Computational Auditory Scene Analysis: A Representational Approach

Guy Jason Brown

Doctor of Philosophy in Computer Science

Department of Computer Science
University of Sheffield

September 1992

Computational Auditory Scene Analysis: A Representational Approach

Guy Jason Brown

ABSTRACT

This thesis addresses the problem of how a listener groups together acoustic components which have arisen from the same environmental event, a phenomenon known as *auditory scene analysis.* A computational model of auditory scene analysis is presented, which is able to separate speech from a variety of interfering noises.

The model consists of four processing stages. Firstly, the auditory periphery is simulated by a bank of bandpass filters and a model of inner hair cell function. In the second stage, physiologically-inspired models of higher auditory organization - *auditory maps* - are used to provide a rich representational basis for scene analysis. Periodicities in the acoustic input are coded by an *autocorrelation map* and a *cross-correlation map.* Information about spectral continuity is extracted by a *frequency transition map.* The times at which acoustic components start and stop are identified by an *onset map* and an *offset map.*

In the third stage of processing, information from the periodicity and frequency transition maps is used to characterize the auditory scene as a collection of symbolic *auditory objects.* Finally, a search strategy identifies objects that have similar properties and groups them together. Specifically, objects are likely to form a group if they have a similar periodicity, onset time or offset time.

The model has been evaluated in two ways, using the task of segregating voiced speech from a number of interfering sounds such as random noise, "cocktail party" noise and other speech. Firstly, a waveform can be resynthesized for each group in the auditory scene, so that segregation performance can be assessed by informal listening tests. The resynthesized speech is highly intelligible and fairly natural. Secondly, the linear nature of the resynthesis process allows the signal-to-noise ratio (SNR) to be compared before and after segregation. An improvement in SNR is obtained after segregation for each type of interfering noise. Additionally, the performance of the model is significantly better than that of a conventional frame-based autocorrelation segregation strategy.

ACKNOWLEDGMENTS

This work has benefitted from discussions with Ray Meddis (Department of Human Sciences, Loughborough University of Technology), Brian Roberts (Department of Psychology, University of York), John Holdsworth and Roy Patterson (MRC Applied Psychology Unit, Cambridge), and Steve Beet (Department of Electrical Engineering, University of Sheffield).

Special thanks must go to my supervisor, Martin Cooke, for his enthusiastic support of this work. His comments and suggestions have been invaluable. Additionally, Martin kindly allowed me to use his database of speech and noise mixtures for the evaluation in chapter 6.

I thank Phil Green for his support at every stage of this project. Thanks also to Malcolm Crawford, who knew all the secrets of the black magic box. For his support and generous assistance throughout the course of this work, Malcolm wins the Richard Dawkins prize for altruism.

This work was supported by SERC CASE award 88501484 with British Telecom Laboratories. Thanks to William Millar.

The support of my friends has been greatly appreciated. Thanks to Brian, Jane, Judy, Nick, Steve, Paul, Karen, Jim and Liz. Special thanks to Stuart for keeping me sane during the final stages of thesis writing.

Finally, I would like to thank Sara. The task of writing this thesis would have been so much more difficult without her love and support.

This thesis is dedicated to my parents,
for love, faith and finance

Contents

Chapter 1

Introduction

In most listening situations, a mixture of different sounds reaches our ears. For example, at a crowded party there are many competing voices and other interfering noises, such as music. Similarly, the sound of an orchestra consists of a number of melodic lines that are played by a variety of instruments. Nonetheless, we are able to attend to a particular voice or a particular instrument in these situations. How does the ear achieve this segregation of concurrent sounds?

Cherry [48] noted this phenomenon in 1953, and called it the "cocktail party problem". Since then, the perceptual separation of sound has been the subject of extensive psychological research. Recently, a coherent account of this work has been presented by Bregman [24]. He contends that the mixture of sounds reaching the ears is subjected to an *auditory scene analysis*, which occurs in two stages. In the first stage, the acoustic signal is decomposed into a number of sensory components. Subsequently, components which are likely to have arisen from the same environmental event are recombined into perceptual structures that can be interpreted by higher-level processes.

Although auditory scene analysis is documented comprehensively in the literature, there have been few attempts to investigate the phenomenon with a computer model. In this thesis, a computational investigation of auditory scene analysis is presented. The model draws upon psychological findings, and also exploits recent advances in the understanding of auditory physiology and anatomy.

In this first chapter, section 1.1 discusses the motivation behind the model. Section 1.2 addresses some theoretical issues that arise in the modelling of perceptual processes, and these are related to auditory scene analysis theory in section 1.3. Previous approaches to sound source segregation are described in section 1.4. Finally, the limitations of these approaches and some possible solutions are discussed in section 1.5.

1.1 Practical Applications

Undoubtedly, computational models of auditory scene analysis will contribute to the understanding of hearing. Additionally, they have a number of practical applications, which are considered below.

Automatic Speech Recognition

The need for efficient methods of communication between humans and information processing machines has stimulated research into automatic speech recognition (ASR) systems. Generally, ASR devices convert the incoming speech waveform into a series of short-term spectral estimates, which are matched with stored templates or statistical models. This approach works well in quiet acoustic environments. However, speech is normally heard in the presence of other interfering sounds, whose spectral characteristics combine with those of the speech signal. As a result of this distortion, ASR systems cannot find the correct match to the incoming speech spectrum, and the recognition performance decreases. For example, Ghitza [93] demonstrates that the error rate of a template-matching recognizer doubles at a signal-to-noise ratio (SNR) of approximately 15 dB.

In contrast, human listeners with normal hearing have no difficulty in recognizing speech at such SNR levels. This observation suggests that models of auditory processing could provide a robust front-end for automatic speech recognition in noise. So far, research in this area has concentrated on modelling the initial spectral analysis that is performed by the auditory system, with limited success (Beet [11], Robinson *et al.* [223], Ghitza [93], Hunt and Lefebvre [123]). If the performance of automatic recognizers is to approach that of human listeners, higher-level processes which underlie the ability of the auditory system to segregate concurrent sounds must also be modelled.

Virtual Reality and Visualization by Ear

The term "virtual reality" describes a collection of human-computer interface technologies that attempt to create interactive environmental simulations. Currently, the majority of virtual reality research is concerned with the generation of visual images (Rheingold [216]). However, it is clear that *auditory* images play an equally important role in the perception of an environment. Hence, models of auditory scene analysis may contribute to the design and testing of convincing virtual worlds.

A related application area is scientific visualization, which aims to provide informative representations of large data sets. Currently, visualization is dominated by the use of three-dimensional colour graphics. However, certain phenomena are difficult to display in this manner. In particular, the temporal structure of a data set cannot easily be interpreted from a sequence of visual images. Conversely, the auditory

system is highly sensitive to temporal structure. Hence, by transforming a data set into the auditory domain, it may be possible to "visualize it by ear" (Kendall [129]). Clearly, the data must be mapped onto an auditory stimulus in a meaningful manner, so that relevant changes in the data are accompanied by changes in the auditory percept. Models of auditory scene analysis may assist in the design of appropriate mappings.

Hearing Impairment

It is well known that many listeners with sensorineural hearing loss have difficulty in understanding speech, particularly in noisy environments (for example, see Festen and Plomp [85]). This condition may arise as a result of defects in many different parts of the auditory system, and is not usually corrected by the use of a conventional hearing aid. Essentially, a hearing aid amplifies both the speech and the interfering sounds, so that no improvement in the SNR is obtained (Duquesnoy and Plomp [81]). Rather, as Summerfield [265] points out,

> "The primary role of the ideal hearing aid would be to attenuate interfering noises, echoes, and the sounds of competing talkers while amplifying a target voice." (page 921)

A model of auditory scene analysis could provide the basis for such a device.

Automatic Music Transcription

Automatic music transcription systems are useful tools for the analysis of musical performance, segmentation of audio recordings and generation of manuscript (Chafe and Jaffe [45], Moorer [189], Chafe *et al.* [46]). In order to assign a pitch and a duration to the notes of each melodic line in a piece of polyphonic music, a transcription system must segregate the sounds of the different instruments. A model of auditory scene analysis could provide a means of achieving this initial segregation (Mellinger [170]).

1.2 Signals, Symbols and Marr

A central problem in the modelling of perceptual processes is how to build invariant representations of continuously varying sensory information. For example, automatic speech recognition aims to relate continuous acoustic evidence to discrete phonetic symbols. The task of bridging this representational gap has been called the *signal-to-symbol transformation* (Green *et al.* [102]).

The following section describes a theory of visual processing proposed by David Marr, in which intermediate representations are used to bridge the gap between

signal and symbols. Subsequently, the influence of Marr's representational scheme on speech and hearing is assessed, and some problems of his approach are considered.

1.2.1 Marr's Computational Theory

Over the last 15 years or so, research in machine vision has been strongly influenced by the computational theory of Marr [159, 160]. The central tenet of Marr's approach is that perceptual information processing must be understood at several levels of explanation. He called the levels the *computational theory* (function), *representation and algorithm* (process) and *hardware implementation* (mechanism). Marr argued that these levels are only loosely related, so that the explanation at one level should be quite independent of the explanation at the other two. For example, the algorithm of an information processing device should be understood independently of the hardware used to implement it.

In particular, Marr stressed the importance of the level of computational theory. Previous approaches to the understanding of vision had drawn strong conclusions from neurophysiology and psychophysics (for example, see Barlow [9]). Marr pointed out that these approaches *described* the behaviour of cells or subjects, but did not attempt to *explain* the behaviour. In contrast, his level of computational theory emphasizes the *function* of perceptual processing:

> "The nature of the computations that underlie perception depends more upon the computational problems that have to be solved than upon the particular hardware in which their solutions are implemented...an algorithm is likely to be understood more readily by understanding the nature of the problem being solved than by examining the mechanism (and the hardware) in which it is embodied." (page 27)

Marr viewed vision as a hierarchy of representational transformations, each of which makes some entity or type of information explicit. At each stage, the input and output of the representation is explicitly defined, together with the algorithm for performing the transformation. His computational framework for vision consisted of three representational stages. Initially, a rich description of intensity level changes is constructed, called the *primal sketch.* The second stage, the *$2\frac{1}{2}$-D sketch*, operates on the primal sketch to make depth and orientation of visible surfaces explicit. Finally, a *3-D model representation* is derived from the $2\frac{1}{2}$-D sketch, which describes the spatial organization of shapes and their three-dimensional structure.

Additionally, Marr [159] proposed four principles for guiding the organization of complex symbolic processes. The *principle of explicit naming* states that a collection of data which is to be described as a whole should first be given a name. This principle is central to the idea of symbolic computation, since it allows the data to be manipulated as a single entity. The *principle of modular design* states that a large computation should be implemented as a collection of small modules, which must

be as independent of one another as possible. The *principle of least commitment* states that nothing should be done which may later have to be undone. Finally, the *principle of graceful degradation* states that a system should be robust when the data is degraded, so that at least some of the required computation is delivered.

1.2.2 Representational Approaches in Speech and Hearing

In contrast with the work in machine vision, research in speech and hearing has largely failed to recognize the potential benefits of a representational approach. Currently, the majority of automatic speech recognition systems employ hidden Markov modelling and connectionist modelling techniques, which do not use an intermediate representation besides an initial spectral transformation. Indeed, some workers have proposed techniques for identifying linguistic units *directly* from the speech waveform (Nulsen *et al.* [192]). Attempts to bridge the gap between signals and symbols with such a giant leap have met with little success.

Nonetheless, a few workers have applied Marr's levels of explanation and principles of organization to automatic speech recognition. Green and Wood [103] describe an intermediate representation for acoustic-phonetic processing, which they call the *speech sketch*, after Marr. The first version of the speech sketch identified formant frequencies by LPC analysis, and described their time-frequency movement with piecewise linear contours. A semivowel recognition task was used to evaluate the system, in which the speech sketch contours were matched against descriptions of typical formant behaviour. Subsequently, Green *et al.* [100] have implemented the speech sketch in an object-oriented knowledge framework. Although the speech sketch does not attempt to model auditory processing, the authors note that an auditory representation could be used as the basis for the system. Preliminary work on such an *auditory speech sketch* is reported in Cooke and Green [53].

A similar representational approach has been described by Riley [220]. He proposed a composite symbolic description of the speech signal, called the *schematic spectrogram*, in which onsets, offsets, formant peaks and gross spectral balance were made explicit. Riley's model identified formant contours by tracking the ridges in a time-frequency representation of speech. Subsequently, the ridges were associated with acoustic correlates by using formant uniqueness and continuity constraints. Seneff [242] has described a similar *skeleton spectrogram*, which is obtained by tracking the peaks in an auditory-based "synchrony spectrum".

Recently, a number of workers have described models of auditory processing that are strongly influenced by Marr's representational approach. For example, Holdsworth *et al.* [117] have applied Marr's principle of modular design to the construction of a multi-representation auditory model. Their system consists of a simulation of auditory nerve activity, which provides the input to *tonic*, *phasic* and *coincidence* processing modules. In the tonic module, periodicities in the auditory nerve firing patterns are emphasized by a triggered temporal integration mechanism (see also Patterson *et al.* [201]). The phasic module simulates an array of cochlear nucleus

onset cells in order to detect the start of acoustic events in different frequency regions (see section 2.2.2). Finally, the coincidence module correlates onset cell responses across frequency in order to sharpen the temporal response.

A similar model has been described by Mellinger [170], in which separate modules extract onset and frequency transition information. Also, Cooke [52] has recently proposed a time-frequency representation of auditory synchrony information, which he calls *synchrony strands.* Both of these models provide the representational basis for auditory scene analysis strategies, and are discussed in section 7.4.

1.2.3 Problems with the Marrian Approach

Although the arguments in favour of the Marrian philosophy are strong, it should be noted that it has some potential problems. The first of these arises from the fact that Marr's three levels of explanation are only loosely related. As Marr [160] observes:

> "There is a wide choice available at each level, and the explication of each level involves issues that are rather independent of the other two." (page 25)

In practice, the rather unconstrained choice at each level of explanation can lead to an unprincipled model. An appropriate choice of representation is especially important, since any particular representation makes certain information explicit at the expense of other information, which may be hard to recover. The solution to this problem adopted here is to employ representations that are motivated by the known topographical organization of the higher auditory system. Effectively, the model uses representations that the auditory system itself is likely to construct. The details of this approach are given in chapter 2.

Edelman [83] has questioned the validity of information processing models of the nervous system, including Marr's approach. In particular, Edelman claims that information processing models place too strong an emphasis on the need for precise point-to-point wiring in neural structures:

> "It is surprising to observe that neurobiologists...can believe that precise algorithms are implemented and that computations and calculations of invariances are taking place inside neural structures. These beliefs persist despite the presence of the enormous structural and functional variances that exist in neural tissue - variances that would doom any equivalent parallel computer to producing meaningless output...the algorithms proposed by these workers to explain brain functions work because they have been designed to work according to ingenious and precise mathematical models...there is no evidence that they actually occur in the brain." (page 42)

Edelman's solution to this problem is the *theory of neuronal group selection.* Instead of assuming that the brain works in an algorithmic manner, the theory proposes that the nervous system functions by 'selecting' groups of neurons that respond to a particular sensory input. As such, it regards the brain as a selective system which operates in a manner resembling Darwinian natural selection. Certainly, Edelman's theory does not have the attractive simplicity of information processing models. Whether it offers a better explanation of the function of the nervous system is an open question.

It is interesting to note that the approaches of Edelman and Marr may not be completely incompatible. Certain aspects of Edelman's theory fit very naturally into a representational scheme, as discussed in section 7.5.

1.3 Auditory Scene Analysis

Over the last 20 years, experimental psychologists have developed a body of knowledge that effectively constitutes a computational theory of hearing. Recently, this effort has culminated in the publication of a comprehensive account by Bregman [24].

Bregman argues that the function of audition is to perform an *auditory scene analysis* of the mixture of sounds reaching the ear. This term describes the task of segregating the sensory components that arise from particular environmental events into separate perceptual representations. Bregman refers to these perceptual representations as *streams*, whereas the term *source* is used to denote a physical entity which gives rise to a series of acoustic events. For example, the playing of a violin is a source, whereas the mental experience of a violin playing is a stream. The phrase *acoustic event* is used to denote a single experienced event which may extend over time. In our example, each note that the violin plays is an acoustic event.

There are a number of similarities between the framework for audition proposed by Bregman and Marr's framework for vision (Williams *et al.* [280]). Firstly, both theories have their foundation in the Gestalt principles of perceptual organization (see chapter 3). However, Bregman notes that

> "Gestalt psychology did not emphasize the relevance of these principles to the practical task of scene analysis." (page 654)

as does Marr [160]. Secondly, both accounts attempt to identify the computational problem that is faced by a perceptual process. Finally, the theories of Bregman and Marr both concentrate on primitive (unlearned) grouping processes rather than schema-based (learned) mechanisms. Given these similarities, it is perhaps rather surprising that Bregman does not make any reference to Marr's influential analysis of visual processing.

In his book, Bregman identifies a number of principles that the auditory system

appears to use to group acoustic components together. For example, components which start and end at the same time are likely to be grouped into the same perceptual stream. Similarly, acoustic components which are harmonically related, have the same pattern of amplitude fluctuation or originate from the same spatial location tend to be fused. Auditory grouping principles are discussed in detail in chapter 4, and implementations of some of these principles provide the basis for the scene analysis strategy presented in chapter 5.

1.4 Previous Work

This section reviews some previous approaches to sound source segregation. Non-auditory attempts to segregate simultaneous talkers and musical sounds are discussed first, followed by approaches that are based on models of auditory processing.

1.4.1 Segregation of Simultaneous Talkers

A number of workers have described schemes for segregating the speech of simultaneous talkers, which are not based on models of auditory processing. For example, Parsons [198] proposes a *harmonic selection* technique, which aims to separate the harmonics of two competing voices. Initially, Parson's system obtains short-term spectral estimates of the acoustic input by application of a discrete Fourier transform. Subsequently, an algorithm attempts to identify two sets of harmonically-related peaks in each short-term spectrum, which are assumed to have arisen from the two voices. The pitch of each voice is computed from its harmonic series, and this information is used to maintain the continuity of the talkers over time. Finally, the system reconstructs a waveform from the harmonics of each voice, which can be assessed for intelligibility and naturalness.

Recently, Denbigh and Zhao [71] have described two techniques for segregating the speech of simultaneous talkers. In the first technique, a single microphone is used to record the voices of the two speakers. Two pitch values are identified from short-term spectral estimates of the recorded signal, as described above, and each value is allocated to the voice which has the closest average pitch. Additionally, Denbigh and Zhao's system compares adjacent spectra in order to identify the onset of a new voice. This allows the pitch tracking algorithm to lock onto weak voiced sounds, and increases the accuracy of pitch estimates when both voices are simultaneously active. Given a pitch contour for a target voice, the system extracts the harmonics of the voice using a comb filter and resynthesizes a waveform for listening tests.

This approach assumes that the two voices have different average pitch values (a male and a female speaker were used for evaluation), and fails when the pitch contours cross. In an attempt to rectify these problems, Denbigh and Zhao describe a second technique which records from two separated microphones. The pitch values are obtained from each recording as before. However, pitch continuity and directional

information (obtained from the phase spectra of the two inputs) are used to assign each pitch value to the appropriate voice. Informal testing of this system suggests that the interfering voice is almost completely removed, but the resynthesized target voice is of low quality.

Varga and Moore [272] and Moore *et al.* [188] describe a technique for segregating simultaneous speakers which matches the input signal with concurrent hidden Markov models. This approach is notable in that it employs a learned (schema-based) mechanism rather than unlearned (primitive) grouping processes (see section 3.4). In fact, the system is not limited to segregating speech, since the Markov models can be trained on other sounds. A similar technique has been proposed by Gramss and Strube [99], which uses a neural network classifier rather than statistical models.

1.4.2 Segregation of Musical Sounds

Moorer [189] describes a non-auditory technique for segregating and transcribing two simultaneous musical instruments. Initially, his system finds candidate pitches by identifying periodicities in the signal waveform. Then, the harmonics of each candidate pitch are extracted by a bank of bandpass filters. A second periodicity detector operates on the output of each filter to determine whether the harmonic is active. Subsequently, an algorithm deduces the pitches that are actually present from the harmonic information, and uses a number of constraints to assign the notes to their correct instruments. Finally, the system prints out a score in conventional music notation. Good results are obtained by Moorer on his test pieces, but the system is limited to segregating harmonic sounds and fails if the pitches are gliding. Additionally, the system cannot segregate coincident harmonics of different fundamentals, so it fails if the two instruments are playing an octave apart.

A more sophisticated transcription system has been described by Chafe *et al.* [46] and Chafe and Jaffe [45]. Initially, their system performs a high-resolution spectral transformation of the acoustic input. Individual harmonics are identified and tracked over time using information about changes in the spectrum and amplitude envelope. Subsequently, the notes belonging to each melodic line are identified from the harmonics present, using *a priori* knowledge about the characteristics of the instrument sounds (for example, their envelope shapes).

Although the system described by Chafe and Jaffe does not attempt to model auditory processing, the authors note that

> "The particular problem in polyphony is to group spectral components into sources in roughly the same fashion as the ear." (page 1291)

A recent auditory-based approach to the segregation of musical sounds has been described by Mellinger [170]. His system is discussed in the following section and in section 7.4.

1.4.3 Models of Auditory Source Segregation

A number of workers have attempted to model the perceptual segregation of double (simultaneous) synthetic vowels. Scheffers [233] investigated the role of pitch differences between two vowels in the segregation process. When the vowels were unvoiced, or both had the same fundamental frequency, listeners were able to identify them with a performance that was above chance level. Presumably, subjects were using information about spectral shape to segregate the vowels in this case. Additionally, the identification performance for voiced vowels improved if a difference in fundamental frequency was introduced. Performance increased with increasing difference in fundamental, and asymptoted at 1-2 semitones. Scheffers concluded that listeners could use differences in spectral shape and fundamental frequency to segregate simultaneous vowels.

In an attempt to model these findings, Scheffers developed a two-stage simulation of auditory processing. The first stage identified the two fundamental frequencies that provided the best fit to the peaks in a simulated auditory spectrum. Subsequently, the auditory spectrum was sampled at integer multiples of each fundamental to give an estimate of the spectral profile of each vowel. In the second stage of the model, the segregated vowel spectra were matched against stored templates. Scheffers found that the overall identification performance of the model was poorer than that of human listeners. Additionally, the model failed to show the same pattern of improvement in performance with increasing fundamental frequency separation.

Assmann and Summerfield [8] have compared four different models of double vowel identification. Firstly, a "place" scheme estimated the spectra of the two vowels from the distribution of power output across an auditory filterbank, in a similar manner to Scheffer's model. Secondly, a "place-time" model was employed, which segregated the vowels on the basis of the periodicities in each filter channel. Each of these two approaches were studied in two versions. A "linear" model used the outputs of the auditory filterbank directly, whereas a "nonlinear" version applied a compressive nonlinearity to the filter outputs. Assmann and Summerfield found that the nonlinear place-time model came close to predicting listener's performance. However, like Scheffer's model, it failed to show a gradual improvement in identification performance with increasing separation of fundamental frequency. Recently, an extension of Assmann and Summerfield's place-time scheme has been described by Meddis and Hewitt [169], which is able to reproduce the effects of fundamental frequency separation. Additionally, other models of double vowel segregation have been proposed by Gardner [88] and Summerfield *et al.* [266]. The details of the Meddis and Hewitt scheme are described later, in section 5.2.1.

The spectral characteristics of vowel sounds are constant over time, and hence schemes for double vowel segregation operate on a single auditory excitation pattern. Relatively few auditory models have been described which are able to segregate time-varying sounds. An early attempt is the system described by Weintraub [277], which uses periodicity information to segregate the voices of simultaneous speakers.

However, his model is speech-specific and uses *a priori* knowledge about the number of voices that are present. Recently, Cooke [52] and Mellinger [170] have described segregation systems that do not make strong assumptions about the number of active sound sources, or their characteristics. The models of Weintraub, Cooke and Mellinger all have some similarities with the system presented here, and they are discussed in detail in section 7.4.

Beauvois and Meddis [18] describe an auditory model in which stream segregation phenomena occur as emergent properties of low-level processing. The model is able to reproduce some simple examples of auditory stream segregation, but does not incorporate a mechanism for grouping components in different spectral regions.

1.5 Summary and Discussion

In this chapter, the concept of auditory scene analysis has been introduced. It has been argued that a representational model of auditory processing should be adopted, in analogy with the theory of visual processing proposed by Marr. The motivation for modelling auditory scene analysis has been stated, and previous auditory and non-auditory approaches to sound source segregation have been reviewed.

There have been surprisingly few attempts to model the phenomenon of auditory scene analysis, despite an extensive literature accumulated over 20 years of psychological research. Previous approaches to source segregation have seldom attempted to model auditory processing, and have been motivated by applications such as speech enhancement rather than by computational studies of hearing.

In general, previous approaches have suffered from two major limitations. Firstly, in an attempt to simplify the problem, strong assumptions have been made about the number and type of sound sources present. Automatic music transcription systems employ *a priori* knowledge of the number of instruments that are playing and their acoustic characteristics. Similarly, schemes for speech processing generally assume that the interfering source is another talker with a different average pitch. These assumptions do not hold in natural acoustic environments, where many sound sources with unknown characteristics may be active at the same time.

A second limitation of previous source segregation systems arises from the fact that they have been heavily influenced by conventional speech processing techniques. Specifically, they represent the acoustic signal as a series of short-term spectral estimates, so that no information about temporal continuity is taken into account. In fact, time and frequency are intrinsically linked in the sounds that we hear. Hence, strategies for source segregation should treat time and frequency as equally important dimensions of the acoustic signal.

The model described here addresses these problems by characterising the auditory scene as a collection of symbolic time-frequency objects. Subsequently, objects which have similar properties are identified by a search strategy and combined into explicit

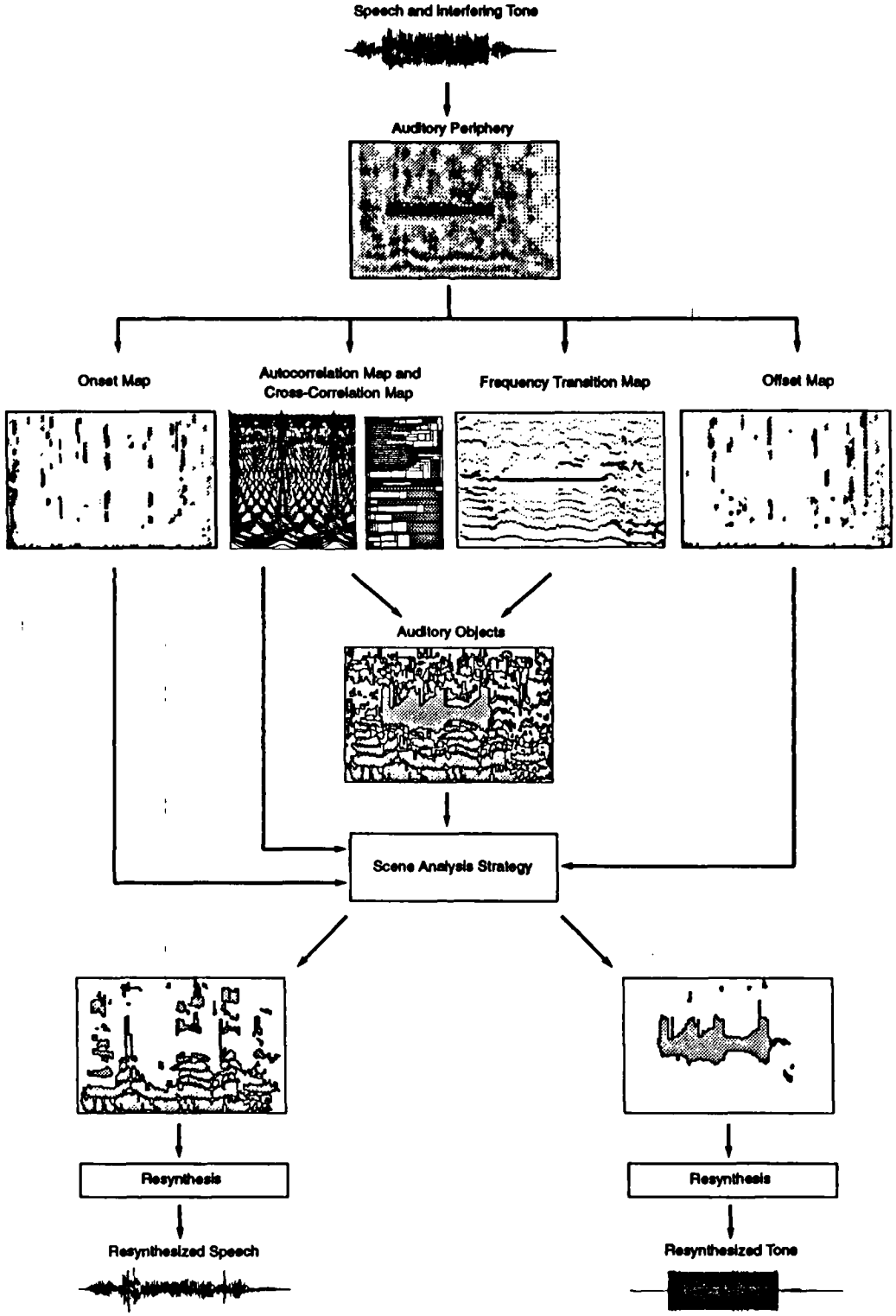

Figure 1.1: Schematic diagram of the model, showing the segregation of speech from an interfering tone. See the text for details.

groups. This approach does not make strong assumptions about the number or type of sound sources present. Additionally, the scheme presented here differs from most models of auditory processing in that it builds *multiple* representations of the auditory scene. Periodicities, frequency transitions, onsets and offsets are made explicit at an early stage of processing by separate auditory representations. Hence, *the system constructs an auditory equivalent of Marr's primal sketch.* Further, the model is physiologically principled, since the representations that are used are based on the known topographic organization of the higher auditory system.

1.6 Overview of the Thesis

A schematic diagram of the model is shown in figure 1.1, which illustrates the auditory representations that are used at each stage of processing.

In the following chapter, the structure and function of the auditory system is reviewed, with particular emphasis on higher levels of the auditory pathway. It is argued that physiological studies of orderly topographic arrays of neurons - *auditory maps* - provide a better means of explaining auditory function than studies of single cells. Chapter 3 reviews those aspects of auditory scene analysis theory that are relevant to the remainder of the thesis.

Chapter 4 describes the auditory map representations that are used in the model, together with the physiological and psychophysical evidence that supports them. A model of the auditory periphery is presented, which provides the input to the auditory maps. In chapter 5, a technique for forming symbolic *auditory objects* from the map representations is presented. A strategy for searching the auditory scene is described, which groups objects that have a similar pitch contour, onset time or offset time. The model is evaluated in chapter 6, using two methodologies. Firstly, a waveform can be resynthesized from a group of objects, allowing the results of the segregation process to be assessed in listening tests. Secondly, a technique for comparing signal-to-noise ratios before and after segregation is presented, which allows the performance of the model to be quantified in an intuitive manner.

Chapter 2

Structure and Function of the Auditory System

In this chapter, the physiology and anatomy of the auditory system is reviewed. The discussion concentrates on those aspects of higher auditory structure and function that are relevant to the following chapters of this thesis, and is not intended to be exhaustive.

Section 2.1 describes the anatomy and physiology of the auditory periphery, including the response properties of auditory nerve fibres. Section 2.2 reviews the higher auditory system, and describes two possible mechanisms by which higher centres may detect features in the sensory input.

2.1 Auditory Periphery

The peripheral part of the auditory system extends as far as the auditory nerve, where neural activity is initiated. Functionally, the auditory periphery can be divided into *outer*, *middle* and *inner* ears. The structure and function of each component is briefly reviewed below. Comprehensive discussions can be found in Pickles [205], Russell [229], Palmer [196] and Wilson [281].

2.1.1 Outer and Middle Ears

The outer ear consists of the pinna (the part we actually see), together with an *external auditory canal* which leads to the *tympanic membrane* (eardrum). The external auditory canal forms an acoustic resonator, which increases the sound pressure at the tympanic membrane between frequencies of 2 kHz and 7 kHz.

The motion of an object in the environment gives rise to pressure changes in air, which travel down the external auditory canal and cause the tympanic membrane to vibrate. These vibrations are transmitted to the *oval window* by three small bones (*ossicles*), which comprise the middle ear. The ossicles act as an energy-coupling mechanism, which matches the low impedance at the tympanic membrane to the higher impedance of the cochlear fluids.

2.1.2 Inner Ear

The inner ear consists of a tapered, spirally-coiled tube called the *cochlea*. Two structures, Reissner's membrane and the *cochlear partition*, divide the cochlea into three fluid-filled chambers along its length. The cochlear partition consists of a thin *basilar membrane*, together with the *organ of Corti*. The basilar membrane varies in stiffness and width along its length, being stiffest and narrowest at the base of the cochlea.

Vibration of the ossicles distorts the oval window, and causes movement of the incompressible cochlear fluids. These pressure changes interact with the varying stiffness of the basilar membrane to produce a *travelling wave*, which is propagated along the length of the membrane. In the case of a pure tone stimulus, a sharp peak occurs in the travelling wave at a distance along the basilar membrane related to the frequency of the stimulus. High frequency tones produce a peak towards the basal end of the basilar membrane, and low frequency tones produce a peak towards the apical end. Hence, the basilar membrane appears to perform a spectral analysis, in which frequency is converted to a place representation.

