercoe, B. L., Quatieri, T. F. (1991). "A perceptual representation of audio for co-channel source separation," Proc. IEEE Workshop on Apps. of Sig. Proc. to Audio and Acous., Mohonk.

[FlanG66] Flanagan, J. L., Golden, R. M. (1966). "Phase vocoder," The Bell System Technical Journal, 1493-1509.

[Gaik93] Gaik, W. (1993). "Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling," J. Acous. Soc. Am. 94(1), 98-110.

[Ghitza88] Ghitza, O. (1988). "Auditory neural feedback as a basis for speech processing," IEEE Intl. Conf. on Acous., Speech & Sig. Proc., 91-94.

[Ghitza93] Ghitza, O. (1993). "Adequacy of auditory models to predict human internal representation of speech sounds," J. Acous. Soc. Am 93(4), 2160-2171.

[Gibson79] Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton-Mifflin.

[GigW94] Giguere, C., Woodland, P. C. (1994). "A computational model of the auditory periphery for speech and hearing research. II. Descending paths," J. Acous. Soc. Am. 95(1), 343-349.

[Gjerd92] Gjerdingen, R. O. (1992). "A model of apparent motion in music," Program of the 2nd Intl. Conf. on Music Percep. and Cog., UCLA.

[GodsB95] Godsmark, D. J., Brown, G. J. (1995). "Context-sensitive selection of competing auditory organisations: a blackboard model," in working notes of the workshop on Comp. Aud. Scene Analysis at the Intl. Joint Conf. on Artif. Intel., Montréal, 60-67.

[Goldh92] Goldhor, R. S. (1992). "Environmental Sound Recognition," proposal by Audiofile, Inc., to the National Institutes of Health.

[GrabB95] Grabke, J. W., Blauert, J. (1995). "Cocktail-party processors based on binaural models," in working notes of the workshop on Comp. Aud. Scene Analysis at the Intl. Joint Conf. on Artif. Intel., Montréal, 105-110.

[HallG90] Hall, J. W. 3rd, Grose, J. H. (1990) "Comodulation Masking Release and auditory grouping," J. Acous. Soc. Am. 88(1), 119-125.

[HansW84] Hanson, B. A., Wong, D. Y. (1984). "The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence of interfering speech," Proc. ICASSP-84, 195-199.

[Hart88] Hartmann, W. M. (1988). "Pitch perception and the segregation and integration of auditory entities," in *Auditory function* ed. G. M. Edelman, W. E. Gall, W. M. Cowan, chap. 21, 623-645.

[Hawley93] Hawley, M. (1993). "Structure out of sound," Ph.D. thesis, Media Laboratory, Massachusetts Institute of Technology.

[Helm77] Helmholtz, H. von (1877). *On the sensation of tone*, trans. A. J. Ellis, Dover 1954.

[Hein88] Heinbach, W. (1988). "Aurally adequate signal representation: The part-tone-time pattern," Acustica 67, 113-121.

[HewM91] Hewitt, M. J., Meddis, R. (1991). "An evaluation of eight computer models of mammalian inner hair-cell function," J. Acous. Soc. Am. 90(2), 904-917.

[HewM93] Hewitt, M. J., Meddis, R. (1993). "Regularity of cochlear nucleus stellate cells: A computational modeling study," J. Acous. Soc. Am. 93(6), 3390-3399.

[Irino95] Irino, T. (1995). "An optimal auditory filter," Proc. IEEE Workshop on Apps. of Sig. Proc. to Audio and Acous., Mohonk.

[Jeli76] Jelinek, F. (1976). "Continuous speech recognition by statistical methods," Proc. IEEE 64(4), 532-555.

[Kaern92] Kaernbach, C. (1992). "On the consistency of tapping to repeated noise," J. Acous. Soc. Am. 92(2), 788-793.

[Kaern93] Kaernbach, C. (1993). "Temporal and spectral basis for features perceived in repeated noise," J. Acous. Soc. Am. 94(1), 91-97.

[Kash95] Kashino, K., Nakadai, K., Kinoshita, T., Tanaka, H. (1995). "Organization of hierarchical perceptual sounds: Music scene analysis with autonomous processing modules and a quantitative information integration mechanism," Proc. Intl. Joint Conf. on Artif. Intel., Montréal. <http://www.mtl.t.u-tokyo.ac.jp/Research/paper/E95conferencekashino1.ps>

[KataI89] Katayose, H., Inokuchi, S. (1989). "The Kansei Music System," Computer Music Journal 13(4), 72-77.

[Keis96] Keislar, D., Blum, T., Wheaton, J., Wold, E. (1996). "Audio analysis for content-based retrieval," Muscle Fish LLC tech. report. <http://www.musclefish.com/cbr.html>

[Klatt83] Klatt, D. H. (1983). "Synthesis by rule of consonant-vowel syllables," Speech Communication Group Working Papers III, Research Lab of Electronics, M.I.T.