The frequency-sensitive movements of the basilar membrane are converted into nerve impulses by *hair cells*, which form part of the organ of Corti. Hair-like processes, called *stereocilia*, extend from the cells and couple with the *tectorial membrane*, which covers the organ of Corti. Movement of the basilar membrane generates a shearing action between the organ of Corti and the tectorial membrane, causing the stereocilia to bend. These deflections modulate the membrane potential of the hair cells, causing depolarization and release of a *neurotransmitter*. The neurotransmitter diffuses across a short gap to an auditory nerve fibre, where it can initiate a nerve impulse.

Hair cells are divided into two groups, known as *inner* and *outer* hair cells. Although the outer hair cells are more numerous, the large majority of auditory nerve fibres innervate inner hair cells. The function of outer hair cells is uncertain, but they may contribute to the frequency selectivity of the basilar membrane by actively amplifying the travelling wave (Ashmore [7]).

2.1.3 Responses of Auditory Nerve Fibres

The properties of auditory nerve fibre responses have been the subject of extensive research, initially with simple stimuli such as pure tones and clicks (Kiang *et al.* [130]) and more recently with speech (Delgutte and Kiang [66, 67, 68, 69, 70]). Here, some of the important properties are reviewed.

Average Rate Response

One method of quantifying the response of an auditory nerve fibre is to measure its *average rate*, which is obtained by counting the number of nerve impulses that occur within a particular time period. In the absence of any stimulation, auditory nerve fibres discharge at a *spontaneous rate*. When stimulated, the response of a fibre remains at the spontaneous rate unless the stimulus intensity exceeds the fibre's *threshold*. Auditory nerve fibres with a high threshold have a low spontaneous rate, and vice versa. Above threshold, the average rate of a fibre increases approximately linearly with the intensity of the stimulus, until the discharge rate becomes *saturated*. Beyond this point, further increases in stimulus intensity do not give concomitant increases in average rate.

Liberman [147] suggests that auditory nerve fibres should be divided into three populations on the basis of their spontaneous rates. He proposes the categories *high* (spontaneous rates greater than 18 spikes/sec), *medium* (rates between 0.5 and 18 spikes/sec) and *low* (rates less than 0.5 spikes/sec). High spontaneous rate fibres account for the majority of Liberman's sample (61%), with the medium (23%) and low (16%) spontaneous rate fibres being less abundant.

Frequency Selectivity

The frequency selectivity of an auditory nerve fibre can be determined by measuring its threshold as a function of the frequency of a tonal stimulus. A *tuning curve* obtained in this manner indicates that the fibre has a low threshold at a particular frequency, called the *characteristic frequency*. The characteristic frequency of an auditory nerve fibre is closely related to the position of its inner hair cell on the cochlear partition, and the sharpness of tuning is similar to that of the basilar membrane if active mechanics are assumed.

Temporal Response

The response of an auditory nerve fibre to a sustained stimulus, such as a tone burst, is not constant over time. After an initial peak of activity at the onset of the tone, the average rate decreases rapidly for 10-20 ms, and then decreases gradually towards a steady state. This phenomenon is known as *adaptation*. At the offset

of the stimulus, activity falls below the spontaneous rate, and then returns to the spontaneous rate after a brief recovery period. This pattern of activity is similar to that shown on the far left of figure 2.1.

So far, the response of auditory nerve fibres has only been considered in terms of their rate of firing. However, information about the stimulus is also carried in the *fine time structure* of auditory nerve responses. Below stimulus frequencies of 4-5 kHz, auditory nerve fibres tend to synchronize to a particular phase of the stimulating waveform. This phenomenon, called *phase-locking*, arises because of two factors. Firstly, the inner hair cells only initiate nerve firings during upward deflections of the basilar membrane. Secondly, the likelihood of firing is greatest when the upward deflection of the basilar membrane is maximal.

Phase-locking is maintained over a wide dynamic range. When an auditory nerve fibre is stimulated by a low-frequency tone at an intensity below its threshold, the spontaneous discharges of the fibre exhibit phase-locking. Similarly, phase-locking is preserved when the average rate of the fibre is saturated. Hence, it is possible that the auditory system uses fine time structure to code certain stimuli at high intensities. For example, Sachs and Young [231, 288] have investigated the representation of vowel sounds in the auditory nerve, in terms of average rate and phase-locked responses. They found that the average rate response saturated at medium and high intensities for the majority of auditory nerve fibres, so that the peaks in a vowel spectrum were poorly defined. In contrast, phase-locking to the frequency components of a vowel was preserved at high stimulus intensities. Note, however, that auditory nerve fibres with low spontaneous rates did preserve the spectral profile in their average firing rates, even up to the highest intensities used.

Masking

In the auditory nerve, one stimulus may obscure or reduce the response to another. This phenomenon is known as *masking*. For example, if the firing rate of an auditory nerve fibre is saturated in response to one stimulus, a superimposed stimulus will not be able to increase the firing rate any further (Smith [255]). This effect is rather like the visual phenomenon of occlusion (see section 3.1). Masking in the auditory nerve is also demonstrated by the phenomenon of *two-tone suppression*. In this effect, the average rate of an auditory nerve fibre that is responding to a tone can be reduced by a second tone with an appropriate frequency and intensity. There is good evidence that this phenomenon arises as a result of nonlinearities in the basilar membrane mechanics (Patuzzi *et al.* [202]).

2.2 Higher Auditory System

This section presents an overview of the anatomy and physiology of the higher auditory system, with an emphasis on possible mechanisms of feature detection.

Comprehensive reviews can be found in Moore [181], Aitken [5], Hackney [105], Palmer [196] and Irvine [124]. Again, early research on the higher auditory system employed simple stimuli such as pure tones (for example, see Pfeiffer [203]). More recently, complex sounds such as speech have been used (Blackburn and Sachs [20]).

2.2.1 General Anatomy

The higher auditory system projects from the cochlea to the auditory cortex through a series of *nuclei*, which process and relay the neural information in parallel. At each stage of the system there are nuclei for the left and right ears, and most ascending neurons may connect with nuclei on the opposite side as well as with nuclei on the same side. These bilateral connections provide the facility for interaural comparisons.

At all stages of the higher auditory system, descending pathways run parallel to the ascending tracts. The function of these projections is poorly understood, and they are not considered in this review. However, it is interesting to note that the scheme for reentry of auditory information discussed in section 7.5 requires ascending and descending neural pathways.

Auditory nerve fibres projecting from the organ of Corti enter the first auditory nucleus of the brainstem, the *cochlear nucleus*. This nucleus is divided into *anteroventral (AVCN)*, *posteroventral (PVCN)* and *dorsal (DCN)* regions, on the basis of differences in the distribution of cell types. Generally, cells of the AVCN have properties that are similar to those of auditory nerve fibres, and may function as a simple relay to higher centres. In contrast, neurons of the DCN have complex response properties, and may play a role in enhancing spectral or temporal information. The properties of neurons in the PVCN are intermediate to those of the other two regions.

Ascending fibres from the AVCN and DCN enter the *superior olivary complex (SOC)*. This nucleus is the lowest in the brainstem that receives input from both ears. Hence, it probably plays a role in sound localization.

The principal nucleus of the auditory midbrain is the *inferior colliculus*, which receives input from the SOC, DCN and PVCN. As such, it appears to combine the complex signal analysis of the cochlear nucleus with the sound localizing ability of the SOC. Hence, the inferior colliculus may simultaneously code the complexity of sounds, and their location in space.

Neurons from the inferior colliculus project via the *medial geniculate body* to the *auditory cortex*. The response properties of cortical cells are very complex, and probably underlie high-level functions such as auditory memory, discrimination of temporal patterns, attention and source segregation.

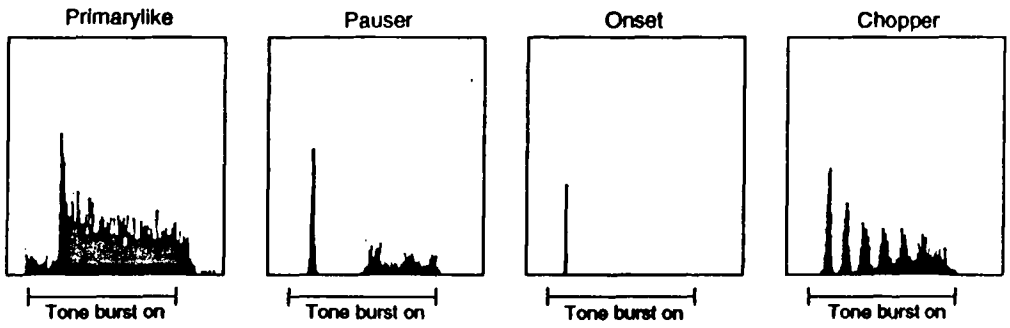

Figure 2.1: Classification of cochlear nucleus neurons on the basis of their temporal response. Time is represented on the abscissa, and average rate is represented on the ordinate. Adapted from Pfeiffer [203].

2.2.2 Responses of Single Cells

The properties of neurons in the higher auditory system differ in two important respects from those of auditory nerve fibres. Firstly, auditory nerve fibres always respond to a tone presented above threshold with an increase in firing rate (*excitatory* response). In contrast, tonal stimuli can reduce or abolish the spontaneous activity of many higher auditory neurons (*inhibitory* response). Secondly, whereas single auditory nerve fibres have qualitatively uniform properties, cells of the higher auditory system exhibit a multitude of different morphologies and response patterns.

In the visual system, neurons appear to respond preferentially to certain features of a stimulus, such as lines and edges of a particular orientation (Hubel and Weisel [120]). Higher auditory neurons with particular response patterns may act as specific *feature detectors* in a similar manner. Hence, a number of workers have attempted to classify the responses of higher auditory neurons, and to relate each response type to a cell morphology and functional role. Two classification schemes are considered below.

Classification by Pattern of Temporal Response

One of the earliest schemes for classifying higher auditory neurons is found in the work of Pfeiffer [203]. He classified cells of the cochlear nucleus on the basis of the pattern of their temporal response to brief tone bursts, delivered just above threshold at the characteristic frequency of the neuron. Pfeiffer originally identified four response types, and gave them the descriptive names *primarylike*, *pauser*, *onset* and *chopper* (see figure 2.1). Primarylike cells have a pattern of temporal response similar to that of auditory nerve fibres, with an initial peak of activity at stimulus onset which adapts to a lower rate. Pauser cells show an initial onset response, followed by a period of no activity and then a gradual increase in average rate. Onset cells produce a sharp peak in activity at the start of the stimulus, and then either no

activity or a low level of sustained activity. Chopper cells fire repetitively during a tone burst, at a rate that is unrelated to the period of the stimulating waveform. Since Pfeiffer's original classification, a number of additional response types have been described, including *primarylike-with-notch*, *buildup*, *negative responder*, *extraordinary*, *off* and *on-off* (Godfrey *et al.* [95], Abeles and Goldstein [1], Shofner and Young [250], Blackburn and Sachs [19]).

Attempts to correlate these response patterns with morphological cell types have met with mixed success. In the cochlear nucleus, primarylike responses appear to originate from the spherical bushy cells of the AVCN (Rhode *et al.* [217]). Similarly, chopper responses probably originate from stellate cells in the AVCN and PVCN (Rhode *et al.* [217]), and onset responses appear to be generated by the octopus cells of the PVCN (Rouiller and Ryugo [228]). However, it has proved difficult to match response patterns with cell types in areas beyond the cochlear nucleus. For example, Ryan and Miller [230] have identified single cells in the inferior colliculus which can show onset, primarylike and pauser responses depending on the intensity of the stimulus. Hence, classification by Pfeiffer's scheme can be potentially misleading. This problem is compounded by the fact that the response of a neuron to a time-varying stimulus cannot generally be predicted from its response to a static tone.

Similarly, it has generally proven difficult to associate a functional role with a particular response pattern. The properties of primarylike units in the cochlear nucleus resemble those of auditory nerve fibres, suggesting that they act as a simple relay to higher centres. Chopper units appear to enhance particular rates of amplitude modulation (Kim *et al.* [132]). However, many diverse functions have been proposed for other response types. For example, it has been suggested that onset units may code periodicity, intensity or interaural onset time disparities (Møller [177], Young *et al.* [289], Rhode and Smith [218]). Clearly, the extent to which different response types code particular features is currently rather uncertain.

Classification by Pattern of Excitation and Inhibition

Evans and Nelson [84] have proposed a scheme for classifying higher auditory neurons according to their excitatory and inhibitory properties, which has subsequently been developed by other workers (Young [287], Shofner and Young [250]). Neurons are classified by their *response areas*, which indicate the distribution of excitatory and inhibitory responses as a function of stimulus frequency and intensity (see figure 2.2). In the cochlear nucleus, five main response area types have been described (*Type I, II, III, IV* and *V*), together with two other response types which do not readily fit into this classification scheme (*Type I/II* and *II/III*).

Correlation of response areas with morphological cell types has been reasonably successful. Type I units have a single excitatory response area without any inhibitory regions, and probably correspond to spherical bushy cells in the AVCN (the primarylike units of Pfeiffer). Similarly, Type III responses probably originate from stellate cells of the PVCN. However, there is a many-to-one relationship between

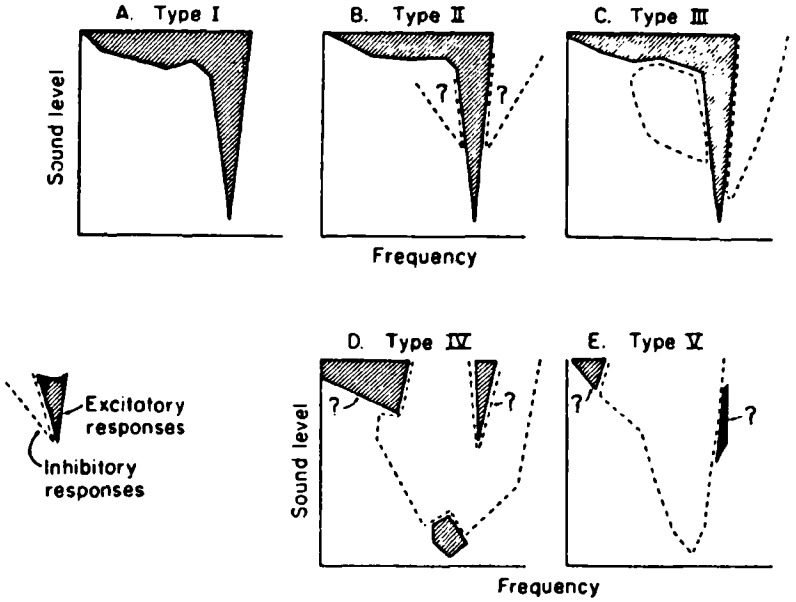

Figure 2.2: Classification of cochlear nucleus neurons on the basis of their excitatory and inhibitory response areas. Question marks indicate variable or uncertain features. From Young [287].

certain cell types and response areas. For example, globular bushy cells and octopus cells of the PVCN may both have a Type I/III response area (Palmer [196]).

The prediction of the function of a neuron from its response area is fraught with difficulties. For example, the strong inhibitory regions of Type IV units might suggest that the cells would respond preferentially to narrowband stimuli, such as pure tones. In fact, Type IV units respond very well to broadband noise (Voigt and Young [274], Spirou and Young [257]). Clearly, it is not currently possible to assess the role of cells with complex response areas in the detection of complex features.

2.2.3 Auditory Maps

So far, it has been assumed that single higher auditory neurons are responsible for detecting features in the sensory input. However, the detection of a feature could also be defined by the pattern of activity over a *group* of neurons. The difference between these two organizations is analogous to the difference between a photograph and a hologram (Pickles [205]). In a photograph, each point represents one point in space, whereas in a hologram each point represents many points in space, and a single point in space can only be reconstructed by the integration of information from many points in the hologram. Certainly, many of the more complex features will only be represented in the latter form. Indeed, complex features that are not represented by the activity of single cells *must* be represented by the activity of many cells.

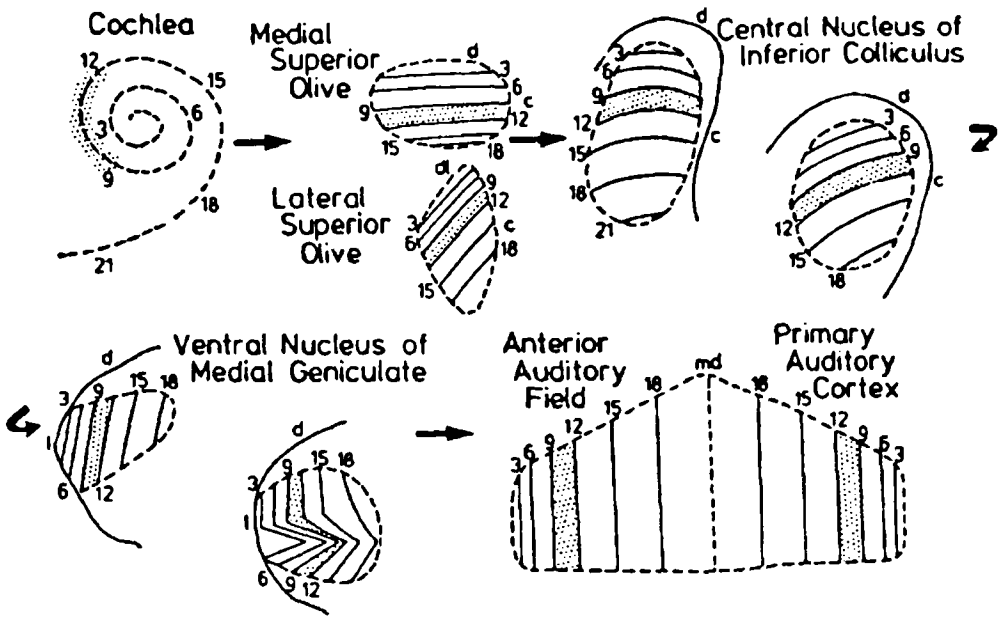

Figure 2.3: Cochleotopic organization in the higher auditory system of the cat. Each drawing represents a section along the organ of Corti superimposed onto schematic views of a higher auditory nucleus. From Moore[181].

In fact, there is good evidence that the functional units of neural processing are groups of cells, rather than single cells. A recurring motif in neurophysiology is the *map*, a term which describes an array of neurons that are systematically tuned for a particular parameter value (Knudsen *et al.* [138]). Some of these maps appear to perform neural computations. Such *computational maps* transform the representation of information into a place-coded probability distribution, so that the computed value of a parameter is represented by the location of maximum activity in the neural array. In contrast, many maps are not computational, and simply reproduce the peripheral representation of information at higher centres.

The majority of computational maps discovered so far appear to process sensory information. For example, line-orientation sensitive cells in the visual cortex are mapped in an orderly manner (Hubel and Weisel [120]). Similarly, maps have been identified in the higher auditory system which appear to code frequency, intensity, spatial location and complex sounds. The physiology of these *auditory maps* is discussed below.

Cochleotopic Maps

The existence of a precise mapping of the cochlea at every level of the higher auditory system has been recognised for some time (Whitfield [278]). This *cochleotopic* organization is characterised by the orderly arrangement of neurons into a systematic progression of characteristic frequency. As such, cochleotopic organization is

an example of a non-computational neural map. Cochleotopic arrangement in the higher auditory system is summarised in figure 2.3.

Within a cochleotopic framework, neurons of similar characteristic frequency are organized into sheets called *iso-frequency laminae* (Clopton [50]). These laminae may take the form of flat planes, or may be arranged in concentric rings rather like the layers of an onion. Hence, cochleotopic organization implies the projection of a one-dimensional structure (the organ of Corti) onto two-dimensional sheets of cells. There is good evidence that the second dimension in higher auditory centres is used to represent other acoustic parameters, in a framework that is orthogonal to the frequency plane (Moore [181]). Clearly, this arrangement gives rise to *computational* auditory maps. A number of these maps are discussed on the following pages.

Maps of Intensity

Recall from section 2.2.2 that many neurons in the higher auditory system have inhibitory regions in their response areas, which may or may not be activated depending on the intensity of a stimulus. Hence, it is possible to define a *best intensity* for these neurons, which is the stimulus intensity at which the maximal average rate is observed.

There is strong evidence for a topographic representation of best intensities in the auditory cortex of the echo-locating bat (Suga and Manabe [263]). Frequency is represented cochleotopically within concentric laminae, and intensity is represented on a circular axis. This arrangement is illustrated in figure 2.4. Whether a map of intensity is found outside of the cortex or in other species is presently uncertain, although there is some evidence for a similar map in the auditory cortex of the cat (Phillips *et al.* [204]).

Maps of Complex Sounds

Many cells in the higher auditory system respond preferentially to certain rates of frequency and amplitude modulation. In some nuclei, these cells appear to be topographically organized.

Schreiner and Langner [237] have described a map of periodicity in the inferior colliculus of the cat, in which neurons are systematically tuned to a particular rate of periodic amplitude modulation. Similar maps may exist in the auditory cortex of the cat (Schreiner and Urbas [239]) and the midbrain of Guinea fowl (Langner *et al.* [144]). These maps are discussed in detail in section 4.2.2. Additionally, Shamma and Chettiar [247] and Mendelson *et al.* [171, 172] have identified maps of preferred frequency sweep rate in the auditory cortex. These maps may play a role in the coding of spectral shape, and are discussed in section 4.4.2.

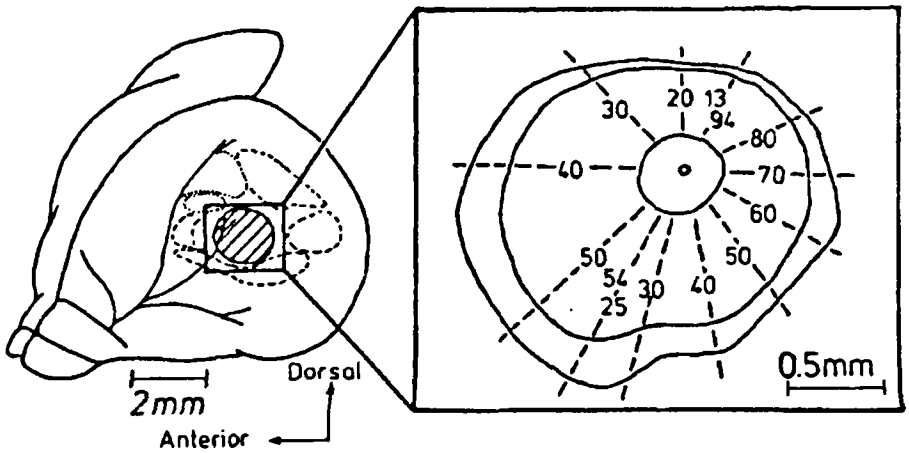

Figure 2.4: Representation of best intensities in the auditory cortex of the echo-locating bat. In the mapped area (right), iso-frequency laminae are shown by solid lines, and iso-intensity contours are shown by dashed lines. Numbers indicate the best intensity, in dB, associated with each circular contour. From Moore[181], re-drawn from Suga and Manabe [263].

Maps of Auditory Space

At the level of the SOC and above, the majority of auditory neurons are influenced by stimulation of either ear (Wise and Irvine [282]). Some of these binaurally-responsive neurons are sensitive to *interaural time differences (ITDs)*, and are thought to encode the spatial location of low frequency sounds. Others are sensitive to *interaural intensity differences (IIDs)*, and may encode the spatial location of high frequency sounds (Hirsch *et al.* [115]).

A very accurate map of auditory space has been identified in the midbrain of the barn owl (Knudsen [137], Knudsen and Konishi [139]). Neurons sensitive to IIDs and ITDs are systematically arranged in separate computational maps, which encode elevation and azimuth respectively. Subsequently, these two maps are combined into a higher-order map of auditory space. In mammals, maps of auditory space appear to be less well defined. However, there appear to be maps of ITD in the guinea pig (King and Palmer [134]) and cat (Bojanowski and Schwarz [22]). Additionally, King and Hutchings [135] have identified a map of auditory space in the midbrain of the ferret.

General Properties of Auditory Maps

All computational maps, including the auditory maps described in this chapter, share a number of fundamental properties (Knudsen *et al.* [138]). Here, the implications of these shared properties are discussed.

Firstly, neurons in a map tend to respond to a broad range of parameter values, although their tuning curves are peaked. For example, neurons in the map of line orientation in the visual cortex respond weakly to almost any orientation, but they respond maximally to a small range of orientations (Schiller *et al.* [235]). Additionally, maps are highly redundant, and contain many neurons that are tuned to the same parameter value. This redundancy may play an important role in learning, as discussed in section 7.5.

Secondly, the neurons of a computational map perform preset computations, in parallel, on the incoming signal. Hence, maps do not require input from higher centres in order to perform their computations, although their behaviour may be modified by descending influences (Middlebrooks and Knudsen [173]). Note that "preset" does not necessarily imply that the maps are genetically hardwired. The basic pattern of connectivity in a map is undoubtably genetically determined, but details can be modified by experience. For example, visual experience contributes to the formation of line-orientation maps in the visual cortex (von der Malsburg and Cowan [158]). Interestingly, maps in one sensory system can be modified by inputs from a different sensory system (Harris [111]). For instance, the development of a map of auditory space can be influenced by visual inputs (King and Moore [133]).

Finally, maps appear to represent parameters, and a range of parameter values, that are functionally significant. For example, Knudsen *et al.* [138] suggest that

> "The generation of a map indicates that the parameter is being evaluated at that particular site in the nervous system, and that the value of the parameter is crucial for subsequent processing." (page 57)

A similar conclusion has been reached by Schreiner and Langner [237]:

> "It could be argued that the identification of a systematically represented response parameter in a given nucleus is evidence that the 'mapped' parameter represents a dimension of perceptual space or contributes to the selection of combinations of information bearing parameters in a subsequent establishment of dimensions of perceptual space." (page 1835)

In an auditory context, we might expect parameters to be mapped that are important for tasks such as pitch analysis and source segregation. The following chapters of this thesis suggest that this is indeed the case.

Advantages of Mapping

It is advantageous to perform computations in neural maps for a number of reasons (Knudsen *et al.* [138]). Firstly, the nervous system requires fast strategies for

processing large amounts of sensory information. Maps, which perform preset computations rapidly in parallel, are ideally suited for such as task. Furthermore, the results are presented in a simple, systematic form.

Secondly, when parameter values are represented in a computational map, further processing can be based on relatively straightforward schemes of connectivity. For example, the map of auditory space in the barn owl is obtained by the simple convergence of maps of ITD and IID. Additionally, the output of any map is always represented as the location of a peak of activity within an array of neurons. Hence, the nervous system can employ a single strategy for reading the information, and can combine information from different modalities in a straightforward manner.

Finally, when a parameter is represented in topographic form, other neural mechanisims can operate to sharpen or modify the response pattern. For example, lateral inhibition can only be applied to mapped information, such as a cochleotopic array (Shamma [245]). Similarly, the general form of auditory maps allows the comparison of particular parameter values at different characteristic frequencies. For example, section 4.2.6 describes a scheme in which adjacent channels of a periodicity map are compared using a simple cross-correlation mechanism.

2.3 Summary and Discussion

In this chapter, the basic physiology and anatomy of the auditory system has been reviewed. Additionally, two possible mechanisms for feature detection in the higher auditory system have been considered. Firstly, features may be represented by the activity of single neurons. Secondly, features may be represented by the pattern of activity over an orderly topographic array of neurons, called a *computational map.*

Levels of Explanation

It is clear from section 2.2.2 that schemes which classify neurons according to the pattern of their temporal or excitatory/inhibitory responses can be misleading. Similarly, the prediction of a feature detecting function for a single neuron on the basis of its response properties has proven to be difficult. As Pickles [205] observes,

> "There seems to be a continuum of response characteristics along every category of response that has been analysed. It is not therefore certain that we are justified in asserting that the degree to which any one complex feature is extracted results from anything other than a random arrangement of excitation and inhibition on the constituent neurons." (page 179)

One possible solution to this problem is to model the detailed physiology of the higher auditory system without attributing functional roles to particular neuron

types. Then, the ability of the model to code certain features can be determined empirically. For example, Pont and Damper [208, 209] have described a model of the DCN, in which models of particular cell types are connected according to a detailed physiological "wiring diagram". Apparently, the model makes a categorical distinction between synthetic speech stimuli with different voice onset times, in the same way as human listeners.

Again, this approach suffers from a number of potential problems. Firstly, the auditory nervous system is highly complex and subject to considerable variation between individuals. Hence, it may never be possible to understand it in terms of a detailed "wiring diagram". Secondly, this problem is exacerbated by the fact that anatomical and physiological data are obtained from animals such as the cat, rather than from the human auditory system. For example, Adams [2] and Moore [182] find substantial differences between the detailed anatomy of the cochlear nucleus in humans and cats. Adams concludes that:

> "Until more evidence becomes available, it is not reasonable to speculate on the functions or targets of cells in the human cochlear nucleus. It is reasonable to assume, however, that the physiology of this nucleus in humans is quite different from that of cats, given the marked difference in the composition of the nucleus between the species." (page 1260)

Clearly, this questions the validity of using a "wiring diagram" model of the auditory system to explain human psychophysical phenomena. Finally, detailed physiological models are generally based on incomplete information. For example, the cochlear nucleus model of Pont and Damper gives realistic responses to pure tones and noise, but has not been calibrated with modulated stimuli. Recall from section 2.2.2 that the response of a higher auditory neuron to a time-varying sound cannot usually be predicted from its response to a static sound. Nonetheless, Pont and Damper's model is used to process speech, which is highly modulated in time and frequency.

These criticisms suggest that the auditory system should be modelled in terms of general principles of *functional* organization. Clearly, auditory maps are an important source of information for such an approach, since they indicate which parameters are likely to be functionally significant, and how the parameters are organized. Additionally, the topography of a map may indicate the mechanism by which the map performs its computation (Knudsen *et al.* [138]).

Maps and Marr

The arguments made in the previous section are similar to those underlying Marr's [160] computational approach to vision (see section 1.2.1). Studies that attempt to ascribe feature detecting roles to neurons on the basis of their physiological properties ignore the questions that Marr stresses in his level of computational theory, namely *"what is the goal of the computation?"* and *"why is it appropriate?"*. Additionally, Marr stresses the need to understand the function of an information

processing device independently of the representations, algorithms and hardware that it uses. Effectively, single unit studies jump straight to the level of hardware, without considering the other levels of explanation. Here, we emphasize the role that the neurons of the higher auditory system play in providing the functional basis for a sensory analysis, rather than their individual feature detecting properties or detailed physiology.

A potential problem with the Marrian approach is that the number of possible representations is large, and the choice of representation is rather unconstrained (see section 1.2.3). Marr [160] notes that neurophysiology can inform the choice of representation, as long as the underlying computational theory is well understood:

> "Neurophysiology...can help us to understand the type of representations being used...But one has to exercise extreme caution in making inferences from neurophysiological findings about the algorithms and representations being used, particularly until one has a clear idea about what information needs to be represented and what processes need to be implemented." (page 26)

Clearly, computational maps indicate the type of representations that the auditory system is using and, together with psychological investigations, the kind of information that is being represented. Hence, computational maps fit naturally into the Marrian philosophy.

Chapter 3

Principles of Auditory Scene Analysis

The term *auditory scene analysis* describes the ability of the auditory system to segregate the acoustic components that arise from different events in the environment into separate perceptual representations (see section 1.3). In Bregman's [24] terminology, the physical entity which gives rise to a series of acoustic events is called a *source*, and the perceptual representation of the events is called a *stream*.

This chapter presents an overview of some of the principles underlying auditory scene analysis. The review is not intended to be exhaustive. Rather, it describes only those aspects of auditory scene analysis theory which are relevant to the following chapters of this thesis. The reader is directed to the book by Bregman for a comprehensive account.

3.1 Principles of Perceptual Organization

In the early part of this century, the Gestalt psychologists (for example, Koffka [140]) formulated a theory describing many of the principles of perceptual organization. The German word "Gestalt" means "pattern", and the theory proposed a number of rules governing the manner in which the brain forms mental patterns from elements of its sensory input. Although the Gestalt principles of perceptual organization were generally described first in relation to vision, they are equally applicable to audition. Here, a number of Gestalt principles are reviewed, and examples from the auditory and visual domains are given.

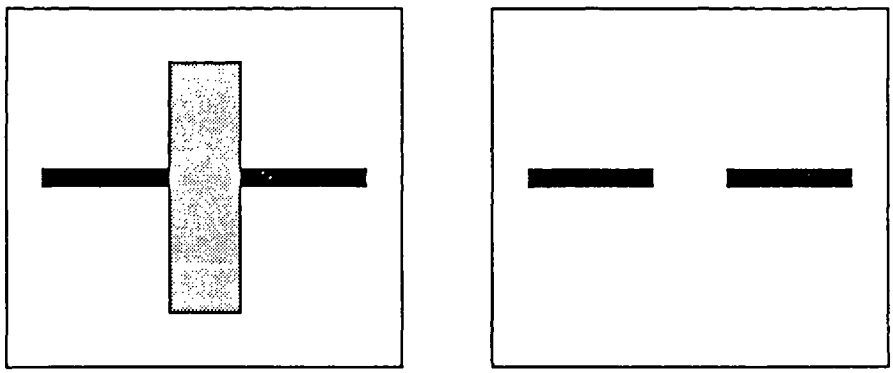

Figure 3.1: Illustration of the Gestalt principle of closure. On the left, there is a tendency to close (complete) the black line behind the grey rectangle. On the right, closure does not occur because there is no evidence that the line is occluded, rather than interrupted.