[KollK94] Kollmeier, B., Koch, R. (1994). "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," J. Acous. Soc. Am. 95(3), 1593-1602.

[KollPH93] Kollmeier, B., Peissig, J., Hohmann, V. (1993). "Real-time multiband dynamic compression and noise reduction for binaural hearing aids," J. Rehab. Res. & Dev. 30(1), 82-94.

[Lang92] Langner, G. (1992). "Periodicity coding in the auditory system," Hearing Research 60, 115-142.

[LazzW95] Lazzaro, J. P., Wawrzynek, J. (1995). "Silicon models for auditory scene analysis," Proc. NIPS*95 (Neural Info. Proc. Sys.). <http://www.pcmp.caltech.edu/anaprose/lazzaro/aud-scene.ps.Z>

[LazzW96] Lazzaro, J. P., Wawrzynek, J. (1996). "Speech recognition experiments with silicon auditory models," Analog Integ. Circ. & Sig. Proc., in review.

<http://www.pcmp.caltech.edu/anaprose/lazzaro/recog.ps.Z>

[LessE77] Lesser, V. R., Erman, L. D. (1977). “The retrospective view of the HEARSAY-II architecture,” Proc. 5th Intl. Joint Conf. on Artif. Intel., Los Altos, 790-800.

[LessNK95] Lesser, V. R., Nawab, S. H., Klassner, F. I. (1995). "IPUS: An architecture for the integrated processing and understanding of signals," AI Journal 77(1). <ftp://ftp.cs.umass.edu/pub/lesser/lesser-aij-ipus.ps>

[Lettv59] Lettvin, J. Y., Maturana, R. R., McCulloch, W. S., Pitts, W. H. (1959). “What the frog's eye tells the frog's brain,” Proc. Inst. Rad. Eng. 47, 1940-1951.

[Lick51] Licklider, J. C. R. (1951). “A duplex theory of pitch perception,” Experentia 7, 128-133, reprinted in *Physiological Acoustics*, ed. D. Schubert, Dowden, Hutchinson and Ross, Inc., 1979.

[Maher89] Maher, R. C., (1989). “An approach for the separation of voices in composite music signals” Ph.D. thesis, U Illinois Urbana-Champaign.

[Marr82] Marr, D. (1982). *Vision*, Freeman.

[McAd84] McAdams, S. (1984). “Spectral fusion, spectral parsing and the formation of auditory images,” Ph.D. thesis, CCRMA, Stanford Univ.

[McAuQ86] McAulay, R. J., Quatieri, T. F. (1986). “Speech analysis/synthesis based on a sinusoidal representation,” IEEE Tr. ASSP-34.

[MeddH91] Meddis, R., Hewitt, M. J. (1991). “Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification,” J. Acous. Soc. Am. 89(6), 2866-2882.

[MeddH92] Meddis, R., Hewitt, M. J. (1992). “Modeling the identification of concurrent vowels with different fundamental frequencies,” J. Acous. Soc. Am. 91(1), 233-245.

[Mell91] Mellinger, D. K., (1991). “Event formation and separation in musical sound,” Ph.D. thesis, CCRMA, Stanford Univ.

[Minsky86] Minsky, M. (1986). *The Society of Mind*, Simon and Schuster.

[Moore89] Moore, B. C. J. (1989). *An Introduction to the Psychology of Hearing*, Academic Press.

[MooreG83] Moore, B. C. J., Glasberg, B. R. (1983). “Suggested formulae for calculating auditory-filter bandwidths and excitation patterns,” J. Acous. Soc. Am. 74(3), 750-753.

[Moorer75] Moorer, J. A. (1975). “On the segmentation and analysis of continuous musical sound by digital computer,” Ph.D. thesis, Dept. of Music, Stanford University.

[NakOK94] Nakatani, T., Okuno, H. G., Kawabata, T. (1994). “Auditory stream segregation in auditory scene analysis with a multi-agent system,” Proc. Am. Assoc. Artif. Intel. Conf., Seattle, 100-107. <ftp://sail.stanford.edu/okuno/papers/aaai94.ps.Z>

[NakOK95] Nakatani, T., Okuno, H. G., Kawabata, T. (1995). “Residue-driven architecture for computational auditory scene analysis,” Proc. Intl. Joint Conf. on Artif. Intel., Montréal. <ftp://sail.stanford.edu/okuno/papers/ijcai95.ps.gz>

[NawabL92] Nawab, S. H., Lesser, V. (1992). “Integrated processing and understanding of signals,” in *Symbolic and Knowledge-Based Signal Processing*, eds. A. Oppenheim and S. Nawab, New York: Prentice Hall.