Common Fate

The Gestalt principle of common fate describes the tendency to group sensory elements which change in the same way at the same time. A visual illustration of this principle is the demonstration by Johansson [126] that, in a collection of randomly moving dots, two dots which have a correlated trajectory are perceived as parts of the same object. In this example, the common motion of the two dots promotes their perceptual fusion.

The principle of common fate can equally be applied to audition, since frequency components which belong to the same acoustic source tend to vary in a coherent manner. Specifically, they tend to start and stop together (common onset and common offset), change in amplitude together (common periodicity and common amplitude modulation) and change in frequency together (common frequency modulation). These factors are discussed in detail in the following chapter.

Additionally, there is some evidence that common fate can combine information from the auditory and visual modalities. For example, it is easier to recognize speech in a noisy environment when the speaker's face is visible (Dodd [78]).

Closure

The Gestalt psychologists noticed a tendency to close (complete) certain perceptual forms. For example, the left panel of figure 3.1 is likely to be interpreted as a line that is obscured by a grey rectangle, even though the line is actually incomplete. Note that in order for closure to occur, there must be some evidence that the perceptual form is obscured, rather than interrupted. This is apparent from the right panel of figure 3.1, where the grey rectangle obscuring the line has been removed. Here,

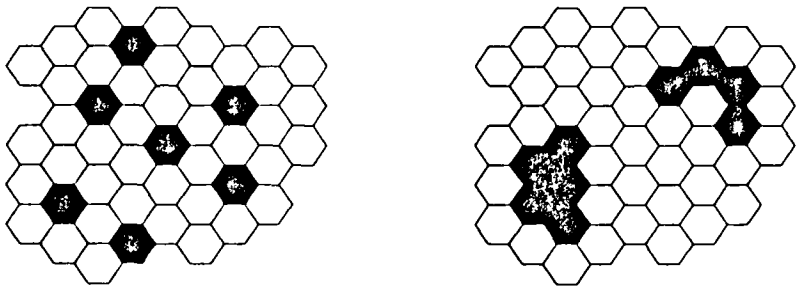

Figure 3.2: Illustration of the Gestalt principles of similarity and proximity. On the left, the black and white hexagons are grouped because they have a similar colour. One the right, two groups of black hexagons are seen, because the members of one group are closer to each other than they are to the members of the other group. Adapted from Bregman [24].

there is little tendency to close the gap, and the figure is perceived as two isolated lines.

The left panel of figure 3.1 can be used to illustrate an equivalent effect in audition. If part of a tone (black line) is deleted and replaced with a brief burst of random noise (grey rectangle), the tone is heard to continue through the noise, even though it is not physically present (Miller and Licklider [174]). This phenomenon is known as the *auditory continuity effect.* If the noise burst is absent, so that the stimulus resembles the right panel in figure 3.1, continuity is abolished and a gap is heard in the tone. A similar continuity effect can be demonstrated when speech is alternated with noise bursts (Dirks and Bower [77]). In this case, the missing speech sounds are perceptually restored. Clearly, this is a useful perceptual mechanism, since the speech of a talker is often interrupted by other sounds in the acoustic environment. The auditory continuity effect is discussed further in section 4.4.1.

Similarity and Proximity

Another Gestalt principle of perceptual organization states that elements will be grouped if they are similar. This effect is illustrated in the left panel of figure 3.2, where the black and white hexagons form different subgroups because of the similarity of their colour. In audition, sounds with a similar pitch, intensity, timbre or spatial location will tend to form a perceptual group. For example, van Noorden [193] presented listeners with a sequence of tones which had the same frequency, but alternated between two intensities. When the difference between the two intensities was large, the sequence of loud tones and the sequence of quiet tones formed different perceptual streams. However, when the intensity of the tones was similar, a single stream was perceived. Hence, a similarity in intensity promoted perceptual grouping.

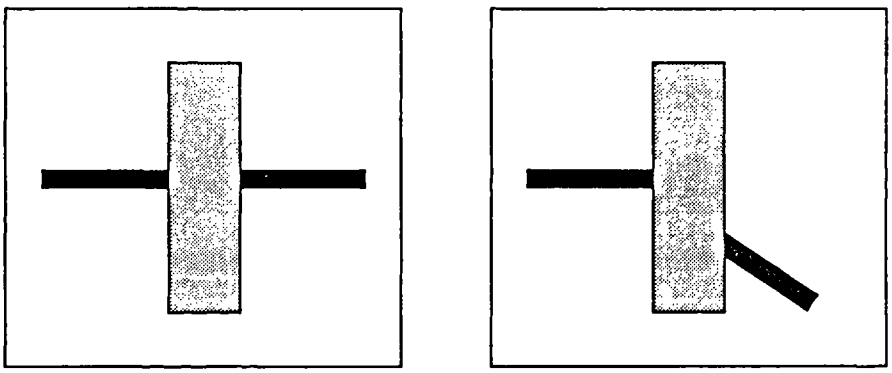

Figure 3.3: Illustration of the Gestalt principle of good continuation. On the left, the black bars have good continuity and are perceived as a single, partly obscured, form. On the right, the black bars have poor continuity and there is no tendency to group them perceptually. Adapted from Bregman[24].

A related factor is the Gestalt principle of proximity. Essentially, this states that the closer the elements of a set are to one another, the greater is the tendency to group them perceptually. A visual illustration of this principle is shown in the right panel of figure 3.2. Here, the black hexagons form two perceptual groups, since the members of one group are closer to one another than they are to the members of the other group.

In audition, acoustic components can be grouped according to their proximity in time or their proximity in frequency. For example, Bregman and Campbell [27] presented listeners with a sequence of alternating high-frequency and low-frequency tones. When the tones were presented slowly, subjects heard the tones in their correct sequence. However, at a faster rate of presentation, the high-frequency and low-frequency tones tended to segregate into different perceptual streams. Hence, the close proximity of the tones in time promoted their perceptual fusion. Similarly, Bregman and Pinker [30] have shown that tones in a repeating sequence are more likely to form a group if they are close in frequency (see section 4.3.1).

Good Continuation

The Gestalt psychologists noted that the smoothness of a change promoted the perceptual integration of changing elements. A visual example is shown in figure 3.3. On the left, the black bars either side of the grey rectangle tend to be perceived as a single partly obscured form, because of the "good continuation" of their lines. On the right, the black bars do not have good continuity and there is no tendency to group them perceptually.

The principle of good continuation can also be applied to audition, since sounds tend to change smoothly in frequency, intensity, location and pitch. Hence, a

smooth change in these properties indicates a continuation of the same sound source, whereas an abrupt change indicates that a new source has become active. Darwin and Bethell-Fox [61] have demonstrated this effect, using repeated synthetic formant patterns that changed smoothly between two vowels. When the pitch of the patterns was constant, a single sound source was heard that contained semivowels and liquid consonants. However, when a discontinuous, stepped pitch contour was imposed on the patterns, they segregated into two perceptual streams that predominantly contained stop consonants. Apparently, the segregation produced illusory silences in each stream during the portions of the pattern attributed to the other stream, and these silences were interpreted, together with the gliding formants, as stop consonants. Darwin and Bethell-Fox confirmed that this effect was due to discontinuities in the pitch contour, rather than concomitant changes in energy.

Some other experimental investigations of good continuation are reviewed in the following chapter. Bregman and Dannenbring [28] have shown that the tendency of a sequence of high and low frequency tones to segregate into two streams can be reduced by connecting successive tones with frequency transitions (see page 79). Also, Ciocca and Bregman [49] have investigated continuity using the auditory equivalent of the visual patterns shown in figure 3.3 (see page 80).

The Figure-Ground Effect

When viewing a visual scene, it is possible to attend to a particular element so that it stands out perceptually from the remainder of the scene. The Gestalt psychologists called this the "figure-ground effect". An analogous effect occurs in audition. For example, in a crowded cocktail party, it is possible to attend to a particular conversation, so that the other voices form a kind of background. Similarly, when listening to a piece of polyphonic music, we attend principally to one melodic line at a time. Hence, it appears that the acoustic environment is segregated into a number of streams, and that a single stream is selected for conscious analysis at a particular instant.

The tendency of listeners to attend separately to different streams has been demonstrated by Bregman and Campbell [27]. They presented subjects with a repeating sequence containing three high-frequency tones and three low-frequency tones, in which the high and low tones alternated. When asked to judge the order of the tones, the majority of listeners reported hearing all of the high tones followed by the low tones, or all of the low tones followed by the high tones. Apparently, listeners were attending separately to the order in each frequency-based stream, and then concatenating the two remembered sequences.

There is good evidence that stream segregation arises principally as a result of preattentive grouping mechanisms. For example, Bregman and Rudnicky [31] have demonstrated that tones in an unattended stream can capture tones from an attended stream. Hence, a second stream can exist even though it is not the subject of conscious analysis. Nonetheless, it is also clear that attention may influence the

formation of perceptual streams. This point has been illustrated by van Noorden [193], using a sequence of alternating high-frequency and low-frequency tones. When the interval between the tones was about seven semitones, frequency proximity cues were ambiguous and listeners could hear segregation or fusion depending on their attentional focus. Specifically, subjects could attend to the high and low frequency tones in separate streams, or could attend to all of the tones in the same stream.

Note that the figure-ground effect is closely related to the Gestalt principle of good continuation. When an abrupt change occurs in the acoustic environment, the attention of a listener is drawn to that change, so that it becomes the "figure" against the other sounds in the acoustic "background". Clearly, this is a useful perceptual mechanism, since attention is directed to new and potentially important events in the acoustic environment.

3.2 Simultaneous and Sequential Grouping

Bregman and Pinker [30] make a distinction between *simultaneous* and *sequential* grouping processes. Simultaneous processes group the components of a source that occur at the same time, but at different spectral locations. Sequential processes group acoustic components that have arisen from the same source over time. In the time-frequency domain, these two phenomena can be regarded as "vertical" and "horizontal" grouping respectively. For example, the visual grouping of the black bars in the left panel of figure 3.3 is a sequential process.

It is useful to distinguish these two aspects of perceptual organization, since they are influenced by different properties of the acoustic input. Simultaneous grouping is affected by frequency proximity and common fate (common onset, offset, periodicity, amplitude modulation and frequency modulation). In contrast, sequential grouping is influenced by many of the factors that define the similarity and good continuation of successive sounds. These include their pitch, temporal proximity, intensity and spatial location.

Although simultaneous and sequential grouping processes are influenced by different factors, it is clear that they can interact to solve a scene analysis problem. Indeed, many experiments have exploited the competition between simultaneous and sequential organization, such as the paradigm used by Bregman and Pinker [30] (see section 4.3.1). A similar paradigm is shown in figure 3.4 (Bregman and Tougas [32]). Here, listeners were presented with a repeating cycle consisting of a tone A followed by a pair of tones B and C. In some conditions, a tone D was also included, otherwise a silent gap was left which was the same length as D. Subjects were asked to judge how clearly A and B could be heard as a repeating pair. When D was present, the AB grouping was more prominent. Apparently, C and D tended to form a sequential group so that the simultaneous fusion of B and C was weakened. Hence, it was easier for A to capture B into a sequential stream, and the AB pattern was heard more clearly. This result implies that simultaneous and sequential grouping

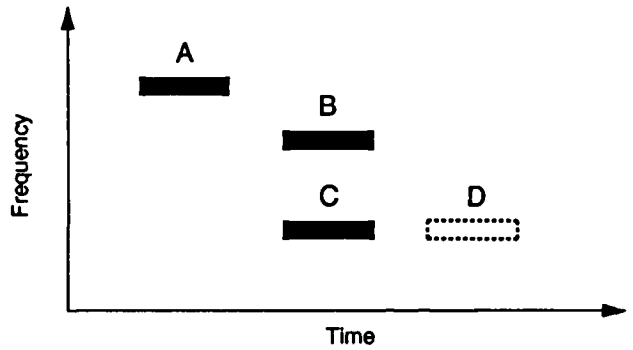

Figure 3.4: Schematic diagram of the stimulus used by Bregman and Tougas. A pure tone A alternates with a pair of tones B and C. In some conditions, a tone D is also included in the repeating cycle. From Bregman and Tougas[32].

processes can compete for the same acoustic event.

Note that in order to group the components of a single source over time, the auditory system has to solve a *temporal correspondence problem.* Specifically, sequential grouping mechanisms must be able to relate the auditory representation of an acoustic event at a particular time with the representation of the same event at a later time. This problem, and a possible solution, are discussed in section 4.4.

3.3 The Principle of Exclusive Allocation

The Gestalt psychologists describe a principle of belongingness, which states that a perceived boundary in the mental representation of a drawing always belongs to some organization in the figure. A familiar visual example of this principle is shown in figure 3.5, which is perceived either as a vase or two faces. When the vase is seen, the lines separating the white and black areas belong to the vase and define its shape. Similarly, when the faces are seen, the same lines belong to the faces.

Note that in figure 3.5, the lines separating the black and white areas never belong to the vase and the faces at the same time. This observation leads to another organizational principle, which Bregman [24] calls the *principle of exclusive allocation.* This principle states that a sensory element should not be used in more than one organization at a time. In fact, this assumption has been implicit in several of the examples already mentioned in this chapter, since competition between grouping processes cannot arise without exclusive allocation. For example, in figure 3.4 sequential and simultaneous streams are competing for B and C. This implies that B and C cannot belong to both streams at the same time. Another example is the Bregman and Rudnicky experiment discussed on page 33.

Although the principle of exclusive allocation always appears to apply when sounds do not overlap in time, it may be violated when sounds are concurrent. For example,

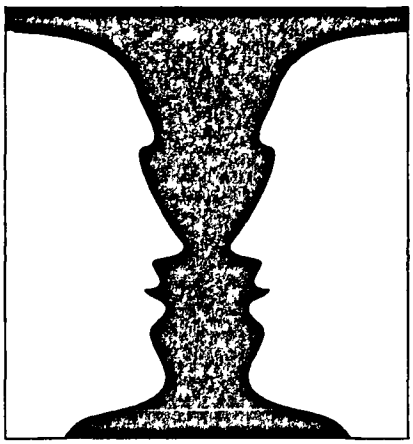

Figure 3.5: The vase-face illusion of the Gestalt psychologists.

a violation of the principle of exclusive allocation occurs in the "duplex" perception of speech (Rand [210]). Here, spectral information is used twice, once to define a speech sound and once to define a non-speech sound. Other violations occur in pitch perception, where a mistuned harmonic of a complex tone can be heard as a part of two sources (Moore *et al.* [185]). These examples, and several others, are discussed in detail in section 5.5.

Intuitively, it seems reasonable that the principle of exclusive allocation should be violated in some circumstances. Since many sources are usually active in the acoustic environment at a particular time, several sounds may contribute to the energy in a certain spectral region (Bregman [23]). Hence, it is desirable to be able to share energy between several perceptual organizations.

3.4 Primitive and Schema-Based Segregation

Bregman [24] makes a distinction between *primitive* and *schema-based* organization. Primitive segregation employs innate grouping rules, such as common fate and good continuation, which use neither past learning nor voluntary attention. In contrast, schema-based segregation employs learned knowledge of familiar sounds in the acoustic environment, such as music and speech.

Schemas appear to extract the spectral components that they require directly from the auditory scene. Hence, they act as a complete scene analysis process, and are not necessarily dependent upon a prior analysis of the sensory input by primitive grouping mechanisms. An example of schema-based grouping occurs in the perceptual restoration of speech that has been interrupted by a brief burst of noise (see page 31). Presumably, a schema for the interrupted word is activated, which selects certain spectral components from the noise burst and interprets them as parts of the missing speech sound.

Another property of schema-based mechanisms is their ability to regroup spectral components that have been previously segregated by primitive processes. This point can be illustrated with a synthetic, two-formant speech sound in which each formant is excited on a different fundamental frequency (Bregman [24]). When presented with this stimulus, listeners hear the two formants individually, but also hear a speech sound corresponding to both formants. Apparently, the two formants have been segregated by primitive processes because they have a different fundamental, but a schema has recombined them to form a speech percept. Other examples of schema-based segregation are discussed in section 5.5.

3.5 Summary and Discussion

This chapter has reviewed some of the principles underlying auditory scene analysis, many of which correspond to the rules of perceptual organization proposed by the Gestalt psychologists.

The basic tenet of Gestalt psychology is that grouping is determined by competition between the "forces of attraction" of perceptual elements. Other theories have stressed rule-based mechanisms of perceptual organization, rather than attractional forces. For example, Jones [127] proposes a theory in which sound is represented within the subjective dimensions of pitch, loudness and time. Rules attempt to predict the position of a new sound in this three-dimensional space from the positions of a sequence of previous sounds. If the prediction is good, the new sound is integrated into the sequence. Theories of this type suffer from a number of difficulties (Bregman [24]). Firstly, rule-based descriptions of a stimulus cannot explain the perception of random sequences. Additionally, the theories fail when there is no rule to guide the organization of a particular stimulus.

Recall from section 2.2.3 that properties such as periodicity, spatial location, intensity and frequency modulation appear to be mapped in the higher auditory system. As might be expected, these properties are also cues which can be used to achieve the perceptual segregation of sound sources. The next chapter presents computational models of several auditory maps, which provide a basis for the scene analysis strategy described in chapter 5.

Chapter 4

Primitives for Auditory Scene Analysis

This chapter describes a computational model of the auditory periphery, and models of periodicity, onset, offset and frequency transition maps. The psychophysical and physiological motivation for each map is discussed in detail, followed by a description of the model and some sample output.

Maps for extracting periodicity information from auditory nerve firing patterns are described in section 4.2. Onset and offset maps are described in section 4.3, followed by a map of frequency transition in section 4.4. Finally, some other grouping primitives that are not employed in the model are discussed in section 4.5.

4.1 Auditory Periphery

Over the past decade, a large number of models of the auditory periphery have been described in the hearing science and speech technology literature. Many of these models have been motivated by the convergence of psychophysical and physiological estimates of auditory frequency selectivity, and by the belief that a model of peripheral auditory processing will provide an improved spectral analysis for automatic speech recognition systems.

Since the auditory periphery has been modelled so extensively, no new modelling work is attempted here. Rather, existing models of each stage of peripheral auditory processing are selected "off the shelf", and assembled into a complete system. Generally, models have been chosen that are in closest agreement with the available experimental data, although computational expense has also been considered.

The auditory periphery can be divided into three main functional components, the

outer and middle ear resonances, basilar membrane filtering and inner hair cell transduction (see section 2.1). A model of each component is described below, and the parameter settings are summarized in table 4.1.

4.1.1 Outer and Middle Ear Resonances

The outer and middle ears are approximately linear for small to moderate sound intensities, and hence their resonances can be modelled by a linear filter. Although it is possible to model the transfer function of the outer and middle ears very closely (Deng and Geisler [75], Meddis and Hewitt [167]), a simple high-pass filter of the form

$$y[t] = x[t] - 0.95 * x[t-1] \tag{4.1}$$

was considered to be an acceptable approximation for the functional approach adopted here. In equation 4.1, $x[t]$ is the amplitude of the input at time t.

4.1.2 Basilar Membrane Filtering

The frequency selective properties of the basilar membrane are generally modelled as a *transmission-line*, or as a *filterbank*. Additionally, a number of models manipulate spectra obtained by a discrete Fourier transform (Goldhor [97], Scheffers [234]). Transmission-line models approximate the basilar membrane by a series of sections that vary in their mechanical properties, and they are usually implemented as a cascade of filters (Lyon [156], Dolmazon [79]). However, models of this type tend to be computationally expensive (Deng and Gesiler [75]), so they are not pursued here.

A more efficient method of modelling the basilar membrane is to employ a filterbank, in which each filter simulates the frequency response of a particular point along the cochlear partition (Ghitza [93], Seneff [243], Cooke [52]). Like any filter, the frequency response of an auditory filter can be completely characterized in the time domain by its response to a brief click, called the *impulse response*. Physiological measurements of auditory nerve fibre impulse responses have been made by de Boer and Kuyper [16] and Carney and Yin [44], using a "reverse correlation" paradigm. The filterbank used here is based on an analytical approximation of their experimental data, the *gammatone* function proposed by de Boer and de Jongh [15]. The gammatone filter of order n and centre frequency f_0 Hz is given by

$$gt[t] = t^{n-1} \exp[-2\pi t] \cos[2\pi f_0 t + \phi] u[t] \tag{4.2}$$

where $u[t]$ is the unit step function

$$u[t] = \begin{cases} 1 & \text{if t} \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

and b is related to the bandwidth. The name "gammatone" is originally due to Aertsen and Johannesma [3], and is derived from the observation that the term before the cosine in equation 4.2 is the statistical gamma distribution, and the cosine term is a pure tone of frequency f_0 Hz and phase ϕ.

Patterson *et al.* [200] have compared the gammatone filter with the psychophysically-derived "rounded-exponential" models of human auditory filter shape. They found that the gammatone filter of order 4 provides a very close fit to the simplest member of the rounded-exponential family, the roex[p] filter, over a 60 dB range. Hence, the gammatone function provides the basis for a model of basilar membrane filtering that is in good agreement with first-order physiological and psychophysical estimates of auditory frequency selectivity.

For the computational model described here, it is advantageous to compensate for the phase delays introduced by the filterbank. Specifically, phase is critical in the comparison of onset and offset times in different frequency channels (see section 4.3), and the performance of the frequency transition map is improved if the filterbank is phase-compensated (see section 4.4). Holdsworth *et al.* [116] describe two methods of phase compensation for the gammatone filter, both of which are used here. Firstly, the peaks of the envelopes of each impulse response can be aligned by introducing a time lead

$$t_c = \frac{n-1}{2\pi b} \tag{4.4}$$

to the output of the filter. Secondly, a peak in the temporal fine structure can be aligned with the peak in the envelope by a phase correction

$$\phi_c = -2\pi f_0 t_c \tag{4.5}$$

Substituting into equation 4.2, this gives the phase-compensated filter

$$gt_c = (t + t_c)^{n-1} \exp[-2\pi b(t + t_c)] \cos[2\pi f_0 t] u_c[t] \tag{4.6}$$

where

$$u_c[t] = \begin{cases} 1 & \text{if t} \geq -t_c \\ 0 & \text{otherwise} \end{cases} \tag{4.7}$$

in which the peak impulse response at $t = 0$ is aligned for each characteristic frequency. Here, a digital approximation of the gammatone filter suggested by Cooke [52] is employed, where an "impulse-invariant" transform is used to convert from the continuous domain to the digital domain. Cooke compares three methods of digital approximation, and concludes that an impulse-invariant transform gives the closest fit to the ideal magnitude, phase and impulse responses of the filter.

Parameter	Description	Value	Units
n	gammatone filter order	4	
	number of filters	128	
	lowest filter centre frequency	50	Hz
	highest filter centre frequency	5000	Hz
	filter spacing	0.2	ERB

Table 4.1: Parameter settings for the peripheral auditory model.

Auditory filters are distributed across frequency according to their bandwidths, which increase quasi-logarithmically with the centre frequency of the filter (see figure 4.27). Here, the gammatone filters are spaced on the equivalent rectangular bandwidth (ERB) scale of Moore and Glasberg [183]. Specifically, 128 overlapping filters were spaced equally in ERB-rate in the range 50 Hz to 5000 Hz, corresponding to a distance of approximately 0.2 ERB between adjacent centre frequencies. Using more than 128 filters does not significantly improve the performance of the model, and introduces an additional computational burden.

4.1.3 Inner Hair Cell Transduction

Mechanical motion of the basilar membrane is converted into spikes in the auditory nerve by inner hair cells. This transduction process gives rise to phase-locking, compression, saturation and adaptation effects (see section 2.1.3).

Movement of the basilar membrane displaces the stereocilia of inner hair cells, causing changes in the cell's receptor potentials. The relationship between stereocilia displacement and receptor potential is nonlinear, and is thought to underlie the phase-locking, compression and saturation behaviour observed in the auditory nerve. Generally, the nonlinearity is modelled as a sigmoidal function (Seneff [243], Shamma [246], Meddis [164], Cooke [52]), although simple half-wave rectifiers have also been used (Lyon [156]). Adaptation phenomena are thought to be caused by depletion of neurotransmitter at the inner hair cell-auditory nerve synapse. Simple models of adaptation employ a single reservoir of neurotransmitter (Schroeder and Hall [240], Oono and Sujaku [195]), but these are unable to reproduce several important aspects of the physiological data. More recent models (Smith and Brachman [256], Schwid and Geisler [241]) employ multiple reservoirs, and give better agreement with experimental findings at the cost (in general) of increased computational expense.

The model of inner hair cell transduction employed here is the multiple-reservoir scheme proposed by Meddis [164, 165, 166]. In a recent review of eight inner hair cell models, Hewitt and Meddis [113] concluded that the Meddis model was in closest agreement with the physiological findings, and had the additional advantage of being computationally efficient. Given the simulated basilar membrane motion from the gammatone filter, the Meddis model returns the probability of a spike occurring in

the auditory nerve. Here, the model is configured according to the parameters given in [165], which simulate an auditory nerve fibre with a high spontaneous firing rate.

4.1.4 Auditory Periphery Representations

Figure 4.1 shows a representation of average firing rate in the auditory nerve for the ten noise sources used in chapter 6. Here, the spike probabilities from the Meddis hair cell model have been integrated over a 20 ms Hamming window, and displayed at 10 ms intervals. Regions of spectral dominance in speech (harmonics and formants) are clearly represented as dark bands of intense firing activity. Note that environmental sounds such as the speech, music and laboratory noise elicit a more complex pattern of response than synthetic sources (noise bursts, 1 kHz tone, telephone and siren).

4.1.5 Summary and Discussion

In this section, a model of the auditory periphery has been described which incorporates outer/middle ear, basilar membrane filtering and inner hair cell transduction effects. The model provides a probabilistic representation of firing rates in the auditory nerve, which forms the input to the auditory maps described in the remainder of this chapter.

The gammatone filters used here provide a good approximation to the frequency analysis performed by the basilar membrane, but they fail to replicate two experimental findings. Firstly, the fourth-order gammatone filter has a symmetric magnitude response, whereas auditory filters are known to be asymmetric, with a longer tail at low frequencies (Patterson *et al.* [200]). However, there is little asymmetry in the *passband* of auditory filters (Lutfi and Patterson [155]), so this may not be a serious deficiency. Secondly, the gammatone is a linear filter which has the same bandpass characteristic regardless of stimulus intensity. In fact, auditory nerve fibres appear to respond in a nonlinear manner, with broader tuning curves at high intensities (Rose *et al.* [227]). Recently, several models of basilar membrane filtering have been proposed which incorporate nonlinear effects (e.g. Jenison *et al.* [125]), and these may provide a more accurate simulation of auditory frequency analysis.

For good performance of the onset, offset and frequency transition maps described later in this chapter, it was necessary to phase-compensate the gammatone filterbank. There is some evidence that the auditory system compensates for its own phase characteristic. For example, the psychophysical experiments of Patterson [199] suggest that the phase lag of the cochlea does not affect the perception of timbre. Patterson concludes that

> "It appears that the auditory system accommodates for the propagation delay in the cochlea and that it can, therefore, be omitted from perceptual models of hearing." (page 1585)

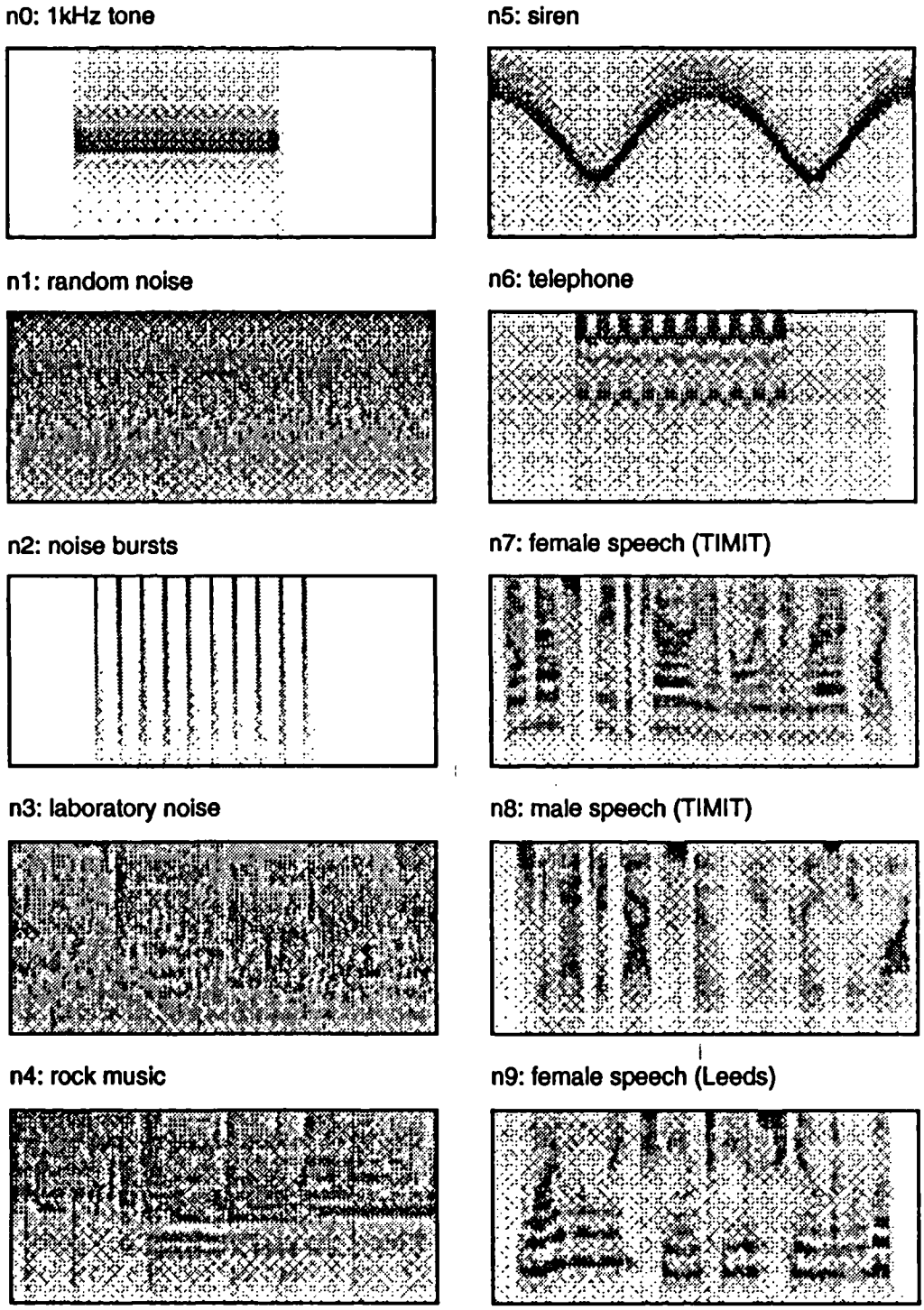

Figure 4.1: Average auditory nerve firing rate representations of the ten noise sources. Time is displayed on the abscissa, and channel centre frequency is displayed on the ordinate.

Note that phase-compensation does not affect the autocorrelation map described in section 4.2.3, since across-channel phase differences in the map are intrinsically corrected by the autocorrelation analysis.

4.2 Periodicity

This section presents models of two maps that extract periodicity information from auditory nerve firing patterns. First, the psychophysical and physiological motivation for the maps is discussed.

4.2.1 Psychophysical Motivation

In this section, evidence is reviewed that perceptual grouping mechanisms integrate regions of auditory nerve activity which have a common periodicity. Since auditory nerve fibres tend to fire at integer multiples of the period of a stimulus (see section 2.1.3), and it is unlikely that two sounds will have identical fundamental frequencies at the same time, neural activity with a related periodicity is likely to have arisen from the same acoustic source.

Since periodic sounds have a pitch, which generally corresponds to the fundamental frequency, auditory processes that group components according to common periodicity are intrinsically linked with mechanisms of pitch analysis. Therefore, the review in this section addresses two related points. Firstly, evidence is presented that the auditory system uses periodicity information in order to calculate pitch. Secondly, evidence is reviewed that differences in pitch assist the perceptual segregation of concurrent periodic sounds.

Evidence for the Role of Periodicity in Pitch Perception

Theories of pitch perception can be categorized into two classes, *pattern recognition models* and *temporal models*. Both types of theory have been proposed to account for the observation that the pitch of a complex tone can be perceived when there is no spectral component at its fundamental frequency.

Pattern recognition models, typified by the theory of Goldstein [98], propose a central mechanism which finds the best fitting harmonic series for a set of resolved frequency components. This type of model is supported by the fact that low, resolvable harmonics tend to dominate the pitch percept (Ritsma [221]). However, pattern recognition theories are unable to explain the (weak) pitches evoked by periodically interrupted noise bursts (Miller and Taylor [175]) and stimuli containing only high, unresolved harmonics (Moore and Rosen [187]). Consequently, alternative models have been proposed which emphasize the role of temporal information in the auditory nerve. Temporal theories, such as the "duplex" scheme proposed

by Licklider [152], assume that adjacent harmonics are not completely resolved, so that the pitch of a complex tone is represented in the periodicity of auditory nerve firings.

Clearly, the pattern recognition and temporal theories disagree on the relative importance of resolved and unresolved harmonics, and neither model can account for all of the experimental findings. Hence, theories of pitch perception have recently been proposed which combine pattern recognition and temporal mechanisms by integrating periodicity information from resolved and unresolved harmonic regions (Moore [179], Meddis and Hewitt [168, 167]). Models of this type are able to explain the majority of psychophysical pitch phenomena, and are supported by recent experimental observations (Carlyon *et al.* [42], Houtsma and Smurzynski [119]). For example, Carlyon *et al.* have shown that listeners can detect differences between the fundamental frequencies of two groups of components in different spectral regions, when only one of the groups is resolved by the auditory periphery.

Evidence for the Role of Pitch in Perceptual Grouping

The previous section suggests that the auditory system combines regions of neural activity which have a common periodicity in order to calculate pitch. Here, evidence is reviewed that similar mechanisms contribute to the perceptual segregation of sounds which have a different fundamental frequency.

One of the first demonstrations of an effect of common fundamental was provided by Broadbent and Ladefoged [34]. They employed a synthetic speech stimulus in which the first two formants were delivered to separate ears of their subjects. When the formants had the same fundamental frequency, the large majority of listeners heard a single voice. However, when the first two formants were synthesized on different fundamentals, listeners reported hearing more than one voice. This result suggests that frequency components which have a common fundamental tend to be grouped into the same perceptual stream.

Broadbent and Ladefoged's results demonstrate the importance of a common fundamental frequency in determining the number of voices that are heard, but they do not suggest whether grouping by fundamental contributes to phonetic categorization. This question has been addressed by Darwin and his colleagues, using a perceptually ambiguous four-formant syllable for which the first three formants are heard as /ru/, and the first, third and fourth formants are heard as /li/. Darwin [58] found that when all four formants were synthesized on the same fundamental, listeners predominantly heard /ru/. However, when the second formant was synthesized on a different fundamental to the others, there was in increase in the number of subjects who heard /li/. Hence, the formant with a different fundamental frequency tended to be excluded from the perceived phonetic category of the syllable. Gardner *et al.* [90] extended this paradigm to include a range of fundamental frequency differences, and found that large differences in fundamental were necessary to produce a change in phonetic percept. For smaller differences, the second formant was heard

as a separate source but the phonetic category of the syllable was unchanged.

Scheffers [233] has investigated the effect of differences in fundamental frequency on the perception of simultaneous (double) vowel sounds. His results suggest that a difference in pitch between two vowels can assist their perceptual segregation (see section 1.4.3). When listeners were presented with vowels on the same fundamental, Scheffers found that they were able to identify both vowels with a performance that was above chance level (45%). However, introducing a difference in fundamental frequency between the vowels improved identification performance, which increased to 62% correct for a 1 semitone difference. Further increases in difference between the fundamentals did not yield a better performance. The generality of these findings has been confirmed by Blokx and Nooteboom [12] using continuous natural speech, and by Chalikia and Bregman [47] using nonspeech pulse trains.

The results of double vowel studies have been interpreted in terms of a "harmonic sieve", which is similar in concept to Goldstein's model of pitch perception mentioned earlier (Duifhuis *et al.* [80], Scheffers [234]). This theory proposes that the components of a harmonic series can be identified by a "sieve" which has "holes" at integer multiples of its fundamental frequency. Harmonics that are aligned with holes in the sieve "fall through" and contribute to the vowel percept, whereas components which are misaligned are blocked. Hence, when the harmonics of two vowels with different fundamental frequencies are present at the same time, each vowel can be isolated by a sieve that is aligned with its fundamental.

Moore *et al.* [185] have quantified the width of holes in the harmonic sieve using a mistuning paradigm. They presented listeners with a harmonic complex in which one component was mistuned, so that its frequency was not an integer multiple of the fundamental. For mistunings of up to 3%, the mistuned harmonic made a normal contribution to the pitch of the complex. However, components mistuned by more than 3% began to be rejected by the harmonic sieve, and made a smaller contribution to the perceived pitch. Additionally, listeners heard components that were mistuned by 2-3% as a separate sound source. Similar findings have been reported by Darwin and Gardner [63] in the context of speech perception. These results suggest that perceptual mechanisms tolerate a small difference in the frequency (or periodicity) of a harmonic when grouping components which have a common fundamental.

In a recent study, Carlyon *et al.* [42] have shown that listeners are able to compare fundamental frequencies across a wide spectral region. Clearly, comparisons of this kind could provide a means of segregating concurrent periodic sounds.

4.2.2 Physiological Motivation

Physiology of Single Cells

Periodicity appears to be coded and enhanced by neurons at many levels of the auditory system. In the auditory nerve, information about the periodicity of a

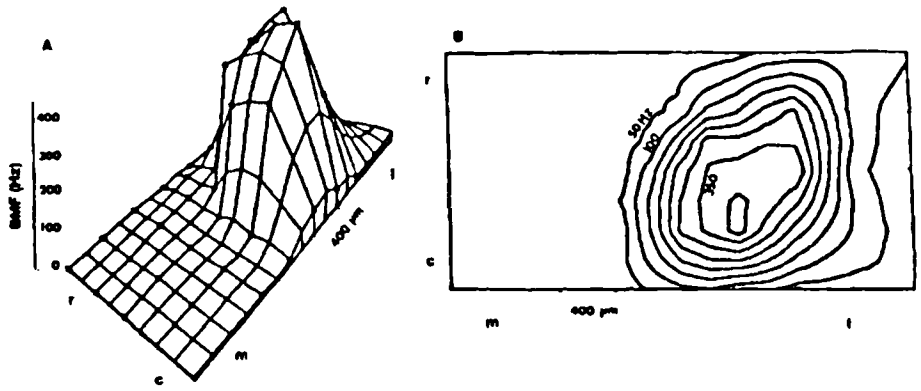

Figure 4.2: Map of periodicity in the inferior colliculus of the cat. Neurons with the same best modulation frequency (BMF) are arranged in concentric rings within iso-frequency sheets, with higher BMFs concentrated in the centre of each sheet. From Schreiner and Langner [237], figure 3.

stimulus is represented by the tendency of fibres to phase-lock to the stimulating waveform (see section 2.1.3). For example, Miller and Sachs [176] have demonstrated that the period of voiced speech is coded in the temporal responses of auditory nerve fibres, and that this representation is stable at high stimulus intensities and in background noise.

There is good evidence that periodicity is actively enhanced at higher levels of the auditory system. Neurons that are tuned to specific rates of periodicity have been identified in the cochlear nucleus (Frisina *et al.* [86]), inferior colliculus (Rees and Møller [214]) and the auditory cortex (Schreiner and Urbas [239]). Additionally, certain types of onset cell in the cochlear nucleus exhibit almost perfect phase-locking to the fundamental frequency of a synthetic vowel, and are said to have "pitch-period following" responses (Kim and Leonard [131], Palmer and Winter [197]). The phase-locking properties of onset cells are discussed further in section 4.3.2.

In the auditory cortex of the bat, Suga *et al.* [264] have identified neurons that respond to combinations of two or more harmonically related tones. Although the bat is highly specialized for echolocation, it is possible that similar neural circuits underlie mechanisms of pitch perception and grouping by common fundamental in the human auditory system.

Topographic Organization

If periodicity is a perceptually important acoustic parameter, then it is likely to be represented in an orderly manner in the higher auditory system (see section 2.2.3). Indeed, there is good evidence that periodicity information is systematically mapped

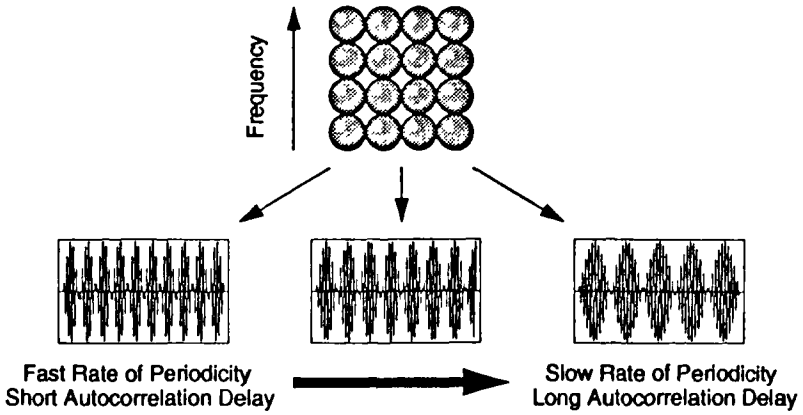

Figure 4.3: Schematic of the autocorrelation map. Each neuron is tuned to a particular rate of periodicity, depending on its autocorrelation delay time.

in at least one auditory nucleus.

In the auditory cortex, Schreiner and Urbas [239] have identified a systematic relationship between preferred rate of periodicity and characteristic frequency. Neurons are tuned to a *best modulation frequency* in the range 3-100 Hz, either in terms of their firing rate or degree of synchronization to a periodic stimulus.

Schreiner and his colleagues have also described a similar organization in the inferior colliculus of the cat and the midbrain of Guinea fowl (Schreiner and Langner [237], Langner *et al.* [145, 144]). In the inferior colliculus, cells with the same best modulation frequency are arranged in concentric rings within each iso-frequency plane (see figure 4.2). Neurons are tuned to rates of periodicity between 10 Hz and 1000 Hz, with higher best modulation frequencies concentrated in the centre of each sheet of cells. Additionally, separate iso-frequency sheets in the inferior colliculus are connected by interneurons (Oliver and Morest [194]), and hence the nucleus has a suitable architecture for comparing periodicities in different frequency regions. This type of organization might provide a basis for pitch analysis or grouping by common periodicity.

4.2.3 A Model Periodicity Map

This section presents a model of a map of periodicity which is based on the "duplex" theory of pitch perception proposed by Licklider [152] (see section 4.2.1). The essence of the "duplex" scheme is that a spectral analysis in the frequency domain is performed simultaneously with a periodicity analysis in the time domain. Licklider suggests that periodicities in the temporal fine structure of auditory nerve firing patterns can be identified by an *autocorrelation* analysis at each characteristic frequency, an operation which is equivalent to building a histogram of the time intervals between each spike and every other spike.

Periodicity

Autocorrelation is a mathematical technique in which a signal is multiplied by a time-delayed version of itself. For a periodic sound, the autocorrelation function shows a peak at the time delay corresponding to the period of repetition. Hence, the map of periodicity employed here is a two-dimensional representation in which neurons at each characteristic frequency are tuned to a series of different autocorrelation time delays (see figure 4.3). Similar computational models based on Licklider's theory have been proposed by Weintraub [277], Gardner [88], Slaney and Lyon [252], Assmann and Summerfield [8] and Meddis and Hewitt [168, 167], and have been named "autocorrelograms" or "correlograms". However, the model described here will be referred to as an *autocorrelation map*, to emphasize the point that Licklider's scheme is compatible with the general framework of auditory map representations described in section 2.2.3.

For an auditory filter with characteristic frequency f, the running autocorrelation $c[\cdot]$ at a time delay Δt is given by

$$c[t, f, \Delta t] = \sum_{i=0}^{\infty} r[t - T, f] r[t - T - \Delta t, f] w[T] \tag{4.8}$$

where

$$T = idt \tag{4.9}$$

and $r[\cdot]$ is the probability of a spike in the auditory nerve, derived from the Meddis hair cell model (see section 4.1). The window $w[T]$ limits the summation over time, and takes the form of a decaying exponential

$$w[T] = \exp\left[\frac{-T}{N}\right] \tag{4.10}$$

with time constant N, as originally suggested by Licklider. Hence, the more distant a spike is in time, the less is contributes to the autocorrelation. When comparing periodicity information in different frequency channels, it is preferable to normalize equation 4.8 so that the autocorrelation function is not influenced by the average firing rate in the auditory nerve. The normalized response of a neuron in the map is given by

$$acm[t, f, \Delta t] = \frac{c[t, f, \Delta t]}{c[t, f, 0]} \tag{4.11}$$

where $c[t, f, 0]$ is equivalent to the energy in the auditory filter channel. Together, equations 4.8 to 4.11 define the computational model. The free parameters of the model are discussed below, and summarized in table 4.2.

Parameter	Description	Value	Units
Δt	autocorrelation time delay	0.0625 to 20.0	ms
N	window time constant	10.0	ms
dt	sample period	0.0625	ms

Table 4.2: Parameter settings for the autocorrelation map.

Range of Autocorrelation Delays

Autocorrelation functions were computed for values of Δt between 0.0625 ms (the sample period dt) and 20 ms (corresponding to a pitch of 50 Hz), in steps of dt. The longest delay of 20 ms was considered to be a sensible upper limit for the period of voiced speech, and the shortest delay was imposed by the sampling rate used in the model.

Window Time Constant

The window time constant N determines the temporal resolution of the autocorrelation map. Autocorrelation functions computed with a long time window have an advantage in extracting the periodicity of static, noisy sounds. Conversely, autocorrelation functions computed with a short time window are able to identify rapid fluctuations in periodicity which are smoothed out by longer windows.

In his original paper, Licklider [152] suggests that an appropriate value for the time constant N is 2-3 ms, although he does not give any justification. More recently, Viemeister [273] has derived a similar value (2.5 ms) using a paradigm in which the detectability of amplitude modulation is used to estimate auditory temporal resolution. However, both of these windows seem too short to give an accurate measurement of the period of sounds with a low pitch, such as male speech. A longer window (10 ms) is suggested by the work of Plack and Moore [206], although this value varies with characteristic frequency and stimulus intensity, and the shape of their window is approximately Gaussian rather than the exponential suggested by Licklider. The longer time window was found to be more satisfactory, and hence N was set to 10 ms in equation 4.10.

4.2.4 Autocorrelation Map Representations

An autocorrelation map for the synthetic vowel /a/, with fundamental frequency 120 Hz, is shown in the upper panel of figure 4.4. Channel centre frequency is represented on the ordinate and autocorrelation delay is represented on the abscissa. The periodicities in each channel are clearly delineated, and common periodicities in different frequency regions form vertical lines in the figure.

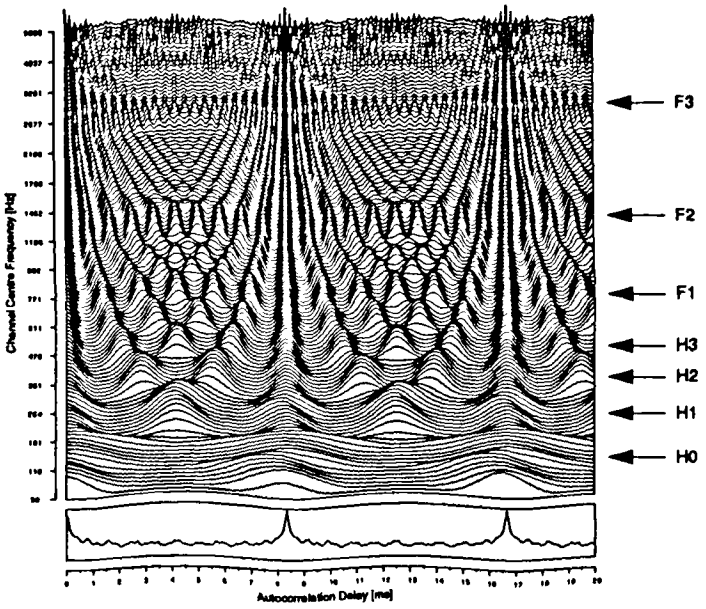

Figure 4.4: Autocorrelation map (top) and summary autocorrelation (bottom) for the synthethic vowel /a/, with fundamental frequency 120 Hz. The positions of the first four harmonics (H0-H3) and three formants (F1-F3) are indicated on the right.

Common periodicities in the individual channels of the map can be emphasized by averaging the autocorrelation functions over frequency. This representation has been called the *summary autocorrelation* by Meddis and Hewitt [168], and is defined by

$$s[t, \Delta t] = \frac{1}{M} \sum_{f=1}^{M} acm[t, f, \Delta t] \tag{4.12}$$

where M is the number of auditory filter channels used in the model (in this case, 128). The summary autocorrelation $s[\cdot]$ for the vowel is shown in the lower panel of figure 4.4, and can be interpreted as showing the relative probability of each pitch period. Hence, a large peak occurs in the summary autocorrelation at a delay corresponding to the pitch period of the vowel (8.3 ms). Peaks also occur at integer multiples of the pitch period, which correspond to subharmonics of the pitch (in this example, only the first subharmonic at 16.6 ms is visible).

Note that this scheme is compatible with the idea that mistuned components of a harmonic complex are rejected by a "harmonic sieve" (see page 46). A component that is slightly mistuned will slightly displace the peak in the summary autocorrelation function. However, a component that is mistuned by a large amount will give rise to a separate peak in the summary autocorrelation, without affecting the position of the pitch period peak.

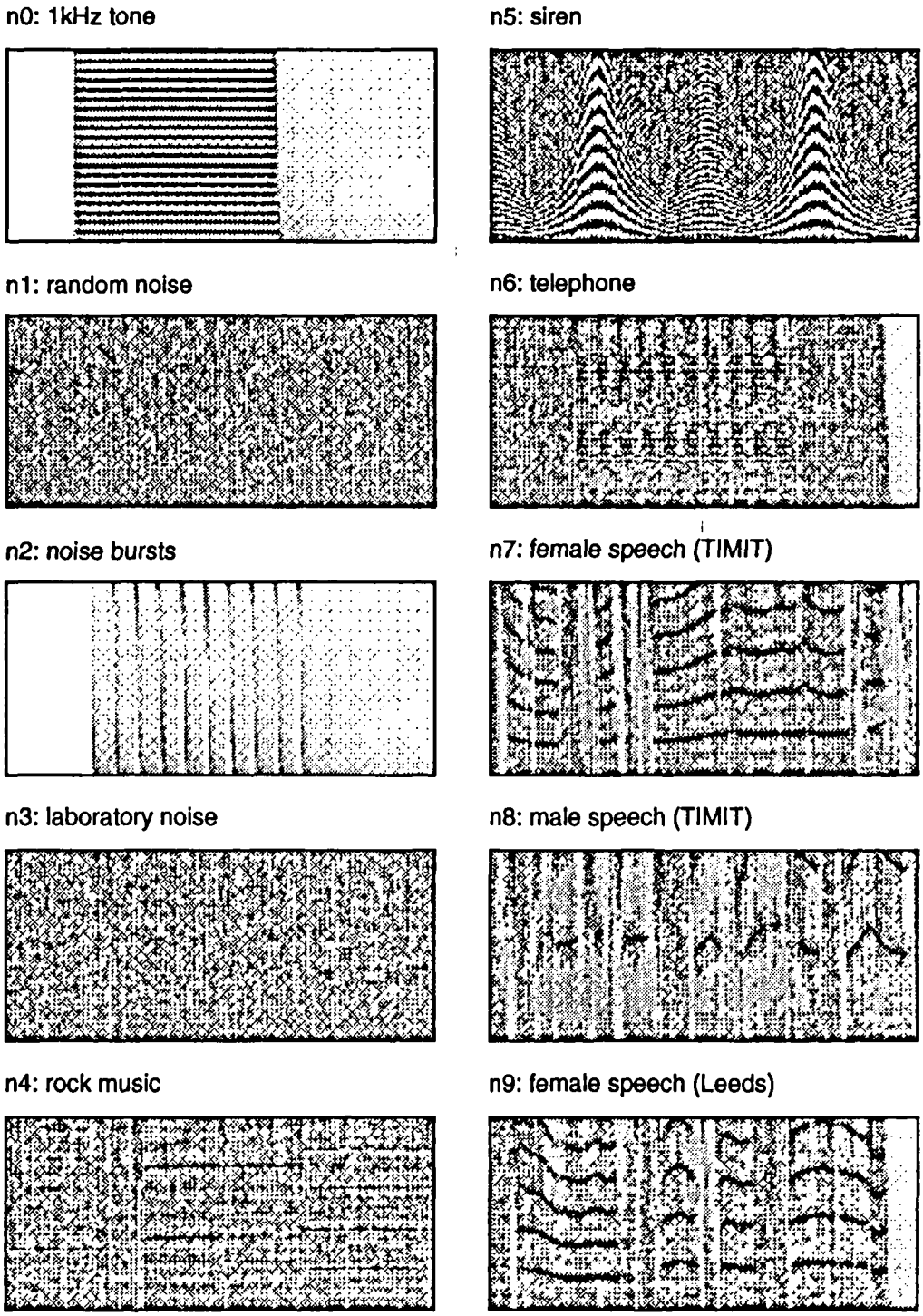

Figure 4.5: Summary autocorrelation representations of the ten noise sources. Time is represented on the abscissa, and autocorrelation delay (corresponding to pitch period) is represented on the ordinate.

It is instructive to display summary autocorrelations as a function of time, in which the probability of a pitch is indicated by the darkness of the image. Summary autocorrelations plotted in this manner are shown in figure 4.5 for the ten noise sources. For periodic sounds such as the speech, siren and 1 kHz tone, variations in the pitch (lowest contour) and its subharmonics (other contours) are clearly represented. Conversely, nonperiodic sounds such as the random and laboratory noise, which do not evoke a pitch percept, fail to generate any coherent activity. The noise bursts are periodic, but they are spaced too far apart in time to evoke a pitch. Note also that sources with a high average pitch (the tone, siren and female speakers) give rise to a greater number of subharmonics than sources with a low average pitch (the male speaker).

4.2.5 Summary and Discussion

This section has presented a model of a map of periodicity, which is based on Licklider's duplex theory of pitch perception. The map codes information about periodicity at each characteristic frequency, and will form the basis for grouping components with a common pitch in the algorithm presented in chapter 5.

Although autocorrelation-based models of pitch perception are able to account for a wide range of psychophysical pitch phenomena (see Meddis and Hewitt [168, 167] for a review), their physiological plausibility has been questioned. For example, Summerfield *et al.* [266] have suggested that autocorrelation is too computationally intensive to form the basis for a mechanism of pitch analysis. This criticism assumes that the auditory nervous system performs calculations in a *serial* manner, like a conventional digital computer. However, biological neural networks operate in *parallel*, and are able to perform many computations simultaneously. Since autocorrelation is highly amenable to parallel computation, it is unlikely to be a prohibitively time-consuming process. Indeed, autocorrelation maps implemented on parallel computer architectures can achieve real-time performance (Slaney [254]). Additionally, other representations have been proposed which are qualitatively similar to autocorrelation maps, but are much less computationally intensive (Patterson [201]).

A more serious problem facing autocorrelation models is that of finding a physiological mechanism which can generate the required time delays. In his original proposal, Licklider [152] suggests that the autocorrelation is performed by a chain of neurons with different synaptic delays, but a physiological counterpart of this arrangement has yet to be identified.

Nonetheless, it is well known that the auditory system is able to create and use time delays. Neurons that are sensitive to differences in the time of presentation of a stimulus at the two ears have been identified in the medial superior olive (Goldberg and Brown [96]), superior colliculus (Hirsch *et al.* [115]) and inferior colliculus (Rose *et al.* [226]). It is likely that these cells encode the spatial location of low frequency sounds by performing a cross-correlation of the temporal response from

each ear (Yin and Chan [284]). Additionally, systematic maps of time delay have been identified in the midbrain of the owl (Knudsen and Konishi [139]) and cat (Bojanowski and Schwarz [22]), as discussed in section 2.2.3. However, the time delays represented in these maps are considerably shorter than the 20 ms or so required for a periodicity analysis. Neurons with longer delays, of up to 24 ms, have been identified in the medial geniculate body of the bat (Suga [262]). The mechanism of these cells appears to involve contributions from axonal delays and self-inhibitory oscillations. These findings suggest that long delay lines are plausible in the mammalian auditory system.

Alternatively, time delays could originate from phase differences along the basilar membrane, as suggested in the "stereausis" model of binaural processing proposed by Shamma *et al.* [248]. Slaney and Lyon [253] describe a modification of the autocorrelation map based on this idea.

Although Licklider's scheme of delay lines is attractive in its simplicity, it is possible that the auditory system codes periodicity information in a different manner. For example, it was noted in section 4.2.2 that neurons in the inferior colliculus are systematically arranged according to their best modulation frequencies, and that this organization could provide the basis for across-frequency comparisons of periodicity information. Hence, the autocorrelation map described here should be regarded as a functional description of periodicity coding in the auditory system, which does not make strong assumptions about the underlying physiological mechanisms.

4.2.6 A Cross-Correlation Map

It is evident from figure 4.4 that the autocorrelation map contains redundant information. Contiguous sections of the auditory filterbank respond to the same spectral dominance, so that channels with centre frequencies close to a harmonic (H0-H3) or formant (F1-F3) have a similar pattern of periodicity. This redundancy provides an early constraint which can be used to group channels of the autocorrelation map that are responding to the same acoustic component. A similar observation has motivated the DOMIN algorithm of Carlson and Granström [40], Cooke's [52] "place groups", Ghitza's [92] "in-synchrony bands" and the "pseudospectrum" described by Deng and Geisler [72].

Regions of the autocorrelation map that have a similar pattern of periodicity can be identified by cross-correlating the responses of adjacent frequency channels. Formally, the similarity at time t of two channels with centre frequencies f_1 and f_2 is given by

$$sim[f_1, f_2, t] = \frac{2 \sum_{\Delta t} acm[t, f_1, \Delta t] acm[t, f_2, \Delta t]}{\sum_{\Delta t} acm[t, f_1, \Delta t]^2 + \sum_{\Delta t} acm[t, f_2, \Delta t]^2} \tag{4.13}$$

The cross-correlation is energy-normalized, so that a difference in the average amplitude of two channels does not affect their similarity score. Consequently, $sim[\cdot]$

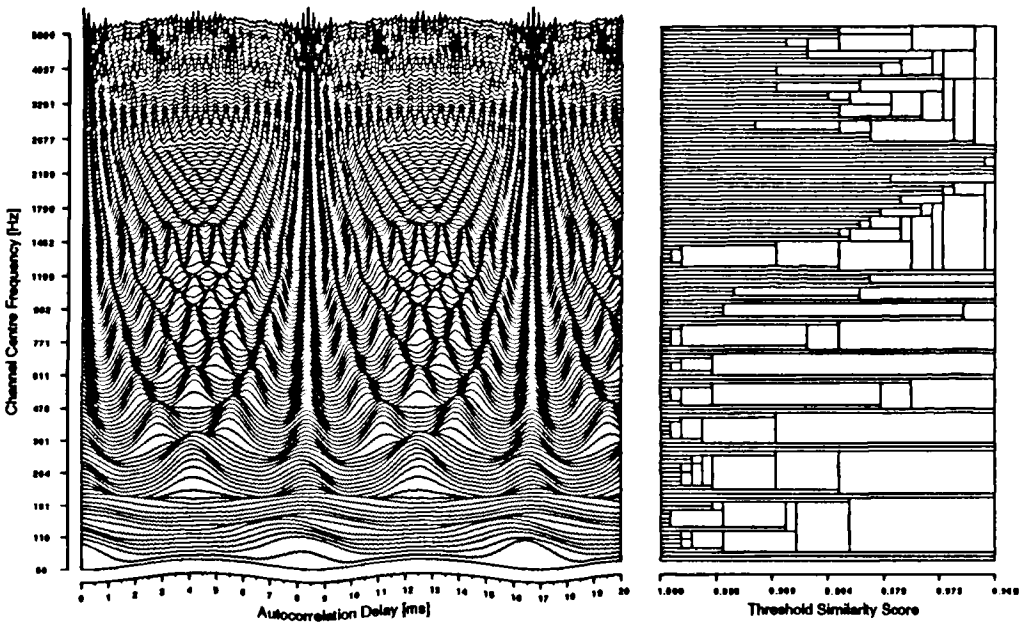

Figure 4.6: Autocorrelation map (left panel) and cross-correlation map (right panel) for the synthetic vowel /a/. Rectangles in the cross-correlation map correspond to groups of channels that extend over frequency and different thresholds of similarity.

has a value between zero (no similarity in periodicity) and unity (identical pattern of periodicity).

Given this metric, it is necessary to decide how good the similarity score of adjacent channels in the autocorrelation map must be in order for them to form a group. The approach employed here is to construct a *cross-correlation map*, which indicates the groups that are formed at a series of different similarity scores. A cross-correlation map for the synthetic vowel /a/ is shown in the right panel of figure 4.6. Like other auditory maps, it is a two-dimensional organization in which characteristic frequency and a tuned parameter (in this case, threshold similarity score) are represented on orthogonal axes. Adjacent channels of the autocorrelation map that have a value of $sim[\cdot]$ equal to or greater than the threshold similarity score are allowed to form a group. At the highest similarity threshold, no groups occur since adjacent channels are not identical. However, as the threshold is relaxed, channels with a similar pattern of periodicity begin to group together. In the figure, groups of channels that extend across frequency and different thresholds of similarity are represented by rectangles, and are referred to as *periodicity groups.*

This technique is reminiscent of the "dendrogram" method of acoustic-phonetic segmentation described by Withgott *et al.* [283] and Glass and Zue [94]. Whereas the dendrogram identifies changes in the spectrum over time, the cross-correlation map identifies changes in periodicity over frequency. However, the principle is the same, since both techniques attempt to find features in a representation that are

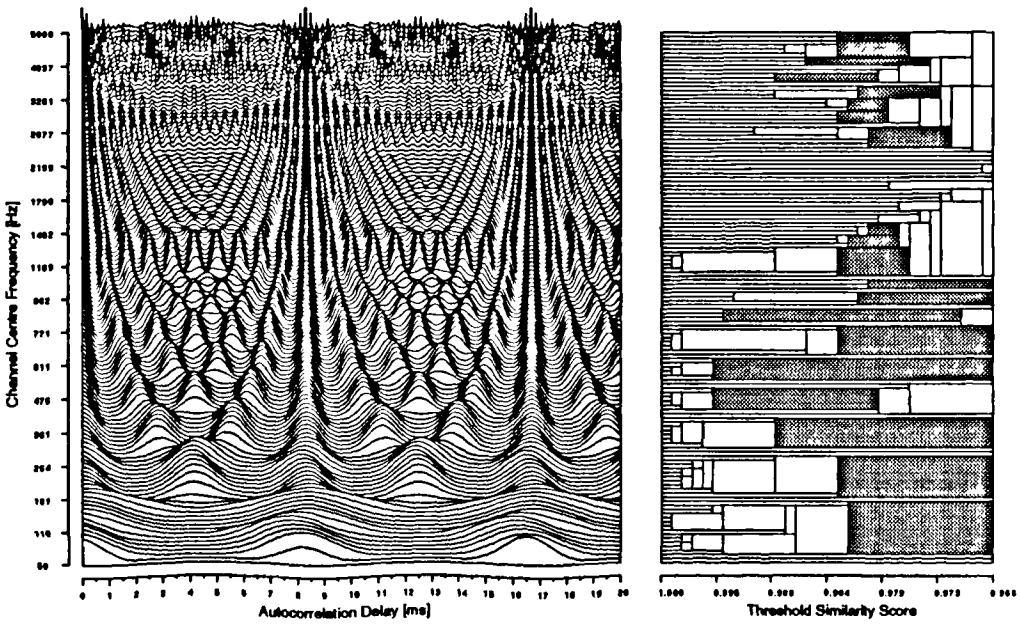

Figure 4.7: Periodicity groups selected by application of a similarity threshold. If adjacent channels have a similarity score equal to or greater than 0.98, then they are selected as a group.

stable across different scales of comparison.

Clearly, the cross-correlation map in figure 4.6 contains many alternative groups at different thresholds of similarity. Three strategies for identifying groups which are most representative of stable areas of periodicity are considered below.

Similarity Threshold

The simplest way of selecting groups from the cross-correlation map is to set a threshold similarity value. For example, in figure 4.7 groups are selected that represent channels with a $sim[\cdot]$ value equal to or greater than 0.98. A problem with this approach is that the choice of threshold is somewhat arbitrary. Additionally, setting a hard threshold may result in the loss of information about good groups, which is a violation of Marr's "principle of least commitment" (see section 1.2.1).

Length Stability Criterion

Generally, groups in the cross-correlation map that represent stable areas of periodicity tend to survive over a number of different similarity thresholds. Hence, good groups can be selected by a criterion which chooses a group if it has no descendents with a greater length along the similarity axis. Figure 4.8 shows periodicity groups selected in this way. In the figure, the descendents of a group lie to its immediate

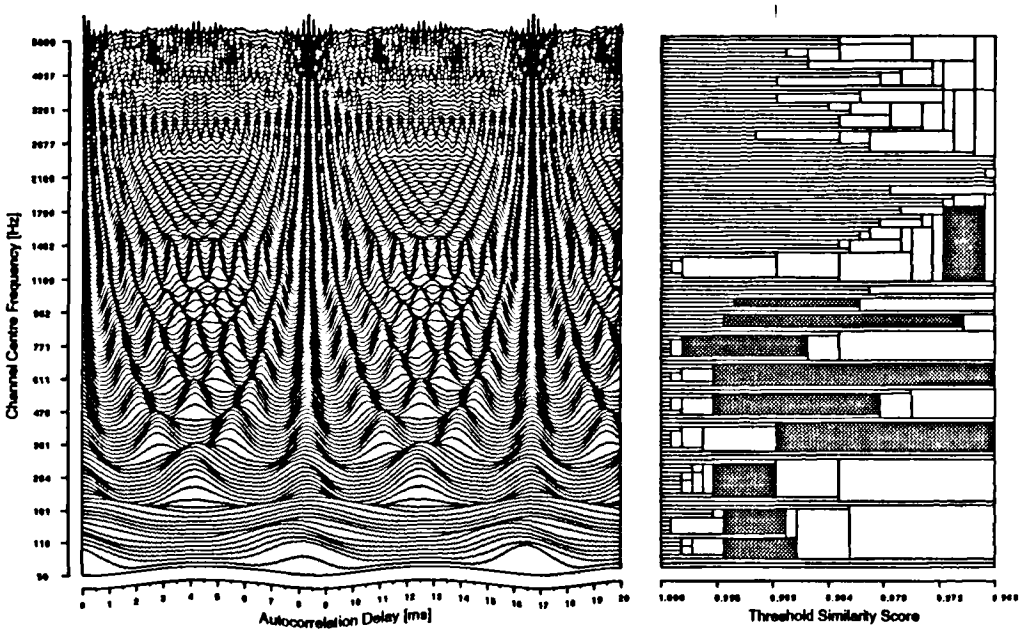

Figure 4.8: Periodicity groups selected by a length stability criterion. Groups that have no descendents of a greater length in frequency-similarity space are chosen.