[NewS72] Newell, A., Simon, H. A. (1972). *Human Problem Solving.* Prentice-Hall.

[Nii86] Nii, H. P. (1986). “Blackboard systems part two: Blackboard application systems from a knowledge engineering perspective,” The AI Magazine 7(3), 82-106.

[OMaHM93] O'Mard, L. P., Hewitt, M. J., Meddis, R. (1993). *LUTEar: Core Routines Library manual*, part of the LUTEar software distribution. <ftp://suna.lut.ac.uk/public/hulpo/lutear/www/linklutear1.html>

[Palmer88] Palmer, C. (1988). "Timing in skilled music performance," Ph.D. thesis, Cornell University.

[Pars76] Parsons, T. W. (1976). "Separation of speech from interfering speech by means of harmonic selection," J. Acous. Soc. Am. 60(4), 911-918.

[Patt87] Patterson, R. D. (1987). "A pulse ribbon model of monaural phase perception," J. Acous. Soc. Am. 82(5), 1560-1586.

[Patt94] Patterson, R. D. (1994). "The sound of a sinusoid: Time-interval models," J. Acous. Soc. Am. 96, 1419-1428.

[PattAG95] Patterson, R. D., Allerhand, M. H., Giguère, C. (1995). "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," J. Acous. Soc. Am. 98(4), 1890-1894.

[PattH90] Patterson, R. D., Holdsworth, J. (1990). "A functional model of neural activity patterns and auditory images," in *Advances in speech, hearing and language processing vol. 3*, ed. W. A. Ainsworth, JAI Press, London.

[PattM86] Patterson, R. D., Moore, B. C. J. (1986). "Auditory filters and excitation patterns as representations of frequency resolution," in *Frequency Selectivity in Hearing*, edited by B. C. J. Moore (Academic, London).

[Pierce83] Pierce, J. R. (1983). *The science of musical sound*, Scientific American Library.

[Pick88] Pickles, J. O. (1988). *An Introduction to the Physiology of Hearing*. Academic Press.

[Port81] Portnoff, M. R. (1981). "Time-scale modification of speech based on short-time Fourier analysis," IEEE Tr. ASSP 29(3), 374-390.

[QuatD90] Quatieri, T. F., Danisewicz, R. G., (1990). "An approach to co-channel talker interference suppression using a sinusoidal model for speech," IEEE Tr. ASSP 38(1).

[QuinR89] Quinlan, J. R., Rivest, R. L. (1989). "Inferring Decision Trees using the Minimum Description Length Principle," Information and Computation 80(3), 227-248.

[RabinS78] Rabiner, L. R., Schafer, R. W. (1978). *Digital Processing of Speech Signals*, Prentice-Hall.

[Riley87] Riley, M. D. (1987). "Time-frequency representations for speech signals," Ph.D. thesis, AI Laboratory, Massachusetts Institute of Technology..

[Riss89] Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*, World Scientific.

[Rosen92] Rosenthal, D. F. (1992) . "Machine rhythm: computer emulation of human rhythm perception," Ph.D. thesis, Media Laboratory, Massachusetts Institute of Technology.

[Ross82] Ross, S. (1982). "A model of the hair cell-primary fiber complex," J. Acous. Soc. Am. 71(4), 926-941.

[Rutt91] Ruttenberg, A. (1991). "Optical reading of typeset music," MS thesis, Media Laboratory, Massachusetts Institute of Technology.

[SchaeR75] Schaefer, R., Rabiner, L. (1975). "Digital representations of speech signals," Proc. IEEE 63(4), 662-667.

[Scharf94] Scharf, B. (1994). "Human hearing without efferent input to the cochlea," J. Acous. Soc. Am. 95(5) pt. 2, 2813 (127th meeting, M.I.T.).

[Scheir95] Scheirer, E. D. (1995). “Extracting expressive performance information from recorded music,” M.S. thesis, Media Laboratory, Massachusetts Institute of Technology.

[Schlo85] Schloss, W. A. (1985). “On the automatic transcription of percussive music - from acoustic signal to high-level analysis,” Ph.D. thesis, Dept. of Music report STAN-M-27.

[Serra89] Serra, X. (1989). "A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition," Ph.D. thesis, Stanford Univ.

[ShahD94] Shahwan, T., Duda, R. O. (1994). “Adjacent-channel inhibition in acoustic onset detection,” Proc. 28th Asilomar Conf. on Sig., Sys. and Comp.

[Sham89] Shamma, S., (1989). “Spatial and temporal processing in central auditory networks” in *Methods in neuronal modelling*, MIT Press

[Shiel83] Sheil, B. (1983). “Power tools for programmers,” Datamation 131-144 (referenced in [Nii86]).

[Sieb68] Siebert, W. M. (1968). “Stimulus transformation in the peripheral auditory system,” in *Recognizing Patterns*, ed. P. A. Kolers and M. Eden, MIT Press, 104-133.