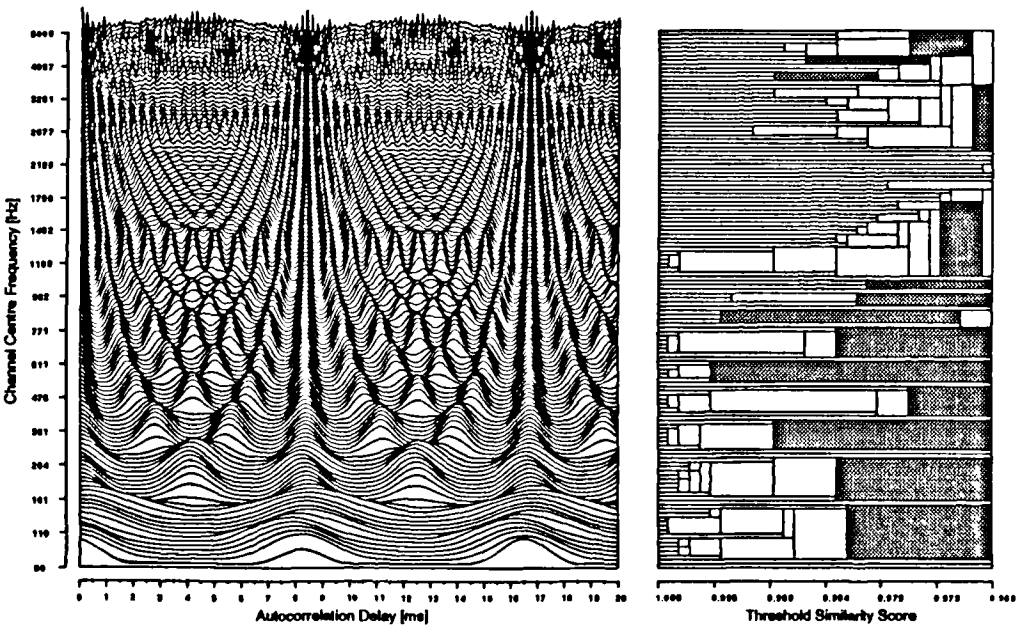

Figure 4.9: Periodicity groups selected by an area stability criterion. Groups that have no descendents of a greater area in frequency-similarity space are chosen.

left (and therefore occur at a higher threshold similarity score).

The length stability criterion performs well at low characteristic frequencies, but good groups in the high frequency channels are missed. This is because there is more variation in the temporal fine structure of adjacent auditory filters at high frequencies, so that single channels survive over many similarity thresholds and are chosen in preference to more representative groups.

Area Stability Criterion

The problems of the length stability criterion can be overcome by identifying groups that are stable across frequency as well as across similarity threshold. Specifically, periodicity groups are selected if they have no descendents with a greater area in frequency-similarity space. An area stability criterion of this type has previously been proposed in a different context by Withgott *et al.* [283].

Groups selected by the area stability criterion are shown in figure 4.9. As required, areas of similar periodicity in the vicinity of harmonics and formants have been identified.

4.2.7 Cross-Correlation Map Representations

Periodicity groups for the ten noise sources, selected from the cross-correlation map by an area stability criterion, are shown in figure 4.10. Harmonics and formants of speech are clearly represented, whereas noise sources give rise to many small, randomly distributed groups. Similarly, the 1 kHz tone and siren produce large groups across a wide frequency range, and periodicities in the telephone and rock music sources are delineated. It is instructive to compare figure 4.10 with the auditory nerve firing rate representations in figure 4.1.

4.2.8 Summary and Discussion

In this section, a cross-correlation map has been proposed which groups auditory filter channels with a similar pattern of periodicity. The map successfully identifies the location of spectral dominances, such as harmonics and formants of voiced speech.

There is good evidence that the auditory system is able to perform a cross-correlation analysis. In the anteroventral cochlear nucleus, Carney [43] has identified neurons which receive convergent inputs from auditory nerve fibres with different characteristic frequencies, and have responses that are consistent with a coincidence or cross-correlation mechanism. Hence, it is possible that a cross-correlation is performed directly on auditory nerve responses, rather than after a periodicity analysis as suggested here.

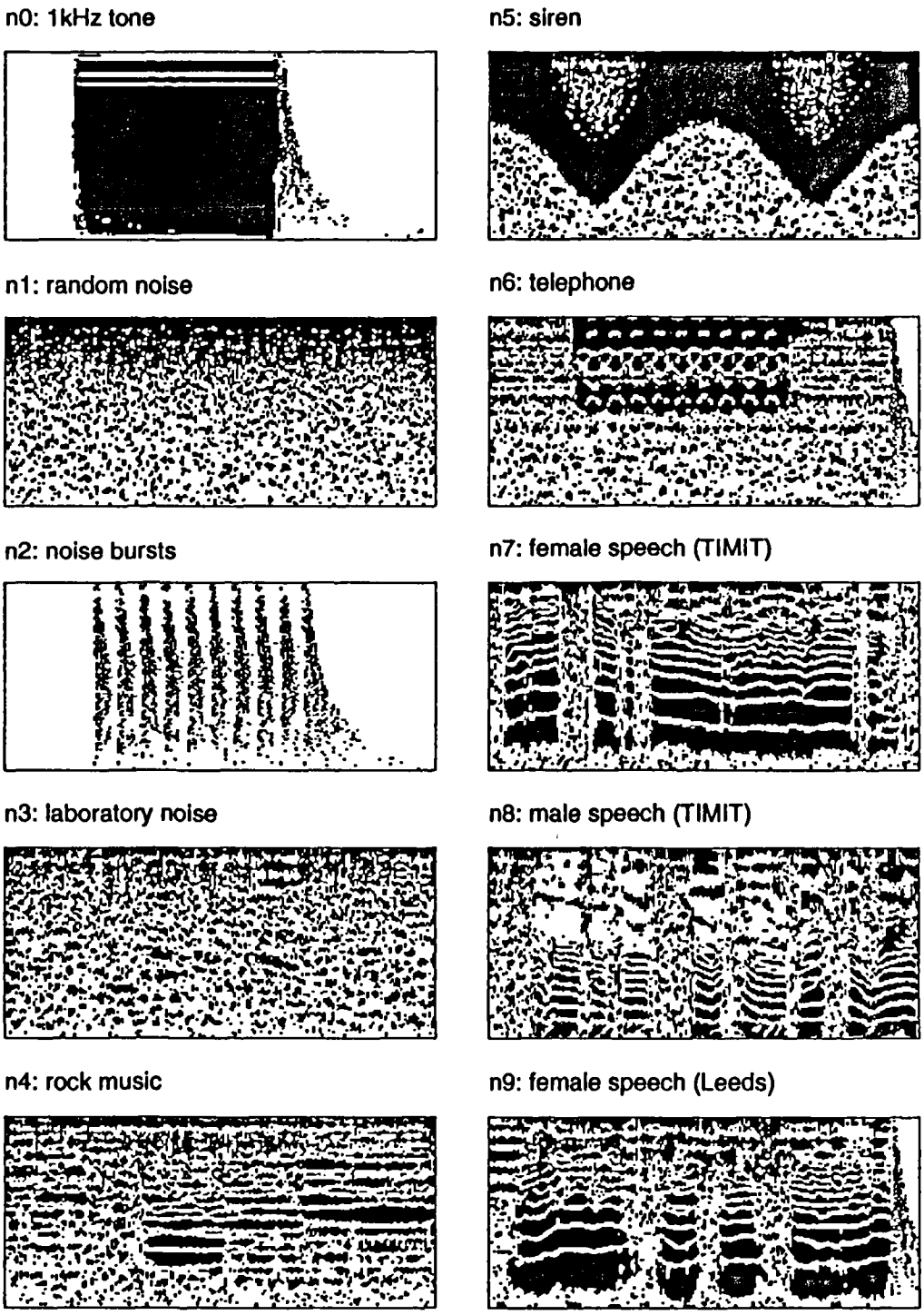

Figure 4.10: Periodicity groups for the ten noise sources, selected from the cross-correlation map by an area stability criterion. Channel centre frequency is represented on the ordinate, and time is represented on the abscissa.

Deng and his colleagues (Deng and Geisler [72], Deng *et al.* [75], Deng and Kheirallah [76]) have proposed an algorithm for cross-correlating auditory nerve responses, which locates spectral dominances in a similar manner to the cross-correlation map described in this section. Their algorithm is less computationally intensive than the one presented here, since it does not require an autocorrelation analysis of each channel. However, the autocorrelation map provides information for mechanisms of grouping and pitch extraction, and has the advantage that phase differences between adjacent channels are intrinsically corrected.

One property of Deng's model is particularly relevant to the work presented here. Specifically, it was found that the ability of his cross-channel correlation algorithm to identify spectral dominances was best when the basilar membrane model incorporated a nonlinear level-dependent damping component. This simulated the tendency of auditory nerve tuning curves to broaden at high intensities (see section 4.1.5), and was necessary to reproduce the *synchrony capture* phenomenon observed in physiological studies (Sinex and Geisler [251], Shamma [244], Deng and Geisler [73], Deng *et al.* [74]). Synchrony capture occurs when a high intensity component produces more response synchrony to itself than is predicted by linear models of auditory nerve tuning curves. For example, the second formant of a nasal consonant-vowel syllable typically captures the synchrony response of fibres that have characteristic frequencies well above the second formant frequency, including those near the third formant (Deng and Geisler [73]). Hence, similar patterns of periodicity occur across wide bands of characteristic frequencies, so that regions of the auditory filterbank that are responding to the same component are very distinct. Therefore, although the cross-correlation map described here is very effective, is it likely that its performance could be improved by incorporating nonlinearities in the auditory filterbank which reproduce synchrony capture effects.

4.3 Onsets and Offsets

This section presents computational models of an onset map and an offset map. The psychophysical and physiological motivation for the maps is discussed first. Subsequently, the models are described.

4.3.1 Psychophysical Motivation

In normal listening situations, it is unlikely that independent sound sources will start and end at the same time. There is good evidence that the auditory system exploits this fact by grouping together acoustic components which have the same onset and offset times. This behaviour is an example of the Gestalt principle of "common fate", which was discussed in section 3.1.

The perceptual effects of onset and offset asynchrony have been investigated by Bregman and Pinker [30], using a paradigm that exploits the competition between

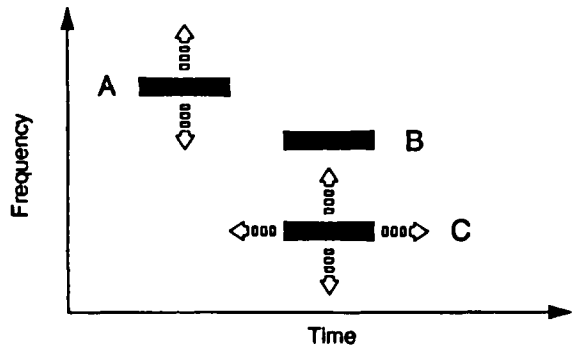

Figure 4.11: Schematic diagram of the stimulus used by Bregman and Pinker. A pure tone A alternates with a pair of tones B and C. From Bregman and Pinker [30].

simultaneous and sequential grouping mechanisms. They presented listeners with a repeating sequence consisting of a pure tone A alternating with a pair of tones B and C, as shown in figure 4.11. When A and B were close in frequency and B was asynchronous with C, A tended to pull B into a sequential stream (A-B-A-B) and C was isolated as a continuous tone (C-C-C-C). However, in conditions where B and C had synchronous onsets and offsets they tended to form a simultaneous group (BC-BC-BC-BC), and A was isolated as a continuous tone (A-A-A-A) regardless of its closeness in frequency to B. Similar results have been obtained by Dannenbring and Bregman [57], using a stimulus in which a pure tone alternates with a complex of three tones. Clearly, common onset and offset times are able to promote the fusion of components that would otherwise belong to separate streams.

Evidence for the importance of onsets and offsets in source segregation has also come from investigations into the perception of musical sound. Rasch [211] played subjects two chords consisting of a high frequency (target) tone which was obscured by a louder (masking) tone at a lower frequency. Listeners were asked to judge whether the frequency of the target tone had moved up or down in the second chord. When the target and masker had synchronous onsets and offsets, so that the only grouping cue was fundamental frequency, thresholds for the task were between 0 and -20 dB. Hence, the target was obscured by a masker that was any louder than 20 dB relative to the level of the tone. However, delaying the onset of the masking tone lowered this threshold by about 10 dB for every 10 ms of onset disparity. When the masker and tone were asynchronous by 30 ms, subjects could still perform the task even though the relative level of the masker to the tone was as high as 60 dB. Rasch observes that onset disparities of 30-50 ms are typical in performed music [212], and that these asynchronies undoubtedly contribute to our ability to hear out the melodic line of each instrument in a polyphonic piece.

Darwin and Ciocca [62] have investigated the role of onset asynchrony in pitch perception. They asked listeners to judge the pitch of a harmonic complex in which

one of the resolved harmonics was mistuned. When the mistuned harmonic started 160 ms before the other components of the complex, it made a reduced contribution to the perceived pitch. Its contribution was abolished if it started more than 300 ms before. These results suggest that mechanisms of pitch perception must take into account the temporal history of the components of a complex in order to exclude those that differ in onset time. Darwin and Ciocca's experiment is discussed further in chapter 5, where a scene analysis algorithm is presented which is compatible with some of their findings.

Onset asynchrony also seems to play a role in speech perception. For example, Darwin [58] describes an experiment in which an onset lead of 300 ms was introduced to the second formant of the synthetic "ru-li" stimulus described in section 4.2.1. Listeners tended to report hearing "li", indicating that the asynchronous formant was making a reduced contribution to the syllable. However, the effect was quite weak.

Rather more convincing evidence for the role of onset and offset asynchrony in speech perception has come from experiments that investigate the conditions under which a tone is integrated into the first formant of a vowel. Darwin [59, 60] and Darwin and Sutherland [65] constructed a continuum of sounds between the vowels /I/ and /e/, which differ only in the position of their first formant peaks. When a tone was added synchronously to the first formant region of each vowel, the vowel percept changed in a manner that reflected the altered position of its formant peak. However, introducing an onset asynchrony of 30 ms to the tone allowed listeners to segregate it from the vowel, so that the original vowel quality was heard. Roberts and Moore [222] have demonstrated that this effect is obtained regardless of whether the tone is added at a harmonic or inharmonic frequency.

So far, it has been assumed that changes in the vowel percept are due to grouping mechanisms which segregate the tone from the vowel. However, the results could also be explained by peripheral adaptation at the frequency of the leading tone. Evidence that adaptation cannot account for all of the effects of onset asynchrony has come from two sources.

Firstly, Darwin and Sutherland [65] demonstrated that the grouping between the leading part of the tone and its continuation into the vowel could be weakened by adding a harmonic of the leading tone which started at the same time, but stopped when the vowel started. Listeners were more likely to hear a change in the vowel colour, indicating that the two leading tones formed a separate perceptual group which ended at the start of the vowel. Darwin and Sutherland quantified the changes in vowel percept, and concluded that grouping mechanisms accounted for at least half of the effects of onset asynchrony.

Secondly, Darwin [59, 60], Roberts and Moore [222] and Darwin and Sutherland [65] have shown that a tone which *stops* 30 ms after a vowel also contributes less to the vowel percept than a tone which is simultaneous with it. Offset asynchrony effects cannot be explained by peripheral adaptation, which is unable to operate

retroactively. It should be noted that in all of these experiments, the effects of offset asynchrony were weaker than those of onset asynchrony. The differences were largest when long (320 ms) vowels were used, presumably because listeners had time to decide the phonetic identity of the vowel before they heard the lagging portion of the tone (Darwin and Sutherland [65]). Studies using shorter vowels of 50-80 ms duration find offset effects that approach the magnitude of onset effects, but are still weaker due to the contribution of adaptation to onset asynchrony (Darwin [59, 60], Roberts and Moore [222]). This point will be discussed again in chapter 5.

Although experiments on speech perception have indicated that onset and offset asynchronies are a powerful cue for perceptual grouping, they raise a paradox. In natural speech, formants move rapidly in frequency so that nearby harmonics are amplified and attenuated at different times. If it is only synchronous harmonics which can be grouped into a formant, then the majority of speech would be unintelligible. Exactly how this problem is avoided is not clear. However, Darwin [60] has shown that there is a limit to the amount of energy which can be incorporated into a harmonic of a vowel without causing it to be segregated as a separate tone. This result suggests the existence of speech-specific constraints which compete with perceptual grouping mechanisms. As Darwin and Sutherland [65] point out,

> "Onset and offset time can be used to separate perceptually harmonics that do contribute to a vowel from those that do not. But such times constitute neither necessary nor sufficient conditions for grouping the harmonics of a single voice. Other principles must be used to ensure that the rapidly modulated harmonics of normal speech are grouped together, and that components occurring simultaneously by accident can be rejected." (page 206)

These considerations have implications for the manner in which onset and offset effects can be incorporated into a model of primitive perceptual grouping, and are discussed again in chapter 5.

4.3.2 Physiological Motivation

Physiology of Single Cells

Cells which respond with a brief burst of activity at the onset or offset of a tonal stimulus are found throughout the higher auditory nuclei, including the cochlear nucleus (Shofner and Young [250]), inferior colliculus (Bock *et al.* [13]) and the auditory cortex (Abeles and Goldstein [1]). A typical onset cell response is shown in figure 4.12 (see also figure 2.1).

In the model presented here, it is proposed that onset and offset cells provide a physiological basis for the perceptual mechanisms which group synchronous acoustic events together. However, although it is reasonable to suppose that neurons which

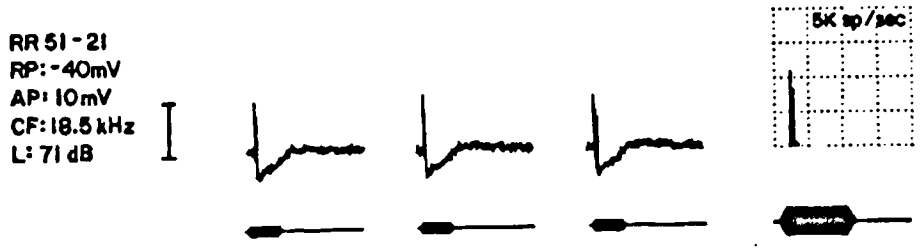

Figure 4.12: Onset cell response to a 25 ms pure tone delivered at its characteristic frequency. The plots on the left show the membrane potential for three presentation of the tone, and the plot on the right shows the number of spikes elicited. From figure 8 of Romand [225].

fire at the start and end of an applied stimulus are coding onsets and offsets in some way, it should be noted that they may have other functions. The precision with which onset cells preserve timing information makes them equally suited to coding periodicity or interaural timing disparities (Rhode and Smith [218]). Additionally, the wide dynamic range of some onset cells suggests that they may encode intensity (Young *et al.* [289]). Clearly, the diversity of functions which have been proposed for onset cells emphasizes the pitfalls of single neuron studies which were discussed in chapter 2.

Onset and offset responses are obtained from a variety of morphological cell types, which have different mechanisms of action (Romand [225]). In this work, it is assumed that onset responses are caused by an *excitatory input* to the cell at the start of the stimulus, followed by an *inhibitory input* which prevents activity throughout the remaining stimulation. Onset cells with this mechanism are common in the cochlear nucleus, and have been classified as type ON-IN by Rhode and Smith [219]. Offset cell responses probably arise through a similar mechanism of action, in which excitation is delayed relative to inhibition.

Topographic Organization

It has already been shown that periodicity information is represented within an orderly framework in the higher auditory system. Although there is no direct physiological evidence, it seems plausible that onsets and offsets are mapped in a similar way.

If onset and offset cells are organized in auditory maps, then the tuned parameters could be *inhibitory delay* for onset maps, and *excitatory delay* for offset maps. The delay before excitation or inhibition determines the rate of amplitude change that the cells are responsive to. For example, an onset cell with a short inhibitory delay will be sensitive to rapid increases in amplitude that occur over short periods of time, but will respond less to a stimulus with a slow rise time. Conversely, an onset

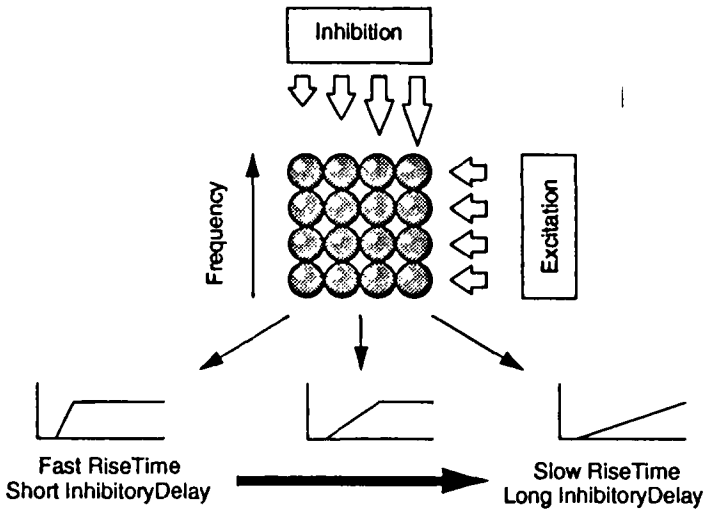

Figure 4.13: Schematic of the Onset map. Cells in the map are tuned according to their sensitivity to the rise time of an onset, which is determined by the delay of inhibition relative to excitation that each cell receives. The length of an open arrow corresponds to the delay of an input.

cell with a long inhibitory delay will detect increases in amplitude that occur over long periods of time, but will locate rapid rises in amplitude with poorer temporal resolution than a short-delay cell. Hence, onset cells could be arranged in a two-dimensional framework with characteristic frequency represented on one axis, and inhibitory delays ranging from short (rapid amplitude rise) to long (slow amplitude rise) represented on the other. Similarly, offset cells could be arranged in a map according to their excitatory delay, varying from short (rapid fall in amplitude) to long (gradual fall in amplitude). Schematics of the onset and offset maps are shown in figures 4.13 and 4.14.

Maps of this form would be valuable, because natural sounds have different rise and fall times. For example, musical instruments have rise times that vary between 20 and 80 ms (Rasch [212]) and fricative speech sounds, such as /s/ and /f/, have slower rise times than stops such as /p/ and /t/ (Stevens [259]). These differences are perceptually important. Rasch [211] has shown that differences in rise time are as effective as onset disparities in promoting the perceptual segregation of simultaneous tones. Additionally, Cutting and Rosner [55] have reported that listeners are able to categorize musical and speech sounds according to rise time.

There is some physiological evidence for the hypothetical onset and offset maps proposed here. Firstly, it is clear that the maps require a series of inputs which are delayed by gradually increasing amounts of time. It might be supposed that these inputs are provided by another map, in which the time delay before the cells respond to stimulation (a property called *latency*) is systematically represented. Such a map has been found in the inferior colliculus of the cat (Schreiner and Langner [237],

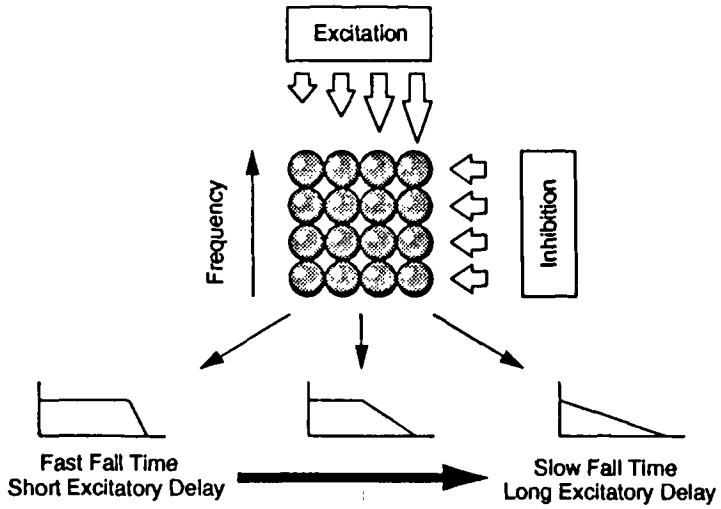

Figure 4.14: Schematic of the Offset map. Cells in the map are tuned according to their sensitivity to the fall time of an offset, which is determined by the delay of excitation relative to inhibition that each cell receives. The length of an open arrow corresponds to the delay of an input.

Langner *et al.* [145]) and rat (Horikawa and Murata [118]). Cells with the same latency are arranged in contours within iso-frequency sheets, as shown in figure 4.15. Secondly, Swarbrick and Whitfield [268] have identified neurons in the auditory cortex which respond selectively to certain slopes of triangularly-modulated noise bursts. This suggests that information about rise and fall times is being coded at high levels of the auditory pathway.

4.3.3 A Model Onset Map

The onset cell model presented here is intentionally abstract. Although it is possible to construct cell models which reproduce the available physiological data very closely (e.g. Hewitt *et al.* [114]), such a level of detail was considered unnecessary for the functional approach adopted in this work.

The response of a typical ON-IN cell to a brief tone burst at its characteristic frequency is illustrated in figure 4.12. Recall that the mechanism of ON-IN cells involves an excitatory input at stimulus onset, followed by an inhibitory input which is delayed by a few ms. Electrical changes caused by these inputs, called excitatory or inhibitory post-synaptic potentials (*EPSPs* or *IPSPs*), are integrated across time by the onset cell membrane. The potential difference across the cell membrane, called the *membrane potential*, is shown on the left of figure 4.12. At the start of the tone, the cell receives EPSPs which raise the membrane potential past firing threshold so that a spike is generated. Following this, the membrane potential falls and is kept low throughout the remainder of the stimulus due to the arrival of IPSPs.

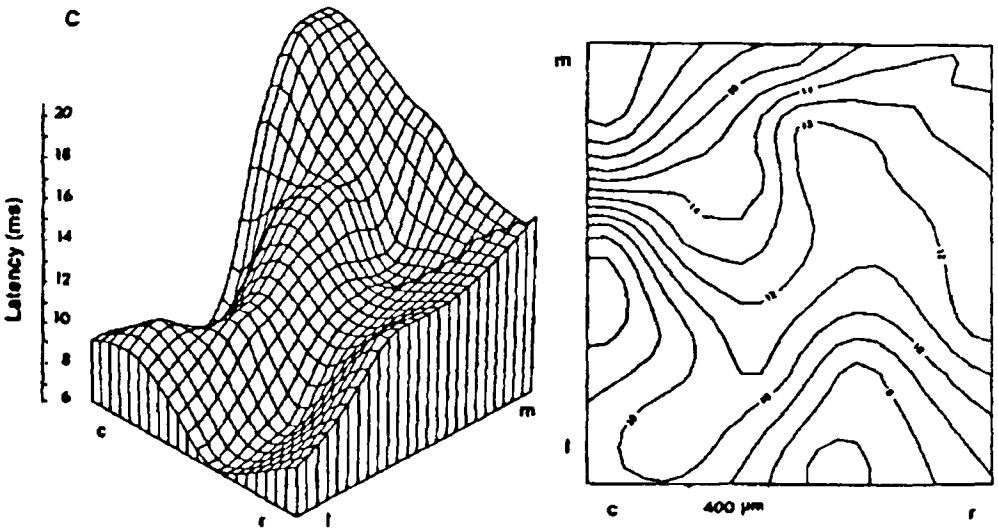

Figure 4.15: Map of latency in the inferior colliculus of the cat, showing the distribution of latencies within a single iso-frequency sheet. Contours on the right are labelled with the latency time in ms. From figure 6 of Schreiner and Langner [237].

At the end of the tone, inhibition ceases and the membrane potential rapidly returns to its resting level.

This mechanism can be approximated by writing the membrane potential $p_{on}[t]$ as a leaky sum of the excitatory and inhibitory inputs to the cell,

$$p_{on}[t] = p_{on}[t-1]c_d + E_{psp}r[t] - I_{psp}r[t-\Delta t_I] \tag{4.14}$$

where

$$c_d = \exp\left[-\frac{dt}{\tau_d}\right] \tag{4.15}$$

The firing rate of the onset cell, $s_{on}[t]$, is determined by the value of the membrane potential when it exceeds a threshold Th,

$$s_{on}[t] = \begin{cases} p_{on}[t] & \text{for } p_{on}[t] > Th \\ 0 & \text{otherwise} \end{cases} \tag{4.16}$$

In equation 4.14, E_{psp} and I_{psp} represent the magnitudes of the excitatory and inhibitory inputs, Δt_I determines the delay before inhibition and τ_d sets the rate at which the membrane potential decays to its resting level. The input $r[t]$ is the envelope of the auditory nerve response at the characteristic frequency of the onset cell, obtained by integrating the output of the Meddis hair cell model over 20 ms with a Hamming window. Using the envelope of the auditory nerve response is a

Parameter	Description	Value	Units
E_{psp}	size of excitatory post-synaptic potential	1.00	mV
I_{psp}	size of inhibitory post-synaptic potential	1.01	mV
τ_d	membrane potential time constant	1.5	ms
Δt_I	delay before inhibition	1-15	ms
Δt_E	delay before excitation	1-15	ms
Th	firing threshold	0	mV

Table 4.3: Parameter settings for the onset and offset cell models.

convenient way of modelling the convergence of many inputs onto an ON-IN cell, which causes a loss of phase-locking at frequencies greater than about 1 kHz (Rhode and Smith [218]).

The value of each parameter in the model is discussed below, and summarized in table 4.3.

Strength of Excitation and Inhibition

When presented with a continuous tone at its characteristic frequency, an ON-IN cell only fires at the start of the stimulus. This suggests that the delayed inhibitory input must be stronger than the excitatory input (Godfrey *et al.* [95]), a view supported by the fact that ON-IN cells in the cochlear nucleus receive strong inhibition from at least two sources (Shofner and Young [250]).

In the model, this effect is approximated by setting the size of the inhibitory input I_{psp} larger than the excitatory input E_{psp}. The values are given in table 4.3. Note that although the parameters are quoted in mV, their values have not been chosen to quantitatively model changes in the membrane potential.

Membrane Potential Time Constant

The parameter τ_d determines the time taken for the membrane potential to decay to 1/e of its maximum deviation from the resting level. Godfrey *et al.* [95] have noted that ON-IN cells are able to fire on every click in a pulse train at rates of up to 400-700 clicks/sec, which suggests that the membrane can reset within a few ms of firing. Accordingly, τ_d was set to a short value (1.5 ms) in the model.

Delay Before Inhibition

As discussed in section 4.3.2, the time delay before inhibition Δt_I is a tuned parameter which varies systematically across one dimension of the onset map. It is

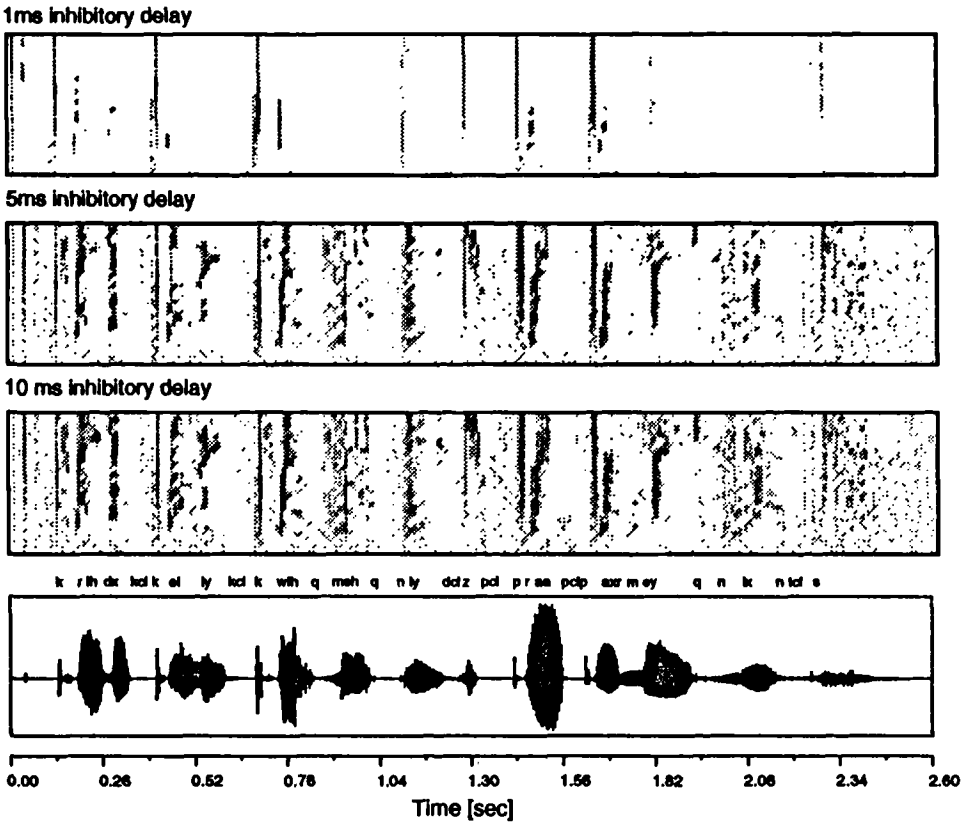

Figure 4.16: Onset map representations of speech, showing the effect of varying the delay before inhibition. The utterance is "critical equipment needs proper maintenance", spoken by a female (TIMIT database). The ordinate of the onset maps is channel centre frequency (50Hz-5kHz).

assumed that the delayed inputs are provided by a map of latency, as described by Langner and his colleagues [145, 237] in the cat and Horikawa and Murata [118] in the rat. These workers found that the latencies represented in the maps varied in the range 5-20 ms. This suggests a maximum *difference* in latency of 15 ms, which provides an upper limit for the delay of inhibition relative to excitation in the onset map. The minimum inhibitory delay is assumed to be 1 ms, which is typical of interneuron latencies. Figure 4.16 shows onset map representations of speech for several different values of Δt_I. The abrupt onsets of the stops /k/ and /p/ are precisely localized in the short-delay (1 ms) map, whereas the gradual onsets of the nasals /m/ and /n/ only appear in the maps with longer inhibitory delays (5 ms and 10 ms).

Firing Threshold

The firing threshold Th is set depending on the inhibitory delay of the ON-IN model. Cells with short inhibitory delays (1-2 ms) only produce positive output at abrupt onsets, so in these cases Th can be set to zero. For cells with longer delays, Th can be increased to remove activity caused by small fluctuations in amplitude over large time intervals (see figure 4.16). In all of the examples shown here, Th was set to zero.

4.3.4 Onset Map Representations

Figure 4.17 shows onset map representations of the ten noise sources used in chapter 6, for onset cells with inhibitory delays of 5 ms. Sources such as the 1 kHz tone, noise bursts, telephone and speech have abrupt onsets which are clearly delineated in the maps.

The siren presents something of a paradox. Although the envelope of the signal has no abrupt increases in amplitude, it elicits a large response from the onset map. This phenomenon arises because the siren changes rapidly in frequency (see figure 4.1). As the concentration of energy moves from one auditory filter to the next, it causes an abrupt increase in activity which is detected by the onset map. Since the map is intended to provide the basis for grouping auditory events by common onset, this situation could potentially cause an onset group to be formed where none should exist. A similar problem has been identified by Mellinger [170] in his model.

Two solutions to this problem are considered here. One is suggested by the fact that ON-IN cells in the cochlear nucleus usually give an inhibitory response to sweeping tones, or fail to respond at all (Rhode and Smith [219]). This is because the cells receive inhibitory inputs from frequencies that are *adjacent* to their characteristic frequency. Hence, tones which sweep up or down in frequency trigger inhibition before they reach the excitatory area of the cell, and fail to elicit an onset response. The model presented here does not have this property, because inhibition is assumed to come only from the same characteristic frequency as the excitatory input. Therefore, a suitable modification of equation 4.14 to prevent the onset cell responding to moving dominances is given by

$$p_{on}[t] = p_{on}[t-1]c_d + E_{psp}r[cf,t] - I_{psp}\sum_{f=-N}^{N} r[cf+f, t-\Delta t_I] \qquad (4.17)$$

where N is the number of frequency channels that contribute inhibitory input to the cell, and $r[cf,t]$ is the envelope of the auditory nerve response at characteristic frequency cf and time t.

A second solution to the problem is to use short (1-2 ms) inhibitory delays in the onset map. Onsets due to sweeping dominances tend to rise gradually (although

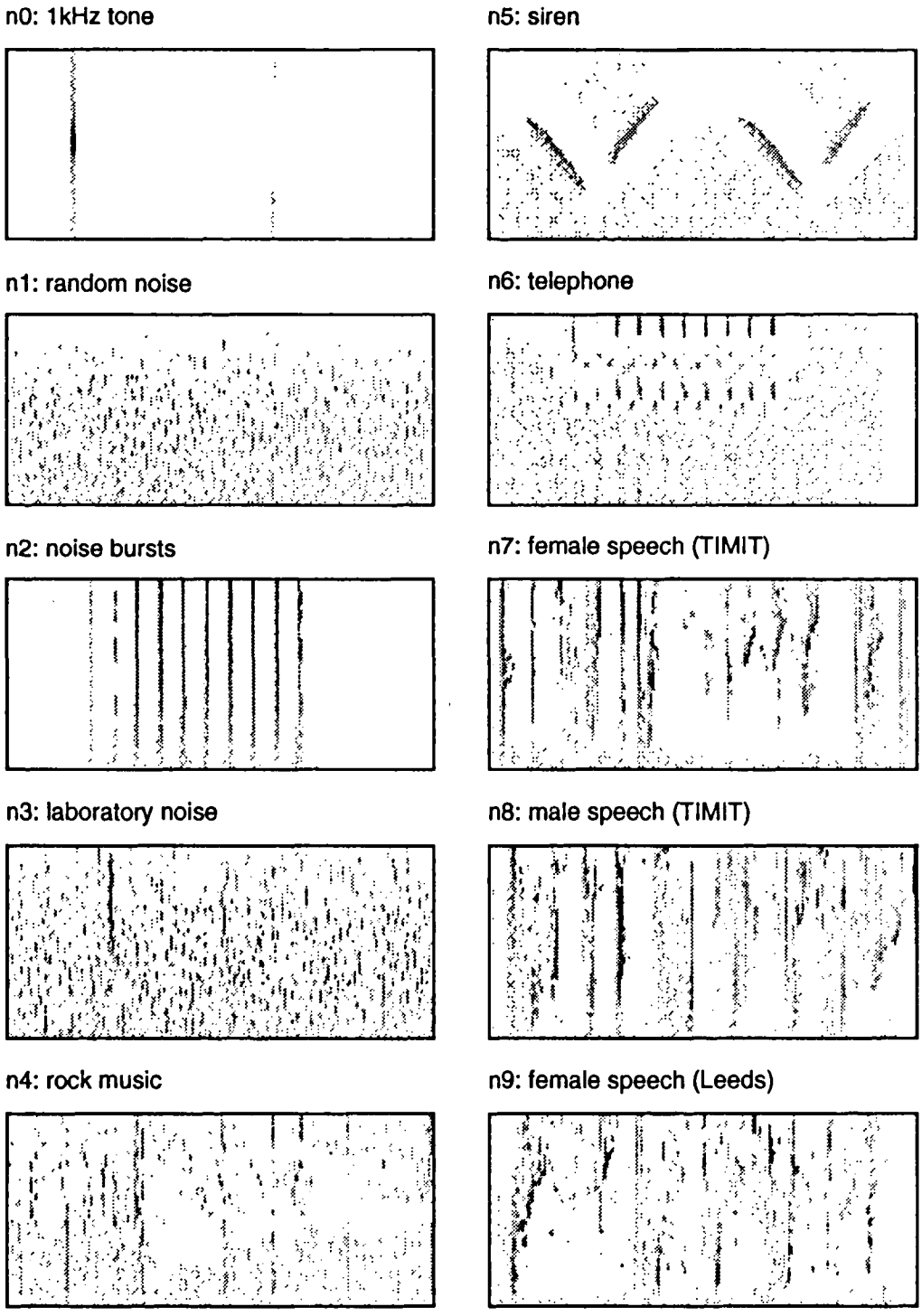

Figure 4.17: Onset map representations of the ten noise sources. Time is represented on the abscissa, channel center frequency (50Hz-5kHz) is represented on the ordinate.

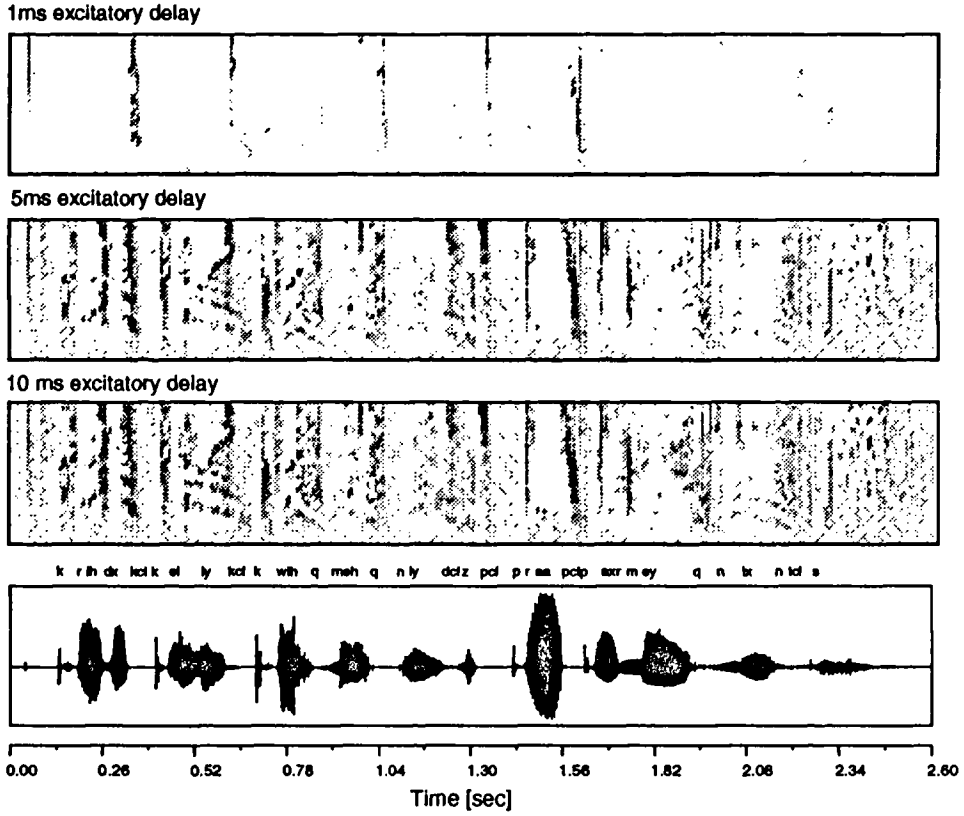

Figure 4.18: Offset map representations of speech, showing the effect of varying the delay before excitation. See the text for details.

this depends on the sweep rate), and hence the response of a short-delay cell will be small.

In fact, neither of these solutions needs to be adopted here, because the scene analysis algorithm presented later is unlikely to be affected by "phantom onsets" caused by a moving dominance. This point is discussed again in chapter 5.

4.3.5 A Model Offset Map

The mechanism of cells which respond at the offset of an applied stimulus is not well understood, and little physiological data is available to inform modelling studies. Intuitively, detecting an offset of energy is rather like the "reverse" of detecting an onset, which suggests that offset cells might receive their excitatory and inhibitory inputs in the opposite order to onset cells. Hence, an offset cell mechanism is proposed here in which excitation is delayed relative to inhibition.

For example, consider the response of an offset cell to a tone burst at its characteristic frequency. At the start of the stimulus, the cell immediately receives strong IPSPs, followed by delayed EPSPs which are swamped by the inhibition. Hence, the cell is unable to fire for the duration of the tone. However, when inhibition ceases at

the end of the stimulus, delayed EPSPs continue to arrive at the cell, raising the membrane potential above threshold and causing a spike to be generated.

Modifying equation 4.14 so that the delayed input is excitatory, the membrane potential $p_{off}[t]$ of an offset cell is given by

$$p_{off}[t] = p_{off}[t-1]c_d + E_{psp}r[t-\Delta t_E] - I_{psp}r[t] \tag{4.18}$$

where c_d is defined by equation 4.15, and the firing rate $s_{off}[t]$ of the cell is given by

$$s_{off}[t] = \begin{cases} p_{off}[t] & \text{for } p_{off}[t] > Th \\ 0 & \text{otherwise} \end{cases} \tag{4.19}$$

Except for Δt_E, the parameters here are the same as those in equations 4.14-4.16, and are set to the values in table 4.3 by similar reasoning. As discussed in section 4.3.2, the delay before excitation Δt_E is a tuned parameter in the offset map, which determines the rate of amplitude fall that each cell is most sensitive to. Figure 4.18 shows offset map representations of speech for several values of Δt_E. Abrupt offsets caused by closures (/kcl/ and /pcl/) are identified by the map with a short (1 ms) excitatory delay, whereas the more gradual offsets of vowels generate a larger response in the maps with long excitatory delays (5 ms and 10 ms).

4.3.6 Offset Map Representations

Figure 4.19 shows offset map representations of the ten noise sources used in chapter 6, for offset cells with excitatory delays of 5 ms. The 1kHz tone, noise bursts, telephone and speech all have abrupt offsets, and these are clearly delineated in the maps.

The siren presents a similar problem for the offset map as it does for the onset map. As the concentration of energy moves out of the response area of one auditory filter and into the response area of the next, a rapid decrease in activity occurs which is detected by the offset map. Although these "phantom offsets" do not present a problem for the scene analysis algorithm described in chapter 5, they could be eliminated by modifying the offset model in a similar way to the onset model, so that inhibition arises from frequencies surrounding the characteristic frequency of the cell. Hence, the modified offset cell membrane potential is given by

$$p_{off}[t] = p_{off}[t-1]c_d + E_{psp}r[cf, t-\Delta t_E] - I_{psp}\sum_{f=-N}^{N} r[cf+f, t] \tag{4.20}$$

where the new parameters N and $r[cf,t]$ are the same as those described for the modified onset cell model in equation 4.17.

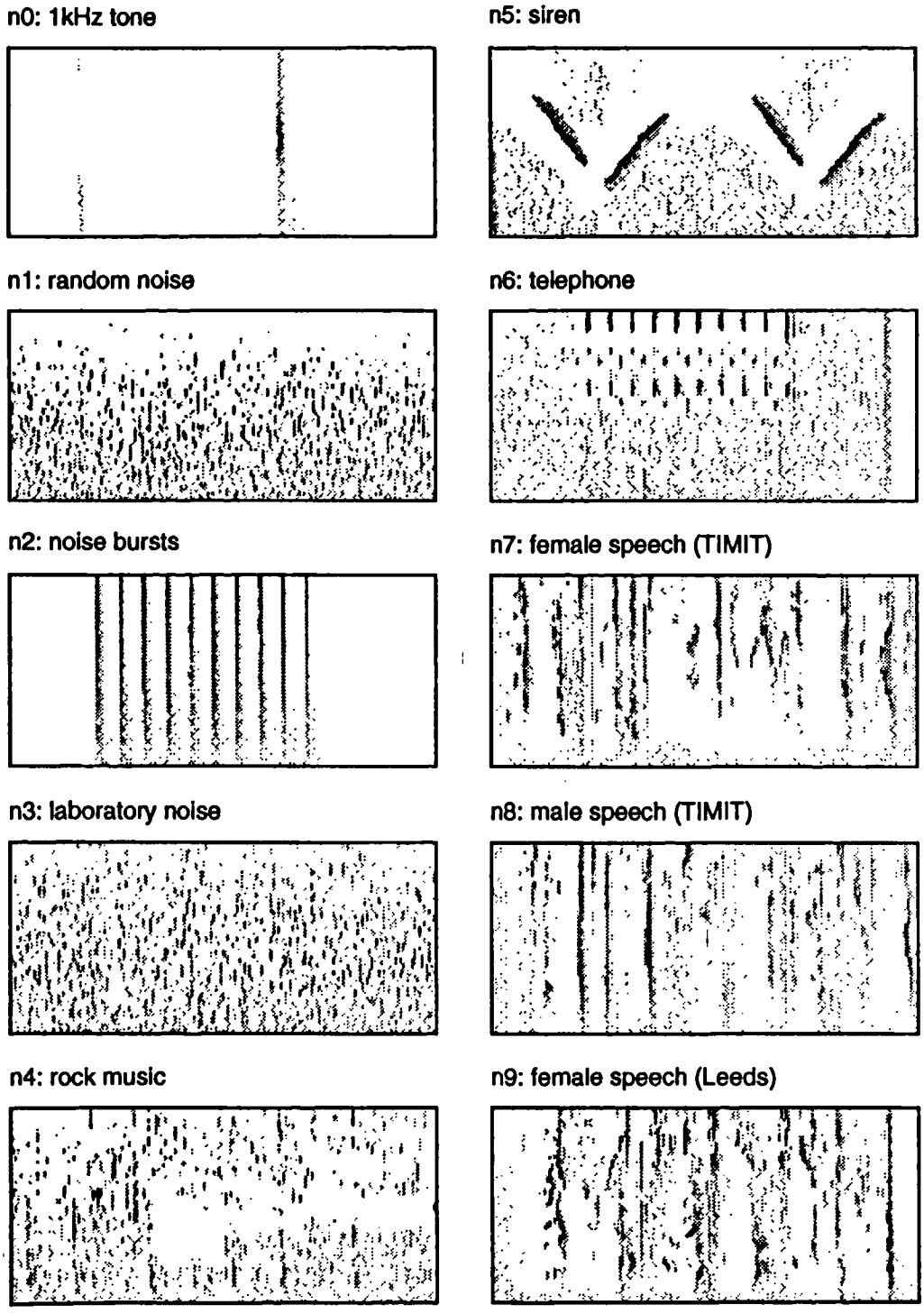

Figure 4.19: Offset map representations of the ten noise sources. Time is represented on the abscissa, channel center frequency (50Hz-5kHz) is represented on the ordinate.

4.3.7 Summary and Discussion

In this section, physiologically-motivated models of onset and offset maps have been described. The maps are effective at identifying the times at which auditory events start and stop, and will provide the basis for grouping synchronous events together in the algorithm presented in chapter 5.

A number of simplifying assumptions have been made in the model presented here. Firstly, it was considered advantageous from an information-processing point of view to separate onset cells and offset cells into two distinct populations (maps). However, it should be noted that onset and offset responses can be recorded from the *same* auditory neuron, depending on the intensity of the stimulus (Shofner and Young [250]). Hence, it is doubtful whether the auditory organization of onset and offset cells is as discrete as the hypothetical maps suggest.

In fact, offset responses can be observed in the onset map and vice versa, as illustrated in the map representations of the 1 kHz tone in figures 4.17 and 4.19. Following the onset of the tone, rapid adaptation occurs in the auditory nerve which registers as an offset. Similarly, when the tone ends there is an increase in activity as the auditory nerve recovers to its spontaneous rate, which registers as an onset. However, these effects are small and do not present a problem for the scene analysis algorithm described later.

Secondly, it has been assumed that the excitatory and inhibitory inputs to the cell models arise from the same characteristic frequency. Although this assumption is physiologically unrealistic, it simplifies the models considerably. A minor problem is that a moving dominance can generate "phantom" activity in the onset and offset maps, but this is unlikely to affect the scene analysis algorithm.

Finally, the architecture of the maps assumes that onsets and offsets are detected by within-channel mechanisms, and that higher-level processes look across frequency channels to identify synchronous components. However, it is possible that low-level neural mechanisms exist which operate across different frequency regions in order to detect common onsets and common offsets directly. Clearly, further physiological evidence is needed to clarify this point.

4.4 Frequency Transition

In this section, a computational model of a frequency transition map is presented. First, the psychophysical and physiological evidence that supports the model is reviewed.

4.4.1 Psychophysical Motivation

An early problem facing perceptual grouping mechanisms is how to match the auditory representation of an acoustic event at a particular time with the representation of the same event at a later time. This task is the auditory analogue of the *correspondence problem* which arises in the perception of visual motion (Ullman [271]).

It is likely that the auditory system uses two cues, *frequency proximity* and *alignment on a common time-frequency trajectory*, to solve the correspondence problem (Tougas and Bregman [270]). These are analogous to the Gestalt principles of proximity and good continuation that were discussed in section 3.1. Since many natural sounds (such as speech) consist of glides in frequency, it might be supposed that trajectory is an important grouping cue. The following sections present evidence that the auditory system measures frequency transitions, and that it uses this information to group frequency components across time according to their trajectories.

Evidence that the Auditory System Codes Frequency Transition

Steiger and Bregman [258] have investigated the auditory coding of frequency transition using a stimulus similar to the one in figure 4.11. Instead of using static tones, a gliding tone A was alternated with a pair of gliding tones B and C. This pattern was repeatedly presented to listeners, who were asked to judge whether A and B were in the same perceptual stream. When A and B glided in frequency at the same rate, A tended to pull B into a sequential stream so that C was heard as a separate tone. However, when the glide rates of A and B differed, B and C tended to fuse into a rich-sounding complex which alternated with A. This result suggests that the auditory system is able to measure the rate of a frequency transition, and that components which would otherwise belong to the same stream can be segregated if they change in frequency at a different rate.

Other evidence for the auditory coding of frequency transition has come from psychophysical adaptation studies. These experiments attempt to demonstrate that a particular sensory "channel" exists as a neural organization by trying to fatigue it with an appropriate conditioning stimulus. Kay and Matthews [128] have found evidence for channels sensitive to frequency modulation using an adaptation paradigm. They presented listeners with a conditioning tone which was sinusoidally frequency modulated at a particular rate, followed by a test tone whose extent of frequency deviation was varied to find the point where the modulation was just detectable.

When the modulation rates of the conditioning tone and test tone were similar, the ability of listeners to detect the modulation in the test tone was greatly reduced. For modulation frequencies between 3Hz and 30Hz, the frequency deviation of the test tone had to be increased to three times that of the conditioning tone before the modulation became detectable. Furthermore, this effect was tuned to modulation rate, so that when the rates of the test tone and the conditioning tone differed by a few Hz, the adaptation effect was reduced.

Kay and Matthews also demonstrated that adaptation was largely independent of the carrier frequencies of the conditioning and test tones, and that the frequency modulation channels were not adapted by amplitude modulated stimuli which had the same periodicity and spectral shape. Similarly, square-wave frequency modulation (abrupt changes in frequency) did not affect the detectability of sinusoidal modulation (Green and Kay [101]). Hence, the channels appear to be tuned to gliding transitions in frequency, rather than some other property of the stimulus. Indeed, channels which are sensitive to other acoustic features may also exist. Regan and Tansley [215] and Tansley and Suffield [269] have found evidence for channels tuned to amplitude modulation, which are similar to frequency modulation channels in their adaptation properties.

The interpretation of Kay and Matthew's experiment is complicated by the fact that it used rapid, periodic frequency modulations of a type which seldom occur in environmental sounds. Gardner and Wilson [91] have investigated frequency modulation channels using more natural stimuli. They presented listeners with tones which swept upward or downward in frequency, and were similar in duration and range to second formant transitions of speech. The threshold for single upward sweeps was increased by a factor of two or three following conditioning by repetitive upward sweeps, but was unaffected by repetitive downward sweeps. Similarly, conditioning by repetitive downward sweeps only increased the threshold for downward sweeps. These results suggest the existence of channels which are specifically tuned to upward or downward frequency transitions. Channels of this kind could provide the basis for the findings of Steiger and Bregman that were discussed earlier.

It should be mentioned that adaptation studies have been the subject of some controversy. For example, Moody *et al.* [178] replicated Gardner and Wilson's experiment and found that the differences between adapted and nonadapted thresholds decreased with continued testing. This result, and a similar investigation by Wakefield and Viemeister [275], suggest the involvement of cognitive factors rather than the adaptation of a neural organization. However, Rees and Kay [213] have presented evidence for the existence of frequency modulation channels without using an adaptation paradigm, so the argument for auditory mechanisms which detect frequency transition remains quite convincing.

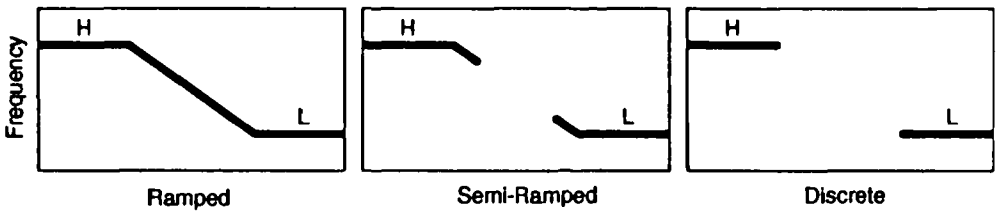

Figure 4.20: Stimuli used by Bregman and Dannenbring, showing the three types of transition between a high frequency tone (H) and a low frequency tone (L). Adapted from Bregman and Dannenbring [28].

Evidence that the Auditory System Uses Frequency Transition

In the previous section, evidence that the auditory system measures frequency transitions was reviewed. Here, evidence is presented that perceptual mechanisms use frequency transitions to group components which lie on a common trajectory.

Bregman and Dannenbring [28] have investigated the role of trajectory in the perception of rapid sequences of alternating high and low frequency tones. They presented listeners with cycles of the three stimuli shown in figure 4.20. In the *ramped* condition, there was a continuous frequency transition between the high and low tones, whereas in the *discrete* condition there was no transition. A *semi-ramped* stimulus was also used, in which there was a partial transition between the two tones. Listeners tended to report that the discrete stimulus segregated into a stream of high frequency tones and a stream of low frequency tones, whereas the ramped stimulus was heard as one stream and the semi-ramped condition gave an intermediate effect. Bregman and Dannenbring concluded that the partial transition in the semi-ramped stimulus helped to pull the two tones into the same stream by "pointing" to the frequency region of the next tone. They also observed that transitions similar to the semi-ramped stimulus occur in speech, and suggested that these trajectories might prevent speech sounds with different spectral characteristics from segregating into different streams.

In a related experiment, Cole and Scott [51] have provided direct evidence for the role of trajectories in speech perception. They asked listeners to judge the order of consonant-vowel syllables in a repeating sequence, for normal syllables and for syllables in which the smooth transitions in frequency between the consonant and the vowel had been removed. Subjects found it more difficult to judge the order of the transitionless syllables, and heard the sequence segregate into separate perceptual streams consisting of consonants and vowels. This result supports Bregman and Dannenbring's hypothesis that trajectories are important in binding speech sounds into a unified stream.

When part of a tone is deleted and replaced by a louder burst of random noise, listeners hear the tone continuing through the noise, even though the tone is not physically present. This phenomenon is known as the *auditory continuity effect*,

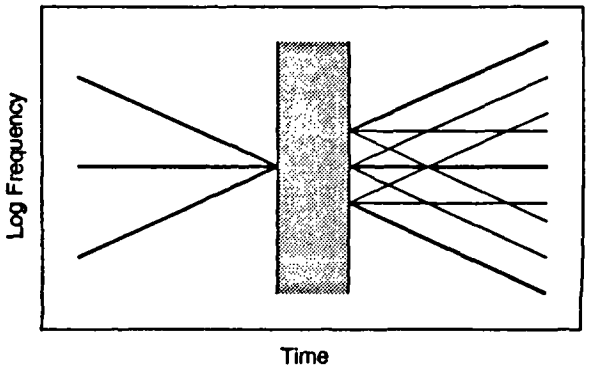

Figure 4.21: Stimuli used by Ciocca and Bregman. The slope of tones entering and exiting a noise burst (grey rectangle) are varied. Subjects report the best continuity though the noise when the prenoise and postnoise tones lie on a common trajectory (bold lines). Redrawn from Ciocca and Bregman [49].

and is an example of the Gestalt principle of closure (see section 3.1). Ciocca and Bregman [49] have presented evidence that the trajectory of the tone preceding and following the noise burst affects the perception of continuity. They played listeners the stimuli shown in figure 4.21. The tone preceding the noise either swept upward in frequency, swept downward in frequency or was static. The slope of the tone following the noise was varied in the same way, but the starting frequency was either the same, higher or lower than the frequency at which the preceding tone entered the noise burst. For each stimulus, listeners were asked to judge how strongly the tone continued through the noise. When the slope of the prenoise and postnoise tones was the same, the strongest continuity was reported for tones which lay along a common trajectory (dark lines in the figure). Ciocca and Bregman concluded that, under certain circumstances, the auditory system is capable of extrapolating time-frequency trajectories in order to make a good continuation. Recently, Kluender and Jenison [136] have performed similar experiments and come to the same conclusion.

It should be pointed out that some of the results reviewed in this section are open to interpretation. For example, Steiger and Bregman [258] have noted that in the semi-ramped stimulus used by Bregman and Dannenbring (figure 4.20), the end of the high frequency tone and the start of the low frequency tone are closer together than they are in the discrete stimulus. Hence, the observed results could be explained by a frequency proximity effect as well as by a frequency trajectory effect. Additionally, several investigations of auditory continuity have failed to find an effect of trajectory (Steiger and Bregman [258], Tougas and Bregman [270], Dannenbring [56]). However, all of these experiments used stimuli consisting of repeating cycles, whereas the studies reviewed here used a single presentation of each stimulus and found a trajectory effect. Since natural sounds (such as speech) do not repeat in a cyclical manner, it would be unwise to discount trajectory effects on the basis of null results from experiments which use repetitive patterns (Kluender and Jenison [136]).

4.4.2 Physiological Motivation

Physiology of Single Cells

Some of the most extensive evidence for the coding of frequency transition by auditory neurons has come from studies on the bat, which uses frequency modulated tones for echolocation (Suga [261]). However, frequency transition also appears to be coded by non-echolocating mammals. In the cat and rat, neurons sensitive to frequency transition have been identified in the cochlear nucleus (Møller [177]), inferior colliculus (Watanabe and Ohgushi [276]) and auditory cortex (Whitfield and Evans [279]). Generally, neurons which respond to frequency modulated stimuli are more prevalent at higher levels of the auditory pathway. Indeed, the majority of cells in the auditory cortex respond preferentially to modulated sounds, and some do not respond to static stimuli at all (Whitfield and Evans [279]).

The psychophysical adaptation experiments of Kay and Matthews, discussed in section 4.4.1, suggest the existence of channels which are tuned to a particular rate of sinusoidal frequency modulation. There is good physiological evidence to support their findings. For example, Whitfield and Evans [279] have described neurons in the auditory cortex which respond optimally to particular modulation rates in the range 2-20 Hz, which is similar to the range of modulation rates which Kay and Matthews used in their study. Also, Britt and Starr [33] and Mendelson and Cynader [171] have described neurons that are tuned to the rate of discrete frequency sweeps.

Similarly, Gardner and Wilson found psychophysical evidence for channels tuned to upward and downward frequency transitions. Neurons which respond selectively to the direction of a frequency sweep have been identified in the auditory cortex by Whitfield and Evans [279] and Mendelson and Cynader [171]. These neurons may provide a physiological basis for Gardner and Wilson's findings.

Again, it must be emphasized that the results of single cell studies should be interpreted with caution. Although it is reasonable to suppose that neurons which respond to specific rates and directions of frequency transition are coding this information in some way, there is no *direct* evidence linking cell responses to perceptual effects.

Topographic Organization

It was argued in section 2.2.3 that acoustic parameters which are perceptually important are likely to be represented within an orderly framework in the higher auditory system. As expected, frequency transition appears to be mapped in this way.

Mendelson and his colleagues have investigated the distribution of neurons sensitive to frequency transition in the auditory cortex of the cat (Mendelson and Cynader [171], Schreiner *et al.* [238], Mendelson *et al.* [172]). They found that neurons sensitive to upward or downward frequency sweeps were concentrated in different

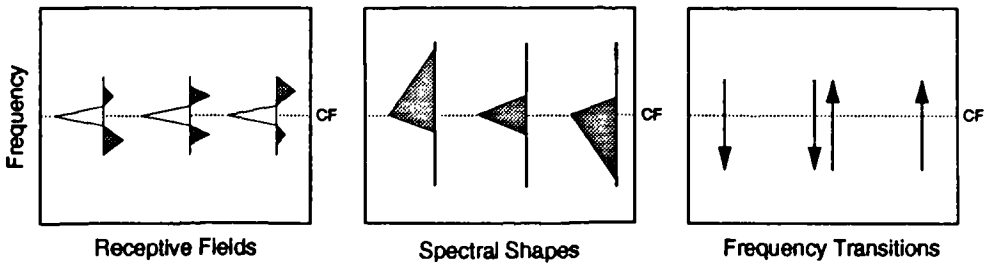

Figure 4.22: Map of frequency transition. As a result of the differences in inhibitory strength above and below characteristic frequency (CF), each cell in the map is tuned to a spectral shape (centre panel) and direction of frequency transition (right panel). Shaded triangles in the left panel represent inhibition, open triangles represent excitation. Redrawn from Shamma and Chettiar [247].

parts of the cortex, and that preferred sweep rate varied systematically across a smooth gradient.

Similar findings have been reported by Shamma and Chettiar [247] and Shamma *et al.* [249] for the auditory cortex of the ferret. Their results suggest that the preference of neurons for particular directions of frequency transition varies systematically along iso-frequency planes in the cortex, as shown in figure 4.22. The selectivity of a cell for the direction of a frequency sweep correlates with the strength of the inhibitory inputs that the neuron receives from above and below its characteristic frequency. For example, neurons sensitive to upward sweeps have strong inhibition at frequencies above the characteristic frequency.