[Slaney88] Slaney, M. (1988). “Lyon's cochlea model,” Technical Report #13, Apple Computer Co.

[Slaney93] Slaney, M. (1993). “An efficient implementation of the Patterson-Holdsworth auditory filter bank,” Technical Report #35, Apple Computer Co.

[Slaney95] Slaney, M. (1995). “A critique of pure audition,” in working notes of the workshop on Comp. Aud. Scene Analysis at the Intl. Joint Conf. on Artif. Intel., Montréal, 13-18.

[SlanL92] Slaney, M., Lyon, R. F. (1992). “On the importance of time -- A temporal representation of sound,” in *Visual Representations of Speech Signals*, ed. M. Cooke, S. Beet & M. Crawford, John Wiley.

[SlanNL94] Slaney, M., Naar, D., Lyon, R. F. (1994). “Auditory model inversion for sound separation,” Proc. of IEEE Intl. Conf. on Acous., Speech and Sig. Proc., Sydney, vol. II, 77-80.

[Smith93] Smith, L. S. (1993). “Sound segmentation using onsets and offsets,” Interface Journal of New Music Research.

[SoedBB93] Soede, W., Berkhout, A. J., Bilsen, F. A. (1993). “Development of a directional hearing instrument based on array technology,” J. Acous. Soc. Am. 94(2), 785-798.

[StadR93] Stadler, R. W., Rabinowitz, W. M. (1993). “On the potential of fixed arrays for hearing aids," J. Acous. Soc. Am. 94(3), 1332-1342.

[Staut83] Stautner, J. P. (1983). “Analysis and synthesis of music using the auditory transform,” S.M. thesis, Dept. of EECS, Massachusetts Institute of Technology..

[SteigB82] Steiger, H., Bregman, A. S. (1982). “Competition among auditory streaming, dichotic fusion and diotic fusion,” Perception & Psychophysics 32, 153-162.

[StubS91] Stubbs, R. J., Summerfield, Q. (1991). “Effects of signal-to-noise ratio, signal periodicity, and degree of hearing impairment on the performance of voice-separation algorithms,” J. Acous. Soc. Am. 89(3), 1383-93.

[Suga90] Suga, N. (1990). “Cortical computational maps for auditory imaging,” Neural Networks 3, 3-21.

[SummA91] Summerfield, Q., Assmann, P. F. (1991). “Perception of concurrent vowels: Effects of harmonic misalignment and pitch-period asynchrony,” J. Acous. Soc. Am. 89(3), 1364-1377.

[Ther92] Therrien, C. W. (1992). *Decision Estimation and Classification*. John Wiley & Sons.

[vBek60] von Békésy, G. (1960). *Experiments in hearing* McGraw-Hill, reprint by the Acous. Soc. Am.

[vdMalS86] von der Malsburg, Ch., Schneider, W. (1986). “A neural cocktail-party processor,” Biol. Cybern. (54) 29-40.

[Wang95] Wang, D. (1995). “Primitive auditory segregation based on oscillatory correlation,” Cognitive Science, to appear. <ftp://ftp.cis.ohio-state.edu/pub/leon/wang95>

[Warren70] Warren, R. M. (1970) “Perceptual restoration of missing speech sounds,” Science 167.

[Warren84] Warren, R. M. (1984) “Perceptual restoration of obliterated sounds,” Psychological Bulletin 96, 371-383.

[Wein85] Weintraub, M. (1985). “A theory and computational model of auditory monaural sound separation,” Ph.D. thesis, Dept. of EE, Stanford University.

[WinN95] Winograd, J. M., Nawab, S. H. (1995). “A C++ software environment for the development of embedded signal processing systems,” Proc. ICASSP, Detroit. <ftp://eng.bu.edu/pub/kbsp/ICP/icp-icassp95.ps.gz>

[Woods95] Woods, W. S., Hansen, M., Wittkop, T., Kollmeier, B. (1995). “Using multiple cues for sound source separation,” in *Psychoacoustics, Speech and Hearing Aids*, ed. B. Kollmeier, World Scientific (in press).

[Woodf92] Woodfill, J. I. (1992). “Motion vision and tracking for robots in dynamic, unstructured environments,” Ph.D. thesis, Dept. of Comp. Sci., report # STAN-CS-92-1440.

[YoungS79] Young, E. D., Sachs, M. B. (1979). “Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers,” J. Acous. Soc. Am. 66, 1381-1403.

[Zurek87] Zurek, P. M. (1987). “The precedence effect,” in *Directional Hearing*, ed. W. A. Yost, G. Gourevitch, Springer-Verlag.

[Zwis80] Zwislocki, J. J. (1980). “Five decades of research on cochlear mechanics,” J. Acous. Soc. Am. 67(5), 1679-1685.