A consequence of this organization is that neurons in the map also respond preferentially to spectral shapes which have least energy at the frequencies where inhibition is strongest (centre panel in figure 4.22). Hence, it is possible that the map codes information about the *gradient* of the acoustic spectrum as well as the direction of frequency transitions. As such, the map is an auditory analogue of the edge orientation columns identified in the visual cortex by Hubel and Wiesel [120]. This point will be mentioned again in the following section.

4.4.3 A Model Frequency Transition Map

In this section, a computational model of a map of frequency transition is presented which is motivated by the psychophysical and physiological evidence reviewed earlier. The map provides a solution to the correspondence problem by extracting information about the movement of spectral peaks across time and frequency from firing patterns in the auditory nerve. It should be emphasized that the map is *not* intended to provide a basis for grouping acoustic components which share a common rate of frequency modulation. This point is discussed in section 4.5.2.

A schematic of the frequency transition map is shown in figure 4.23. Cells in the map

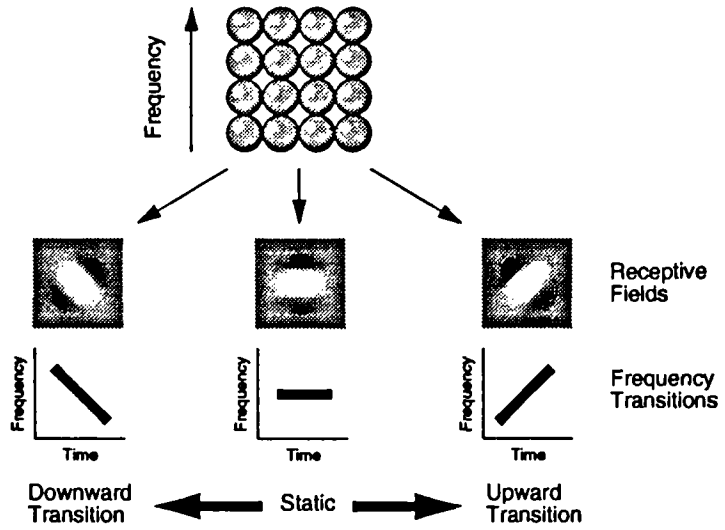

Figure 4.23: Schematic of the frequency transition map. Cells in the map are tuned to different rates of frequency transition, according to the time-frequency orientation of their receptive fields.

are arranged in a two-dimensional framework, with characteristic frequency represented on one axis and frequency transition represented on the other. Each neuron is tuned to a particular rate and direction of frequency sweep, depending on the orientation of its receptive field. Similar schemes have been proposed by Mellinger [170] and, in a non-auditory context, by Riley [220]. Also, some preliminary work on the map has been reported in Brown and Cooke [35].

The firing rate of each neuron in the map is determined by convolving its receptive field with the simulated auditory nerve response. Hence, for a cell with characteristic frequency f and receptive field orientation θ, the firing rate $s[t, f, \theta]$ at time t is given by

$$s[t, f, \theta] = \sum_{i=-N}^{N} \sum_{j=-M}^{M} r[t+i, f+j] g_\theta[i, j] \tag{4.21}$$

where $2N$ and $2M$ define the width in time and frequency of the receptive field $g_\theta[\cdot]$, and $r[\cdot]$ represents the probability of firing in the auditory nerve. The parameters of the model are discussed in detail below, and their values are summarised in table 4.4.

Form of the Receptive Field

Each neuron in the map is required to be tuned to a particular rate of frequency transition. This implies that the receptive field of a cell must elicit a maximal

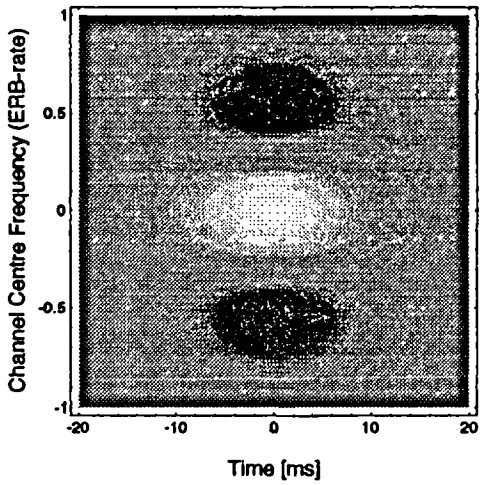

Figure 4.24: Form of the receptive field employed in the map. The function consists of an excitatory central lobe (light area), and two flanking inhibitory lobes (dark areas) which confer directional selectivity.

response when it is aligned with a dominance which is moving at the cell's preferred rate. Several forms of receptive field could be used to satisfy this condition. The one used here is based on a function suggested by Riley [220], defined as

$$g[t,f] = -\frac{\partial^2}{\partial f^2}G[t,f] \tag{4.22}$$

$$= \frac{1}{\pi\sigma_f^2}G[t,f] - \frac{f^2}{[\pi\sigma_f^2]^2}G[t,f] \tag{4.23}$$

where G[t,f] is a two-dimensional Gaussian

$$G[t,f] = \exp\left[-\frac{t^2}{2\pi\sigma_t^2} + \frac{f^2}{2\pi\sigma_f^2}\right] \tag{4.24}$$

of standard deviation σ_t and σ_f in time and frequency. A plot of the receptive field $g[t,f]$ is shown in figure 4.24. It consists of a central excitatory (positive) region, and two flanking inhibitory (negative) regions which confer directional selectivity. The width of the receptive field was taken to be eight standard deviations, therefore $N = 4\sigma_t$ and $M = 4\sigma_f$ in equation 4.21.

In the form presented in equations 4.22-4.24, $g[t,f]$ responds maximally when it is centred on a dominance that is static in frequency, such as a pure tone (see figure 4.23). Receptive fields that are tuned to particular rates and directions of frequency transition are obtained by rotating $g[t,f]$ in the time-frequency plane. The operator

$$R_\theta[t,f] = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} t \\ f \end{bmatrix} \tag{4.25}$$

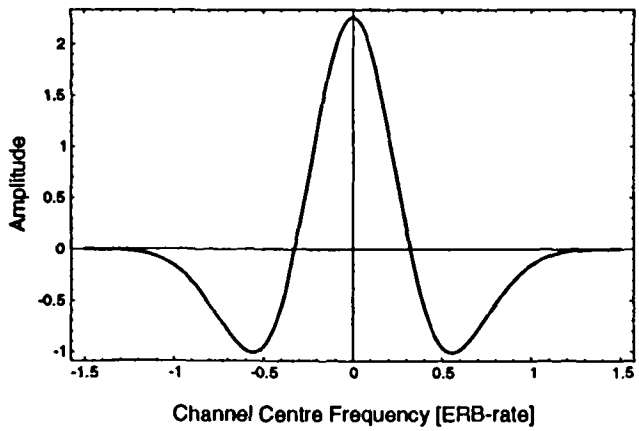

Figure 4.25: Slice through $g[t, f]$ at $t = 0$. The arrangement of a central excitatory region and two inhibitory regions is similar to the receptive fields in the frequency transition map described by Shamma.

rotates a point by θ radians in time and frequency, and thus the receptive field at a particular orientation may be written as

$$g_\theta[t, f] = gR_\theta[t, f] \tag{4.26}$$

For example, figure 4.23 shows three receptive fields at rotations of $-\pi/4$, 0 and $\pi/4$ radians, which are tuned to downward sweeping, static and upward sweeping dominances respectively.

The form of receptive field used here can be justified in two ways. Firstly, a cross-section in frequency through $g[t, f]$ reveals a pattern of excitation and inhibition which is similar to the receptive fields of neurons in the frequency transition map described by Shamma (figures 4.22 and 4.25). Hence, the map proposed here can be seen as an extension of Shamma's map into two dimensions, in which spectral gradients are measured across time as well as across frequency. Secondly, $g[t, f]$ is a physiologically plausible function. As discussed in section 4.4.2, the visual analogue of a map of frequency transition is the edge-orientation map identified by Hubel and Wiesel [120] in the striate cortex. Neurons in the edge-orientation map have receptive fields which are very similar to the ones used here (figure 4.26).

Relative Width in Time and Frequency

The parameters σ_t and σ_f in equations 4.23 and 4.24 determine the width of the receptive field in time and frequency. The choice of σ_t and σ_f depends on a tradeoff between time resolution and frequency resolution, which has been expressed as the *uncertainty principle*

$$\Delta t \Delta f \simeq 1 \tag{4.27}$$

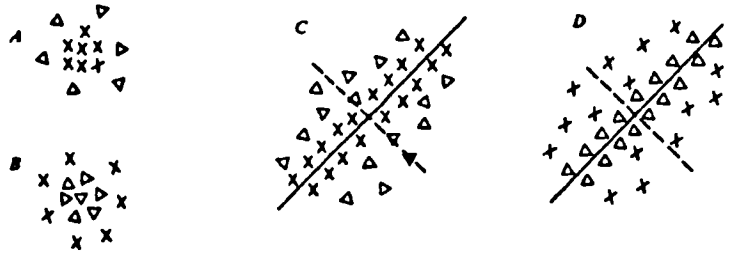

Figure 4.26: Receptive fields of edge-orientation sensitive cells in the visual cortex. Type C has a central excitatory lobe (crosses) and flanking inhibitory lobes (triangles), similar to the receptive field used here. From Hubel and Wiesel [120], figure 2.

by Gabor [87]. This identity states that a fine frequency analysis is simultaneously associated with a coarse time analysis, and vice versa. In the specific case here, the values of σ_t and σ_f reflect a compromise between time-frequency localization and directional selectivity (Riley [220]). When $\sigma_t \neq \sigma_f$, the receptive field has good selectivity in time-frequency, which gives it an advantage in separating crossing dominances. Such situations can arise when formants "cross" in natural speech (Kuhn [142]), and when several acoustic sources are active at the same time. However, when $\sigma_t = \sigma_f$ optimum localization in time-frequency results, and bends in the trajectory of a dominance are resolved more effectively. Since many environmental sounds change rapidly in frequency (such as the speech and siren noise sources), accurate localization was considered important and therefore σ_t was set to the same value as σ_f.

Absolute Width in Frequency

When considering the width of the receptive field across frequency, it is convenient to assume that the spacing of auditory filters is *logarithmic*. On a logarithmic scale, frequency transitions which move at the same rate have the same slope, irrespective of the initial and final frequency. Hence, it can be assumed that neurons centred on different characteristic frequencies in the map have receptive fields which occupy the same number of auditory filter channels, and sweep rates can be specified in convenient units such as octaves per second (oct/sec).

Clearly, this assumption is an approximation because the auditory filters in the model are spaced on an ERB-rate scale, which is not perfectly logarithmic (see section 4.1.2). Figure 4.27 shows the relationship between channel number and channel centre frequency for filters spaced on logarithmic and ERB-rate frequency scales. Although similar, the ERB-rate scale is rather more linear than a strictly logarithmic scale. This discrepancy would present a problem if the map formed a basis for grouping components with common rates of frequency modulation, be-

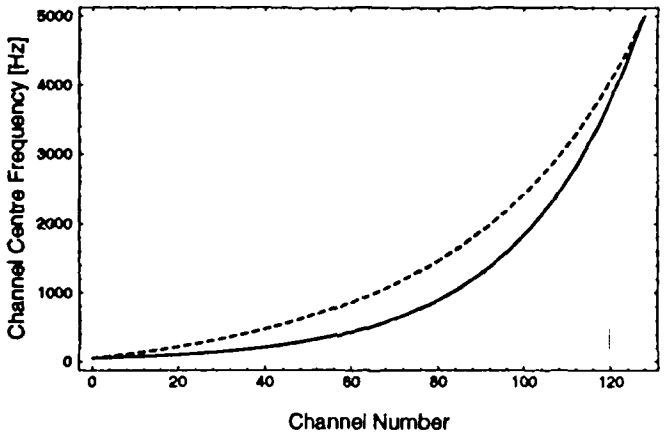

Figure 4.27: Comparison of auditory filter spacing on logarithmic (solid line) and ERB-rate (dotted line) frequency scales. Although similar, the ERB-rate scale is rather more linear.

cause dominances moving at the same rate in different frequency regions will not have exactly the same slope. However, the map is concerned only with tracking dominances across time, so the error between the ERB-rate and logarithmic scales was considered acceptable.

Unfortunately, there is little psychophysical or physiological data to inform the choice of frequency width for the receptive field. However, there are some practical considerations. Receptive fields that are wide in frequency do not localize spectral peaks as well as receptive fields which are narrow in frequency. Conversely, the narrowness of a receptive field in frequency is limited by the number of auditory filters used in the model. For a filterbank with 128 channels in the range 50 Hz to 5 kHz, a frequency spread of 7 channels (approximately 1.4 ERB) was found to be a good compromise. In this case, the central excitatory lobe of $g[t, f]$ occupies 3 channels, and the inhibitory lobes occupy 2 channels either side.

Absolute Width in Time

Several factors determine the choice of time width for the receptive field. Firstly, it is desirable for the receptive field to be at least as wide as the longest pitch period expected for a periodic source, otherwise it will be integrating auditory nerve activity over an uneven temporal window. Since the lowest pitch expected is 50 Hz (see section 4.2.3), this suggests a lower limit of 20 ms for the time width. An experiment by Steiger, reported in Bregman [24], suggests an upper limit. Steiger found that listeners perceived short (50 ms) glides that were aligned on a common trajectory and separated by 10 ms noise bursts as a continuous sweep in frequency. However, if static tones were substituted for the glides, subjects heard an uneven "staircase" rising in frequency rather than a smooth transition. Hence, it appears

Parameter	Description	Value	Units
N	half-width in time of receptive field	15	ms
M	half-width in frequency of receptive field	0.7	ERB
σ_t	standard deviation in time	$N/4$	ms
σ_f	standard deviation in frequency	$M/4$	ERB
	range of frequency transition rates	-20 to 20	oct/sec
	spacing of frequency transition rates	1.82	oct/sec

Table 4.4: Parameter settings for the frequency transition map.

that the auditory system is able to measure the slope of a frequency glide in 50 ms or less.

Similar findings have been reported by Nabelek and Hirsh [190], who found that the ability of listeners to discriminate between different rates of frequency transition was optimal for sweep durations of 30 ms. Additionally, formant transitions in speech usually occur over a duration of 30-40 ms (Lehiste and Peterson [146]), and Liberman *et al.* [149] have shown that transitions in this range are important for the identification and discrimination of different speech sounds.

Consequently, the time width of the receptive field was set to 30 ms in the model. Since the frequency spread was 7 auditory filter channels, 30 ms of auditory nerve firings were collapsed into 7 bins in order to give the receptive field an equal width in time and frequency.

Range of Frequency Transition Rates

Since the map is intended to detect frequency transitions in many types of environmental sounds, it is necessary to know the maximum rate at which sweeps in frequency are likely to occur. Some of the most rapid changes in frequency are observed in formant transitions of speech, which can be as fast as 50 oct/sec (Liberman *et al.* [150]). However, the majority of formant transitions occur at rates of less than 20 oct/sec (Lehiste and Peterson [146]), and this seemed a reasonable upper limit to use in the map. Hence, neurons were tuned to a maximum upward transition rate of 20 oct/sec, and a maximum downward transition rate of -20 oct/sec.

Spacing of Receptive Fields

A final consideration is the distribution of receptive fields along the frequency transition axis of the map. If many closely-spaced receptive fields are used, nearby neurons will be tuned to very similar rates of frequency sweep, and the map will be redundant. Conversely, if the receptive fields are spaced too far apart, the range of frequency transition rates will not be adequately covered.

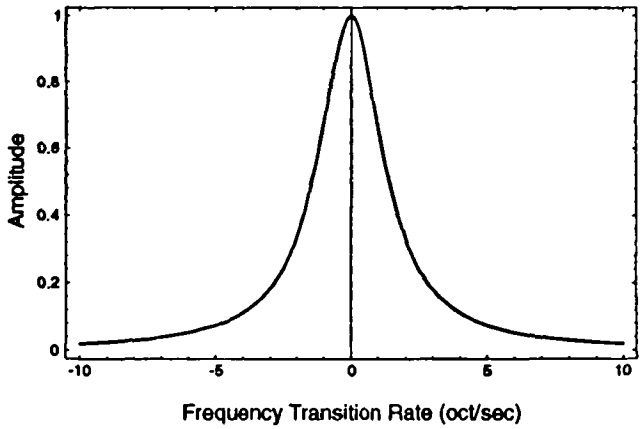

Figure 4.28: Plot of the tuning curve $\Gamma[\theta]$ for the receptive field $g[t, f]$, where θ has been converted into oct/sec. The receptive field elicits a maximum response at the sweep rate to which it is tuned (0 oct/sec).

In psychophysical terms, the spacing between receptive fields in the map is equivalent to the smallest difference in frequency transition rate that listeners are able to detect. Unfortunately, it is difficult to measure a "pure" difference in detectability between two sweep rates, because sweeps at different rates also differ in their duration or extent of frequency deviation. Hence, listeners may use time or frequency differences to distinguish between two frequency transitions, rather than differences in sweep rate *per se* (Pollack [207]).

Since there is a lack of reliable psychophysical data, the spacing of the receptive fields in the map has to be determined by practical considerations. The approach used here is to define a *tuning curve* for the receptive field $g[t, f]$, which quantifies its selectivity to different rates of frequency sweep. Riley [220] has shown that the tuning curve $\Gamma[\theta, \xi]$ for $g[t, f]$ is given by

$$\Gamma[\theta, \xi] \propto \frac{\cos^2 \theta}{\sqrt{1 + [\xi^2 - 1] \sin^2 \theta}} \tag{4.28}$$

where

$$\xi = \frac{\sigma_t}{\sigma_f} \tag{4.29}$$

and θ is the slope (in radians) of a pure tone rising linearly in log frequency. Since σ_t and σ_f are set to the same value in the model, ξ is unity and 4.28 reduces to

$$\Gamma[\theta] = \cos^2 \theta \tag{4.30}$$

A plot of the tuning curve $\Gamma[\theta]$ is shown in figure 4.28, where θ has been recalculated as sweep rate in oct/sec. As expected, the receptive field elicits an optimal response

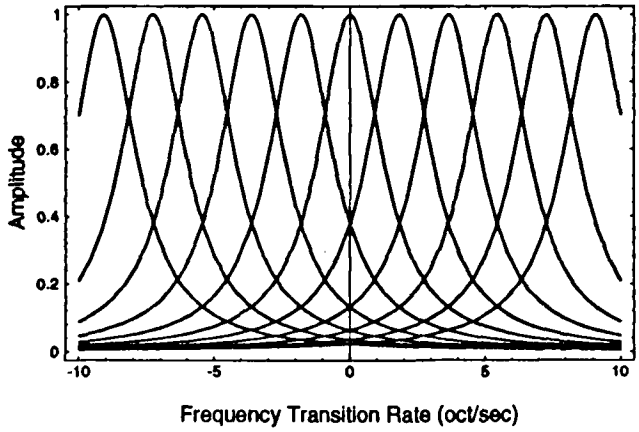

Figure 4.29: Receptive fields are spaced so that their tuning curves overlap at their 3dB points. A reduction of 3dB from the peak response occurs at a sweep rate of 0.91 oct/sec.

when it is aligned with a frequency transition which moves at its preferred rate (0 oct/sec, a static tone).

When spacing bandpass filters, it is conventional to overlap them at a frequency where the output of each filter has fallen by 3 dB relative to its output in the passband. This point corresponds to a reduction in power by a factor of 2, and a reduction in amplitude by a factor of $\sqrt{2}$. The same principle can be used here, since $\Gamma[\theta]$ is a bandpass function. Hence, the receptive fields are spaced so that their tuning curves overlap at the sweep rate where the response has fallen by a factor of $\sqrt{2}$ relative to the response at the neuron's preferred rate. Solving from figure 4.28, the amplitude of response falls to $1/\sqrt{2}$ at a sweep rate of 0.91 oct/sec. Therefore, neurons in the map are tuned to rates of frequency transition at intervals of 2 × 0.91 = 1.82 oct/sec, in order to make their tuning curves overlap at their 3 dB points (figure 4.29).

4.4.4 Frequency Transition Map Representations

The siren noise source provides a good demonstration of the frequency transition map. Plots of the distribution of activity in the map for the first cycle of the siren, taken at 20 ms intervals, are shown in figure 4.30. This diagram can be interpreted by referring to the auditory nerve representation of the siren (figure 4.1) and the schematic of the map shown in figure 4.23. Initially, the concentration of energy is in the centre of the map, indicating that the siren is static in frequency. As the siren falls in frequency, neurons tuned to downward sweeps are activated on the left of the map. Subsequently, the concentration of energy moves to the right of the map as the siren sweeps upward in frequency. Hence, for a sinusoidally frequency modulated tone, the peak of activity in the map follows an elliptical path over time.

Frequency Transition

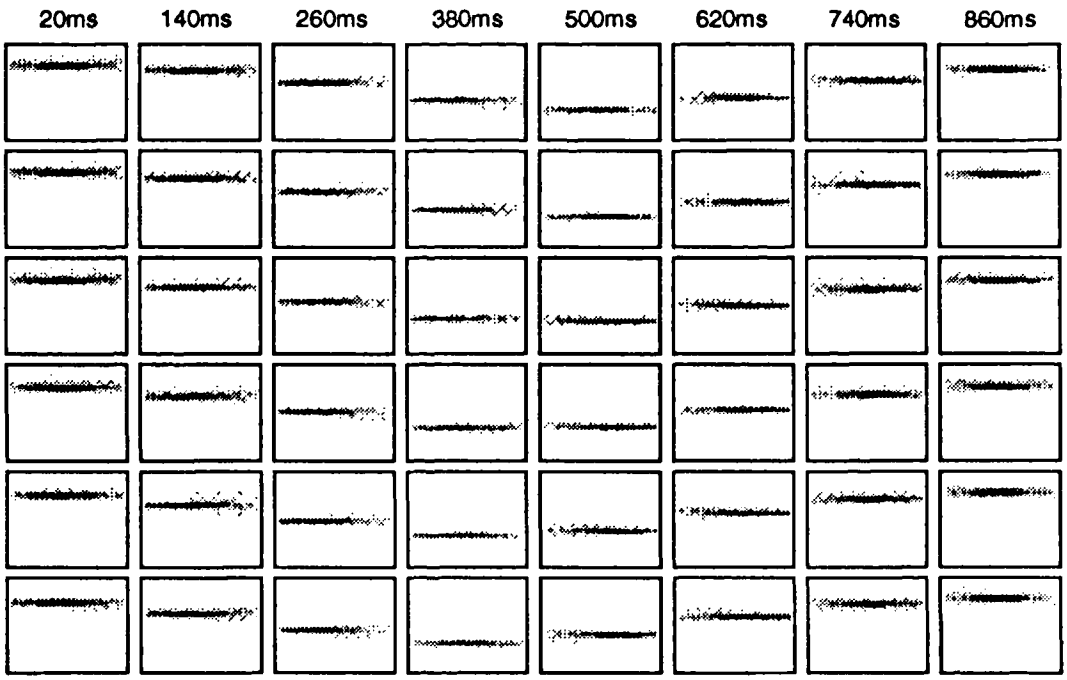

Figure 4.30: Frequency transition map representations of the siren noise source. In each plot, channel centre frequency (50Hz-5kHz) is represented on the ordinate and sweep rate (-20 oct/sec to 20 oct/sec) is represented on the abscissa.

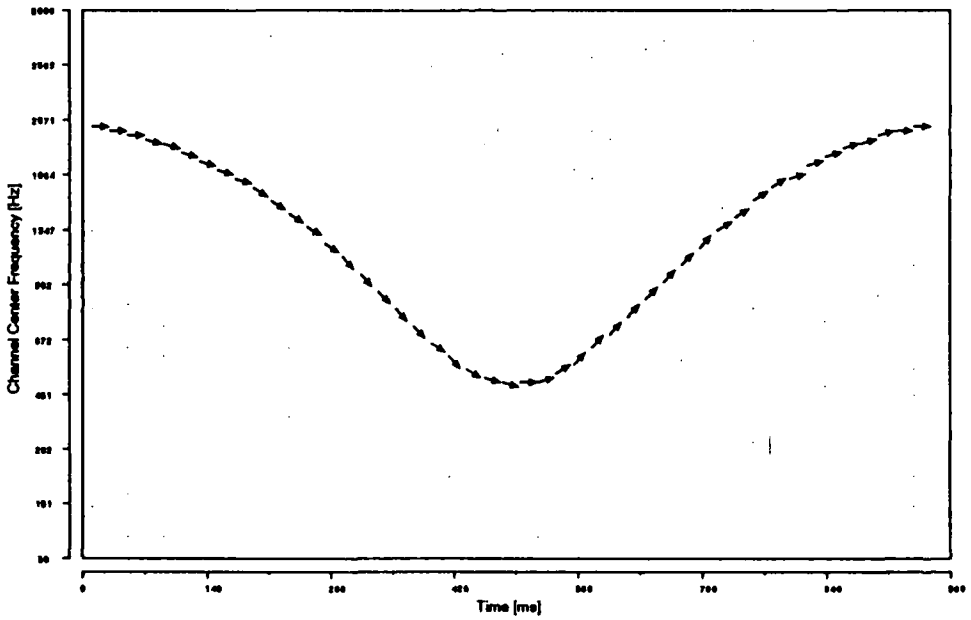

Figure 4.31: Location and orientation of dominances for the first cycle of the siren noise source. Each vector corresponds to the position of a maximum in the frequency transition map representations shown above.

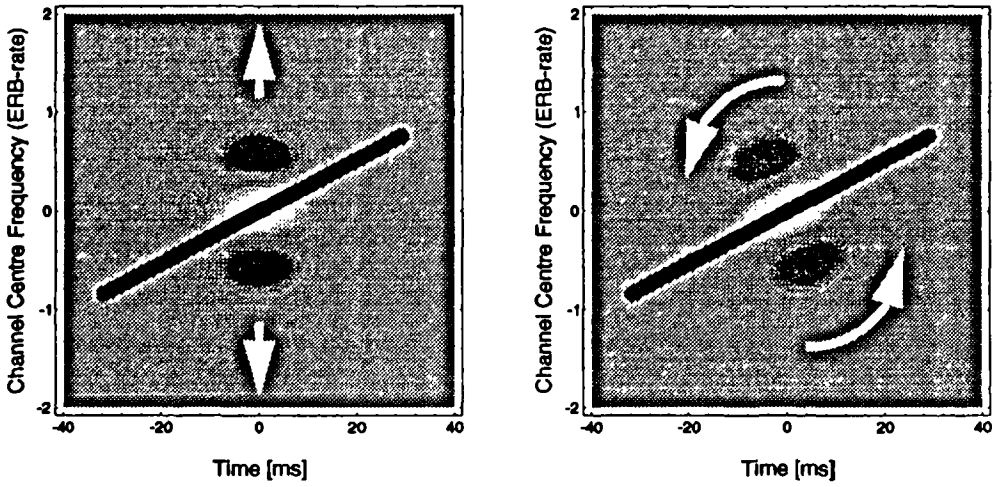

Figure 4.32: Spectral dominances are identified by locating maxima along the frequency axis of the map, equivalent to sliding the receptive field across frequency (left panel). The orientation of a dominance is found by locating the maximum along the frequency transition axis of the map, equivalent to rotating the receptive field in time-frequency (right panel).

Clearly, not all of the information in figure 4.30 is useful, since the map is intended to track *peaks* in the auditory nerve response, whereas the map measures the rate of frequency transition at every characteristic frequency. Dominances can be located in the map by looking for maxima along the frequency axis when $\theta = 0$. This operation can be visualized as sliding the receptive field $g[t, f]$ across frequency, in order to find the channels at which it responds optimally (left panel of figure 4.32). Formally, spectral peaks in the map occur at characteristic frequencies which satisfy the condition

$$\frac{\partial}{\partial f} s[t, f, 0] = 0 \tag{4.31}$$

This technique for identifying spectral peaks is generally very reliable, since $g[t, f]$ is sufficiently wide to ensure that a moving dominance generates activity in the map along the line where $\theta = 0$.

The direction in which a dominance is moving is determined by locating the maximum along the sweep rate axis at the characteristic frequency of the spectral peak. This operation can be visualized as rotating the receptive field $g[t, f]$ in time-frequency, centred on the spectral peak, in order to find the orientation at which it elicits an optimal response (right panel in figure 4.32). Formally, the rate of frequency transition at a particular characteristic frequency is given by

$$\frac{\partial}{\partial \theta} s[t, f, \theta] = 0 \tag{4.32}$$

Therefore, points in the map which satisfy conditions 4.31 and 4.32 define the position and sweep rate of a spectral dominance. These points can be identified by using a finite difference approximation to the differentials, checking the sign of zero crossings to ensure that a maximum has been found rather than a minimum. Figure 4.31 shows the position and orientation of dominances for the siren noise source. Spectral peaks are represented by vectors, which have a size related to the amplitude of the peak in the map, and a direction related to the rate of frequency transition. Clearly, the map forms an ideal basis for tracking spectral peaks across time.

Similar representations of the ten noise sources are shown in figure 4.33. The movement of dominances in the 1kHz tone, siren, telephone and music are clearly delineated, and the harmonics and formants of male and female speech are well represented. Conversely, noise sources do not generate any coherent activity in the frequency transition map.

4.4.5 Summary and Discussion

In this section, a physiologically and psychophysically-motivated model of a map of frequency transition has been presented. The map provides a solution to the auditory correspondence problem by coding information about the movement of spectral dominances across time and frequency.

The model assumes that the auditory system extrapolates trajectories in order to track spectral peaks across time. Although there is reasonable psychophysical evidence for this, it should be noted that trajectory effects are much weaker in audition than they are in vision. Bregman [24] suggests that this is because visual objects tend to have a momentum associated with them, whereas auditory objects do not. For example, there is no inertia in the vocal tract which ensures that a falling transition in frequency will continue to fall. However, the absence of inertia for frequency transition does not rule out the possibility that a primitive auditory process measures and extrapolates time-frequency trajectories.

A related point concerns the analogy, discussed in section 4.4.2, between the frequency transition map and edge-orientation columns in the visual cortex. Kay and Matthews [128] suggest that some caution is needed when comparing auditory and visual organization. In the visual system, time and space constitute two separate characteristics of a stimulus arriving at the retina. However, frequency and time are intrinsically linked in acoustic stimuli, and there is a relationship along the cochlear partition between space and frequency (see section 2.1.2). Hence, in the visual pathway temporal and spatial characteristics of stimuli may be completely separable, whereas in the auditory pathway they may not. However, despite this complication it is still instructive to compare similarities in the organization of the auditory and visual systems.

It has also been assumed in the model that the input to the frequency transition map is auditory nerve firing rate. A problem with this approach is that separate spectral

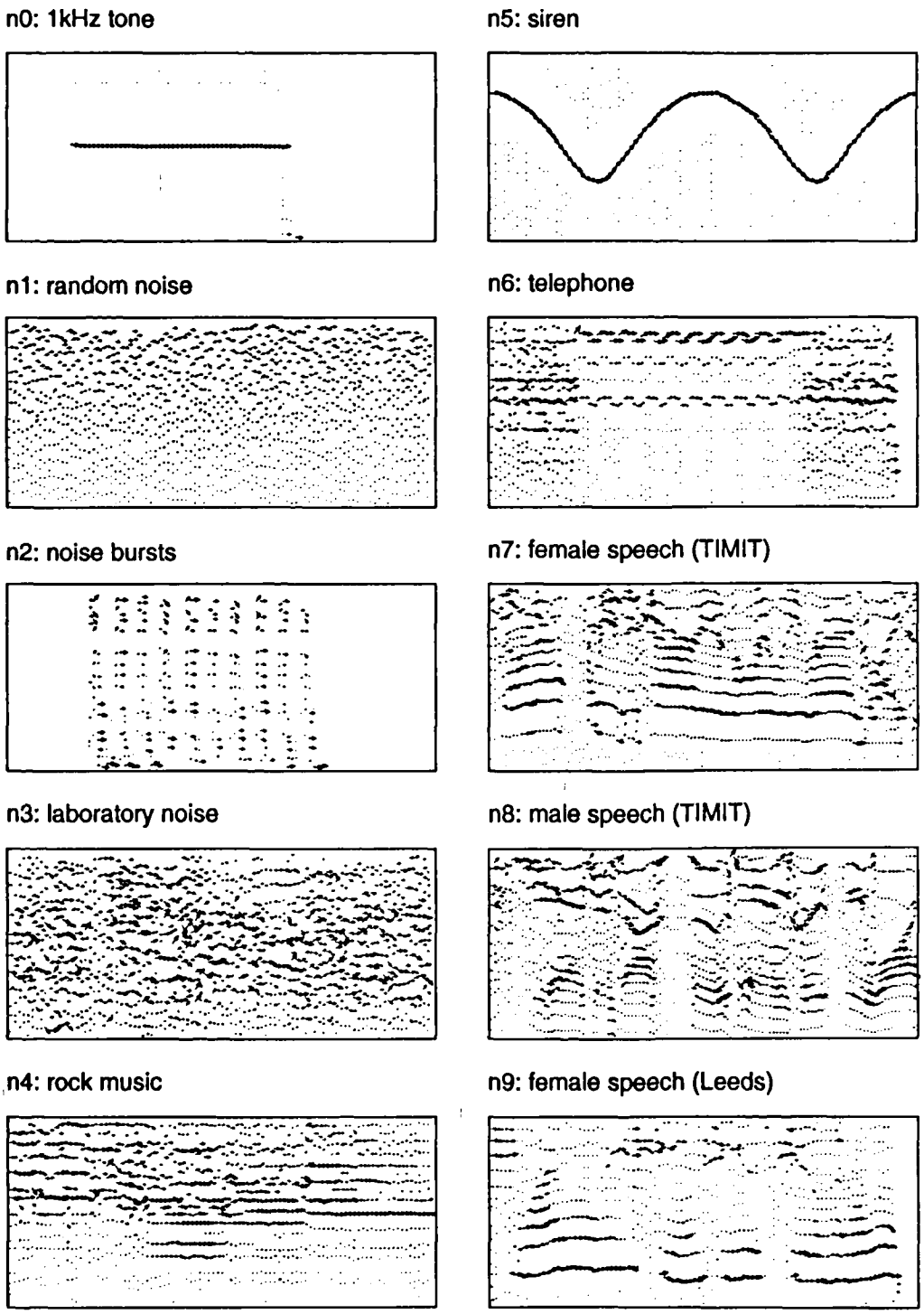

Figure 4.33: Frequency transition map representations of the ten noise sources. Vector size is related to amplitude, direction is related to sweep rate. Time is represented on the abscissa, channel centre frequency (50Hz-5kHz) is represented on the ordinate.

peaks are not resolved in the auditory nerve at medium to high stimulus intensities. Sachs and Young [231, 232] have studied this effect using synthetic vowels, and have suggested two contributing factors. Firstly, auditory nerve fibres saturate at high intensities, so that small variations in amplitude are not accompanied by changes in firing rate. Secondly, high intensity components are able to suppress the auditory nerve response to other components, reducing the contrast in the spectral pattern (see section 2.1.3).

Since the gammatone filterbank does not model nonlinear suppression effects, only the problem of firing rate saturation is considered here (see section 4.1.5). One solution is to assume that the frequency transition map receives its input from auditory nerve fibres which have a low spontaneous firing rate (high threshold). Sachs and Young [231] found that these fibres, which account for approximately 16% of the auditory nerve population, maintain peaks in the rate spectrum at high stimulus intensities when lower threshold fibres become saturated. Another solution would be to use a representation of temporal information in the auditory nerve as input to the map, since this does not degrade at high intensities (Young and Sachs [288]). For example, the synchrony spectrum described by Seneff [243] would provide a suitable input.

4.5 Other Grouping Primitives

This section reviews three other potential grouping primitives, *common amplitude modulation*, *common frequency modulation* and *common spatial location*. These cues are examples of the Gestalt principle of "common fate", which was discussed in section 3.1. They are not included in the computational model either because the experimental findings are insubstantial or controversial, or because their implementation is beyond the scope of the current work.

4.5.1 Common Amplitude Modulation

The components of a single sound source tend to vary in amplitude in a coherent manner, and it is likely that the auditory system exploits this fact by grouping spectral regions which have a common pattern of amplitude modulation (AM). In principle, grouping by common AM is similar to the onset and offset grouping effects described in section 4.3, except that it involves a smaller scale of amplitude change. Also, note that common AM is distinguished from common periodicity by the fact that the amplitude modulation need not be periodic. Rather, it appears that grouping by common AM involves the correlation of instantaneous changes in amplitude, rather than a comparison of repetition rate.

Bregman *et al.* [26] have investigated the effects of common AM using a stimulus similar to the one shown in figure 4.11, in which an amplitude modulated tone A was alternated with a pair of amplitude modulated tones B and C. When B and C

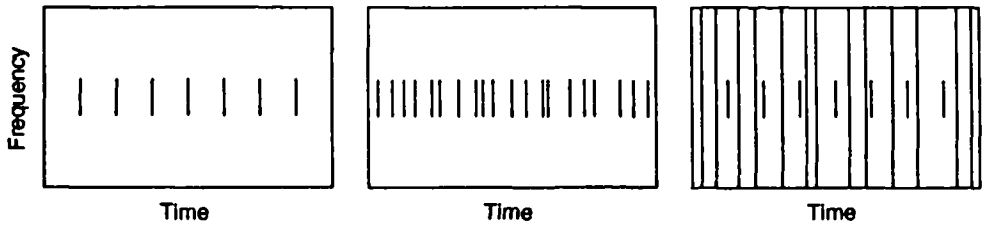

Figure 4.34: Visual analogy of co-modulation masking release (CMR). A tone (left panel) is masked with a narrow band of amplitude modulated noise (centre panel). When co-modulated noise is added at flanking frequencies, the detectability of the tone is improved (right panel). After Bregman [25].

were modulated at the same rate, they tended to fuse into a single perceptual group. Although this effect was weak, Bregman *et al.* [29] repeated the experiment using a larger range of AM mistuning between B and C, and found a correspondingly larger effect. Hence, it appears that common AM can promote the perceptual grouping of spectral components.

Two observations suggest that this result cannot be explained by a pitch analysis mechanism, such as the autocorrelation map described in section 4.2.3. Firstly, B and C tended to fuse if they had common AM regardless of whether their frequency components were aligned in a harmonic series. Secondly, if B and C were modulated at the same rate but with a different phase, their tendency to form a group was reduced. However, mechanisms which group components for the purpose of deriving a pitch are known to be unaffected by phase. This is evident in the autocorrelation map, where across-channel phase differences are corrected by the autocorrelation analysis. Therefore, Bregman's experiment probably illustrates a process which correlates instantaneous changes in amplitude at different characteristic frequencies. Further experiments are required to confirm this hypothesis.

Hall *et al.* [108] have also described an effect which could be due to grouping by common AM. They reduced the detectability of a tone by *masking* it with a narrow band of noise that was amplitude modulated at a slow, irregular rate. The masking effect was maximal when the noise had the same bandwidth as the auditory filter centred on the frequency of the tone. Further increases in the bandwidth of the noise reduced the masking effect, so that the tone became more detectable. This effect only occurred when the noise was amplitude modulated, suggesting that listeners were able to compare the pattern of AM in different auditory filters in order to improve the detectability of the tone. Hence, Hall *et al.* named the phenomenon *co-modulation masking release (CMR).* Additionally, they noted its potential importance as a mechanism of perceptual grouping:

> "Many real-life auditory stimuli have intensity peaks and valleys as a function of time in which intensity trajectories are highly correlated across frequency. This is true of speech, of interfering noise such as

> 'cafeteria' noise, and of many other kinds of environmental stimuli. We suggest that for such stimuli the auditory system uses across-frequency analysis of temporal modulation patterns to help register and differentiate between acoustic sources."(page 56)

A visual analogy of Hall's experiment is shown in figure 4.34. CMR effects have also been demonstrated by Schooneveldt and Moore [236] using separate bands of noise, one centred on the tone and another at a distant (flanking) frequency. When the noise bands were coherently amplitude modulated, a CMR was obtained.

Two explanations for these findings have been suggested by Buus [38]. Firstly, the auditory system may correlate AM across different characteristic frequencies, so that the auditory filter centred on the tone can be identified by its different modulation pattern. This mechanism is compatible with a scene analysis interpretation, in which channels with common AM fuse into a perceptual group and channels with different AM are segregated. Alternatively, Buus suggests that the noise at flanking frequencies might indicate when the amplitude of the masker is at a minimum. Given this information, the auditory system could "listen in the dips" to improve the detectability of the tone. There is some support for both of these explanations, but neither can account for all of the experimental findings (see Moore [180] for a review). For example, Hall and Grose [106] have shown that a CMR can be obtained in situations where across-frequency modulation disparities and dip-listening cues are absent. This result and others (Hall *et al.* [107], Grose and Hall [104]) suggest that CMR occurs as a result of flexible mechanisms which exploit many different cues.

An effect which appears to act in the opposite direction to CMR has been identified by Yost and Sheft [285] and Moore *et al.* [186], in which the outputs of auditory filters tuned away from the frequency of a signal *degrade* signal detection. Yost and Sheft found that the threshold for detecting sinusoidal AM of a component was increased in the presence of another component sinusoidally-amplitude modulated at the same rate, even when the second component was remote in frequency from the first. They called this phenomenon *modulation detection interference (MDI).* Moore *et al.* describe a similar effect. They masked a tone with a sinusoidally-amplitude modulated complex, and added another sinusoidal AM complex at a flanking frequency. When the masker and flanking components had the same modulation rate, the detectability of the tone was impaired. Since this phenomenon could not be explained in terms of a single auditory filter, Moore *et al.* called it *across-channel masking (ACM).*

Yost and Sheft [286] interpret MDI/ACM in terms of channels which are tuned to a particular rate of AM (see section 4.4.1). Spectral components that have a similar rate of modulation are assumed to excite the same AM channel, which promotes their perceptual fusion. However, this interpretation is controversial. Moore *et al.* [184] have found that ACM is broadly tuned for AM rate, a result which is not compatible with a grouping explanation. If components with substantially different AM rates were allowed to fuse, many errors in grouping would be made. Also, it was

found that a flanking frequency modulated component could degrade the detection of a signal masked by an AM complex, and vice versa. Hence, it seems unlikely that channels specific for AM can account for the MDI/ACM phenomenon.

In conclusion, there is reasonable evidence that common AM promotes the perceptual grouping of spectral components. However, the underlying mechanisms are not well understood, so no modelling work was attempted.

4.5.2 Common Frequency Modulation

When the fundamental frequency of a complex sound is modulated, all of its components change frequency in the same direction at the same time. The auditory system might exploit this regularity by grouping components which have a similar pattern of frequency modulation (FM), since they are likely to belong to the same acoustic source. However, evidence for a grouping mechanism based on common FM is weak.

It should be noted that the principles of grouping by common FM and grouping by harmonicity are closely related. Although the harmonics of a modulated fundamental exhibit common FM, they also maintain their harmonicity over time. Therefore, a demonstration of grouping by common FM must show that it is the *motion* of frequency components which promotes their perceptual fusion, rather than the fact that they are harmonically related.

McAdams [162] has examined the effects of common FM using two different types of fundamental frequency modulation, *vibrato* (periodic) and *jitter* (aperiodic). Listeners were presented with a mixture of three vowels synthesized on different fundamentals, and were asked to judge the prominence of each vowel. When vibrato and jitter were imposed on the fundamental of a vowel, it became more prominent. However, the relation between the modulation applied to one vowel and the modulation applied to the other two vowels did not influence prominence. Therefore, common FM did not help to segregate the components of the different vowels from the mixture.

A similar result has been reported by Gardner and Darwin [89], using the /I/-/e/ phoneme boundary paradigm described in section 4.3.1. Recall that this technique quantifies the contribution of a harmonic to a vowel by measuring changes in the vowel's phonetic quality. When one harmonic of the vowel was frequency modulated at a different rate or phase from the others, it still made a full contribution to the vowel percept. Hence, common FM does not appear to affect the grouping of frequency components into a phonetic category. Further evidence supporting this conclusion has come from an experiment which used the synthetic "ru-li" stimulus described in sections 4.2.1 and 4.3.1 (Gardner *et al.* [90]). When the second formant of the syllable was frequency modulated at a different rate from the others, it still contributed fully to the syllable percept. However, subjects reported hearing the second formant as a separate sound source.

In conclusion, these results suggest that common FM has no influence on the perceptual grouping of frequency components. It may, however, influence the number of sound sources that are heard.

Two reasons for the absence of an effect of common FM in perceptual grouping have been proposed. Firstly, Summerfield [267] suggests that harmonicity may be such a powerful grouping cue that it achieves all the segregation that can occur, leaving nothing for common FM to contribute. To test this hypothesis, he investigated the detectability of a "target" vowel in the presence of a "masker" vowel, under various conditions of FM. The vowels were synthesized with their components randomly displaced from their harmonic frequencies, so that harmonicity cues were reduced. However, when there was a difference in FM between the two vowels, no increase in the detectability of the target vowel was obtained. Therefore, common FM was not used to group the components of a vowel in a situation where harmonicity cues were sub-optimal.

A second explanation for the absence of a common FM effect, which is supported by Summerfield's findings, is that listeners are simply unable to detect across-frequency differences in FM. Carlyon [41] has recently addressed this point. He presented subjects with a pair of complex tones which were frequency modulated in a coherent or incoherent manner. When the tones were inharmonic, listeners were unable to distinguish coherent FM from incoherent FM. Carlyon concluded that there is no across-frequency mechanism specific for the detection of common FM. Note that this interpretation is consistent with the MDI/ACM effect described in the previous section, in which the presence of modulation in one frequency region impairs the detection of modulation in another frequency region.

4.5.3 Spatial Location

One of the strongest scene analysis principles is common spatial location. Acoustic components that originate from the same location in space tend to be assigned to the same stream, whereas those arising from different locations tend to be assigned to different streams.

There is good evidence that masking is reduced when a signal and a masker are perceived to be at different spatial locations (see Moore [179] for a review). This spatial release from masking is illustrated by the phenomenon of *binaural masking level difference (BMLD)*. For example, consider a situation in which a mixture of a tone and white noise is presented to both ears of a listener. The level of the tone is adjusted until it is just masked by the noise. Let the level at this point be L_1. Now, if the phase of the tone is shifted by 180 degrees, it becomes audible again. As before, the tone can be adjusted to a new level L_2, so that it is just masked by the noise. The difference between the two levels $L_2 - L_1$ is known as the binaural masking level difference. For low frequency signals (around 500 Hz), the BMLD may be as large as 15 dB. A BMLD may also be obtained when there is a difference in intensity between the two ears. For example, consider a case in which the tone

and noise are presented to one ear only, and the level of the tone is adjusted so that it is just masked. When the noise alone is now added to the other ear, the tone becomes audible again.

The BMLD phenomenon has been observed for complex tones, speech sounds and clicks as well as for pure tones. In general, it appears that the ability of a listener to detect a masked signal is improved when the signal and masker have a different phase or intensity at the two ears, relative to the case where they have the same phase and intensity. Such differences only occur in natural acoustic environments when the signal and masker have a different spatial location. Hence, an implication of the BMLD effect is that a signal will be easier to detect when it is located in a different position in space from a masking sound.

It should be noted that a BMLD can be obtained in situations where the signal and masker are not subjectively well separated in space (Carhart *et al.* [39]). Additionally, the largest BMLDs are obtained for phase differences which are greater than those that occur in natural listening situations. In fact, the only necessary condition for obtaining a BMLD is that the signal or masker should be out of phase at the two ears. Hence, in scene analysis terms, the BMLD phenomenon can be explained by a mechanism that groups components which have a common interaural phase difference, rather than a mechanism that groups components by common spatial location *per se* (Bregman [24]).

Binaural grouping processes have also been illustrated by Kubovy *et al.* [141], who presented subjects with a mixture of eight continuous tones at both ears. The frequencies of the tones corresponded to the notes in a musical scale, and initially the tones had the same phase in each ear. However, when the phase of a tone in one ear was shifted relative to its phase in the other ear, the tone stood out perceptually from the other components and was heard to originate from a different spatial location. By shifting the phase of a sequence of tones, the experience of a melody could be created. Listeners were not able to hear the melody if the stimulus was presented monaurally. This result suggests that a difference in the relative phase of a tone at the two ears allowed it to be perceptually segregated from the other components in the mixture.

No attempt to simulate binaural auditory processing has been made in the model presented here, and hence the scene analysis algorithm described in the following chapter does not use information about spatial location. However, note that interaural time and intensity differences appear to be represented in auditory maps (see section 2.2.3). Models of these maps could be incorporated into the system at a later stage.

Chapter 5

A Strategy For Auditory Scene Analysis

In this chapter, primitives from the auditory maps are used to construct a representation of auditory objects. Subsequently, a search strategy is described which groups objects that have similar properties.

The formation of auditory objects from periodicity and frequency transition primitives is described in section 5.1. Techniques for grouping objects by common periodicity and common onset/offset are proposed in sections 5.2 and 5.3. In section 5.4, a strategy for searching the auditory scene is described. Finally, section 5.5 discusses some theoretical and computational issues.

5.1 A Representation of Auditory Objects

5.1.1 Motivation

In the previous chapter, evidence was presented that perceptual grouping mechanisms use properties of common onset, offset and periodicity to group acoustic components together. But properties of *what*? So far, the auditory representation of an acoustic source has been considered only in terms of the activity in separate neural maps over time. Clearly, a representation of auditory events is required which combines the information from the different maps, and is amenable to the application of grouping principles in a scene analysis strategy.

An important issue to be considered here is the representation of time. The majority of auditory models that have been described in the literature employ a *frame-based* representation of time, such as the one shown in figure 4.1. In a frame-based scheme, the activity at each characteristic frequency is coded as a one-dimensional vector of

parameters at regular time intervals. Generally, this representation is used because the output of the auditory model is required in a form that is compatible with frame-based automatic speech recognition systems (Beet [11], Ghitza [93], Hunt and Lefebvre [123]). Similarly, frame-based representations of time have been employed in the majority of systems which attempt to segregate simultaneous sounds (Parsons [198], Scheffers [234], Stubbs and Summerfield [260], Varga and Moore [272]).

Although the frame-based representation in figure 4.1 provides a good *visual* description of acoustic events, it is inadequate as a basis for auditory scene analysis. Specifically, it does not contain any information about the way in which acoustic components vary across time. The importance of temporal continuity has been noted by a number of workers, notably Darwin and Gardner [64], Riley [220], McAulay and Quatieri [163], Heinbach [112] and Cooke [52]. For example, Riley observes that

> "When we look at a spectrogram, we are not confined to examining them one-dimensionally along single frequency slices, but instead we see a two-dimensional time and frequency surface. In other words, time is not used as a parameter that varies over a family of spectra, but as one of the intrinsic dimensions of the representation."(page 18)

Consequently, the approach described here employs an auditory representation in which time is made explicit. The auditory scene is characterized as a collection of *auditory objects*, which describe the movement of spectral components in time and frequency. The work of Green *et al.* [102], Riley [220] and Cooke [52] suggests that a representation of this type is very expressive in computational terms. For example, Riley considers a situation in which two time-frequency contours (such as speech formants) are competing for the same label. In a frame-based representation where the contours are sampled over n frames, there are 2^n ways of labelling the points along each component. However, if each contour is represented by a single temporally-extended object, there are only two possible labellings.

The following section describes a strategy for the formation of auditory objects, which uses primitives supplied by the frequency transition and periodicity maps. In section 5.1.3, auditory object representations of the ten noise sources are discussed.

5.1.2 Formation of Auditory Objects

It was noted in section 4.4.1 that a correspondence problem must be solved in order to make temporal relations explicit. Specifically, it is necessary to match the auditory representation of an acoustic event at a particular time with the representation of the same event at a later time. The solution to this problem described in section 4.4.1 was a *frequency transition map*, which extracts information about the movement of spectral peaks over time. Additionally, it was noted in section 4.2.6 that spectral peaks tend to recruit the response of a contiguous section of the auditory filterbank. A cross-correlation map was proposed, from which *periodicity groups* are

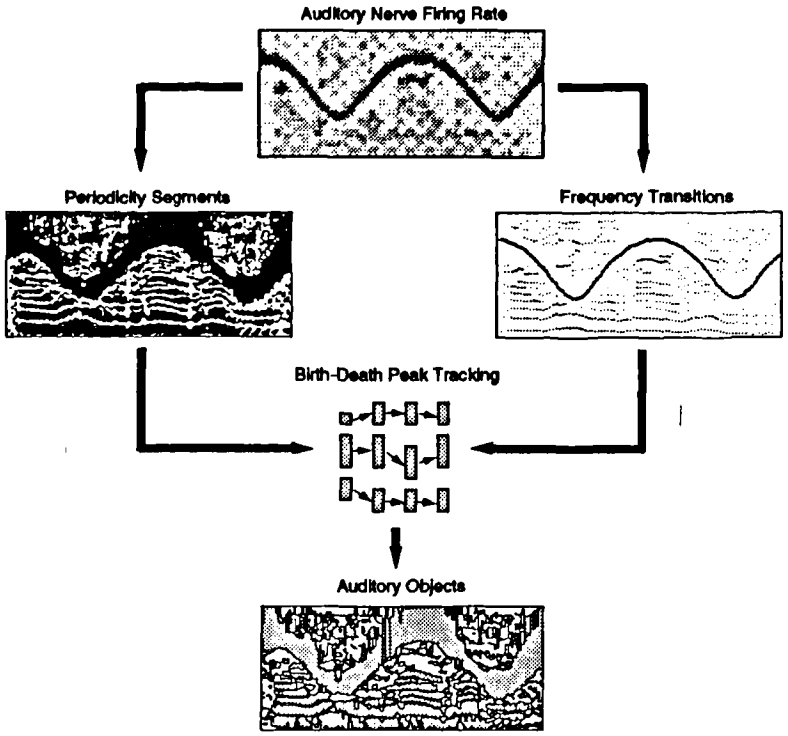

Figure 5.1: Formation of auditory objects. Spectral peaks are tracked across time using information about frequency transition and periodicity groups.

selected by an area stability criterion. The periodicity groups delineate areas of the auditory filterbank which are responding to the same spectral dominance.

Given the frequency transition and periodicity group primitives, auditory object formation proceeds as shown in figure 5.1. The locations and directions of movement of spectral peaks are derived from the frequency transition map at each time instant, and the frequency spread of each peak is determined by matching it with a periodicity group. Spectral dominances are tracked across time by a *birth-death* strategy, which predicts the movement of peaks from one frame to the next. The resulting auditory objects are shown at the bottom of the figure. Here, each grey shape is a single object which defines the path of a spectral dominance across time and frequency. Note that the width of an object corresponds to the number of auditory filter channels that it dominates.

Birth-Death Peak Tracking

Spectral peaks are tracked across time by a birth-death strategy, similar to the procedures described by McAulay and Quatieri [163] and Cooke [52]. Four iterations of the process are shown in figure 5.2.

Initially, the location and orientation of each spectral peak in a particular time slice

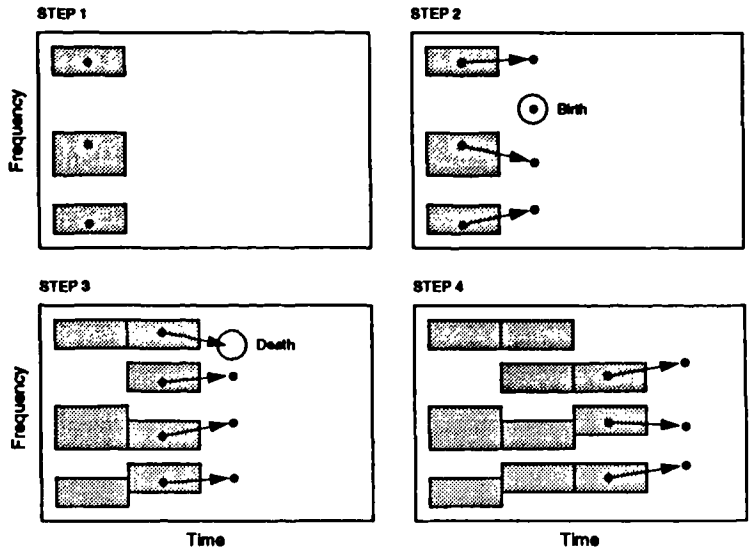

Figure 5.2: Formation of auditory objects by a birth-death tracking procedure. Spectral peaks (black dots) which lie within the predicted acceptance region of an auditory object are recruited together with their corresponding periodicity groups (grey rectangles). Peaks that are not recruited are "born" as new objects (step 2), and objects "die" if they are unable to recruit new peaks (step 3).

are derived by finding the maxima in the frequency transition map, as described in section 4.4.4. Subsequently, the movement of a peak at time t to a new frequency channel f at time $t+1$ is predicted by a simple linear extrapolation of the peak's orientation. In practice, it is desirable to allow some tolerance in the predicted position of the peak, so an *acceptance region* $\omega[f]$ is computed which is centred on f. Formation of an auditory object then proceeds according to the following three rules:

Rule 1: For an existing object at time t, a peak that lies within the acceptance region $\omega[f]$ at time $t+1$ is recruited to the object. If the recruited peak falls within the boundaries of a periodicity group, then the frequency spread of the object at time $t+1$ is taken as the width of the periodicity group. Otherwise, the object is assumed to occupy one channel of the filterbank at time $t+1$ (steps 2,3 and 4 in the figure).

Rule 2: If an existing object at time t is unable to recruit a new peak at time $t+1$, the object "dies" (step 3 in the figure).

Rule 3: Peaks at time $t+1$ which do not fall within the acceptance region of an existing object are "born" as new objects. Periodicity groups are matched to the new object as described in the first rule (step 2 of the figure).

The use of an acceptance region around the predicted location of a peak is consistent with the findings of Ciocca and Bregman [49], which were discussed in section 4.4.1. They found that when asked to judge the continuity of a glide through a band of

noise, listeners tolerated a disparity in the starting frequency of the post-noise glide (see figure 4.21). Unfortunately, Ciocca and Bregman did not quantify the width of the acceptance region for different glide slopes, so their data cannot be used to calibrate the model here. Instead, the width of $\omega[f]$ was derived empirically. A tolerance of one channel either side of the predicted peak position (corresponding to a $\omega[f]$ of 0.6 ERB) was found to be suitable. Wider acceptance regions tended to produce longer objects, but increased the number of tracking errors.

It should be noted that two types of conflict can occur during the birth-death tracking process. Firstly, two objects may compete for the same spectral peak. In this case, the longest object recruits the peak, and the other "dies". Secondly, two spectral peaks may occur within the same periodicity group. This condition occurs infrequently, and is generally due to a failure of the cross-correlation map to identify the boundary between two spectral dominances. When this conflict arises, the same periodicity group is matched with *both* peaks.

Not all of the auditory objects are retained for further processing. Specifically, objects that span fewer than two time frames are eliminated. Although this is a violation of Marr's "principle of least commitment" (do nothing which may have to be undone later), it is desirable to "clean up" the representation at this stage. Very short auditory objects are unlikely to have a significant acoustic correlate, and removing them eases the computational load on the subsequent scene analysis strategy.

5.1.3 Auditory Object Representations

Auditory object representations of the ten noise sources are shown in figure 5.3. It is instructive to compare this figure with the periodicity group and frequency transition map representations in figures 4.10 and 4.33. Note that individual harmonics and formants of speech are generally represented as a single object. Noise sources give rise to many small, randomly distributed objects, although some structure is visible in the laboratory noise source.

5.2 Grouping by Common Periodicity

In the model described here, periodicities in the firing patterns of auditory nerve fibres are extracted by an *autocorrelation map* (see section 4.2.3). This section proposes a strategy for segregating concurrent periodic sounds, which partitions the channels of the autocorrelation map into groups that are likely to have the same fundamental frequency. The basic principles of this strategy are illustrated first, using a simple double vowel stimulus. Subsequently, the application of the strategy to the grouping of auditory objects is described. Before proceeding, it is instructive to consider some previous autocorrelation-based approaches to source segregation.

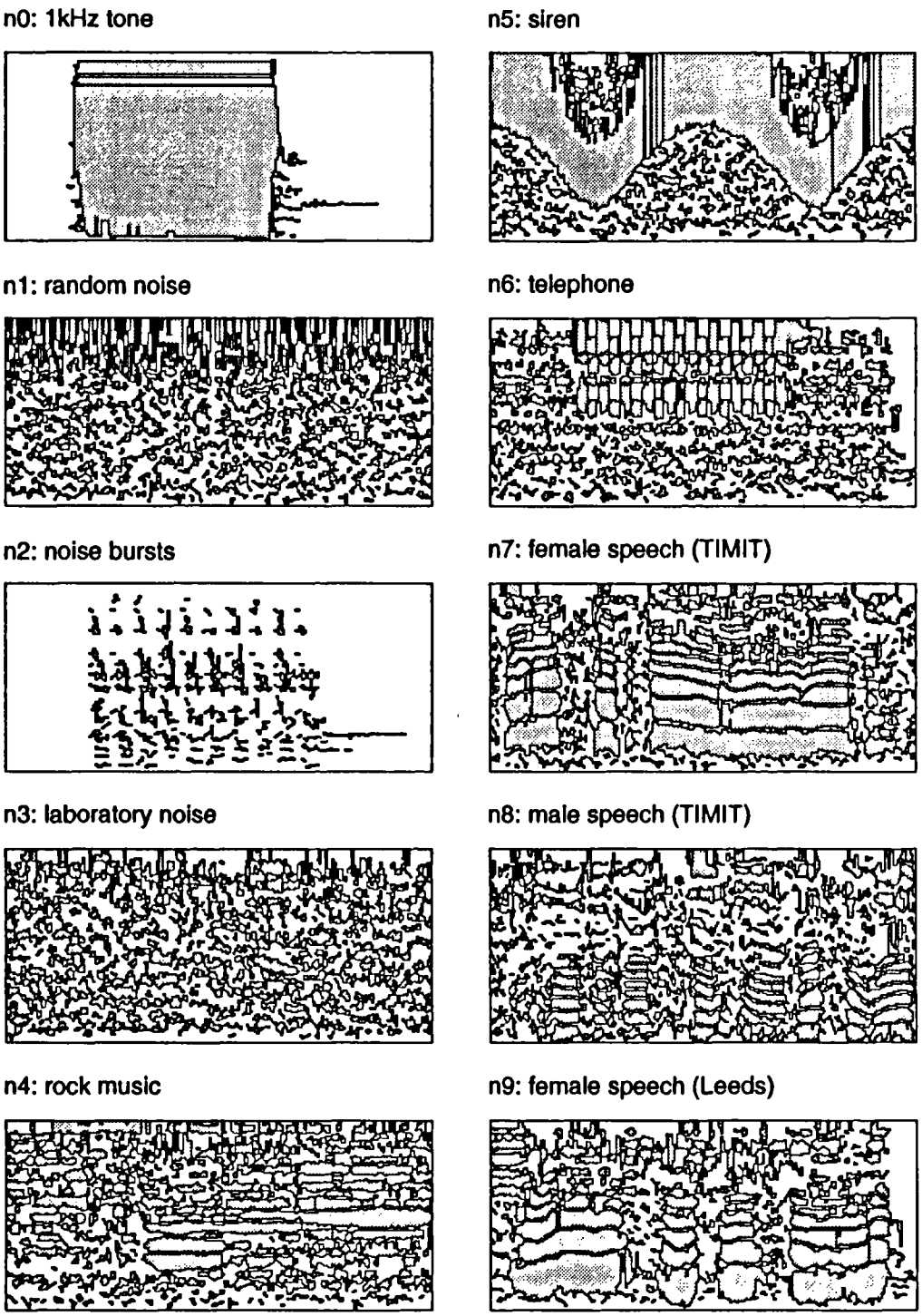

Figure 5.3: Auditory object representations of the ten noise sources. Time is displayed on the abscissa, and channel centre frequency (50Hz-5kHz) is displayed on the ordinate.

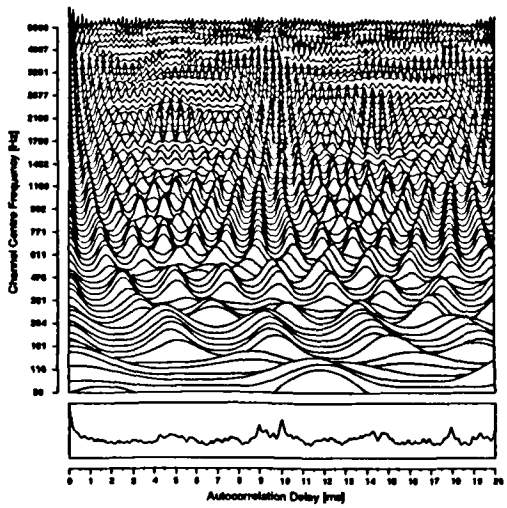

Figure 5.4: Autocorrelation map for the double vowel /a/ (fundamental frequency 112 Hz) and /e/ (fundamental frequency 100 Hz). Note that a peak at the pitch period of each vowel occurs in the summary autocorrelation function.

5.2.1 Previous Work

Two autocorrelation-based segregation strategies are reviewed here, which have been proposed by Assmann and Summerfield [8] and Meddis and Hewitt [169]. Both attempt to model the perceptual processes underlying the ability of human listeners to identify concurrent vowels with different fundamental frequencies (see sections 4.2.1 and 1.4.3). As such, they are limited to processing static sounds, and operate on a single frame of an autocorrelation map. However, the strategies could equally be applied to successive frames of a map in order to segregate time-varying stimuli. Weintraub [277] describes a separation system based on this principle, which is discussed in chapter 7.

Assmann and Summerfield (A&S) propose several schemes for segregating double vowels. Their "nonlinear place-time" model is considered here, which employs an autocorrelation map of the form described in section 4.2.3. Initially, a summary autocorrelation function is formed from the map (see section 4.2.4), and the two largest peaks in the summary are identified. The delays at which these peaks occur are assumed to correspond to the pitch periods of the two vowels. Subsequently, the spectrum of each vowel is estimated by sampling the channels of the autocorrelation map at the delay corresponding to the vowel's pitch period. Hence, two "synchrony spectra" are obtained, which indicate the degree of synchronization to each vowel in the auditory nerve. By matching these spectra against reference templates, vowel identification performance can be quantified. In fact, the A&S model comes close to predicting the overall accuracy of listeners responses. However, it is unable to replicate the finding of Scheffers [233] that identification performance improves with larger differences in fundamental frequency between the two vowels.

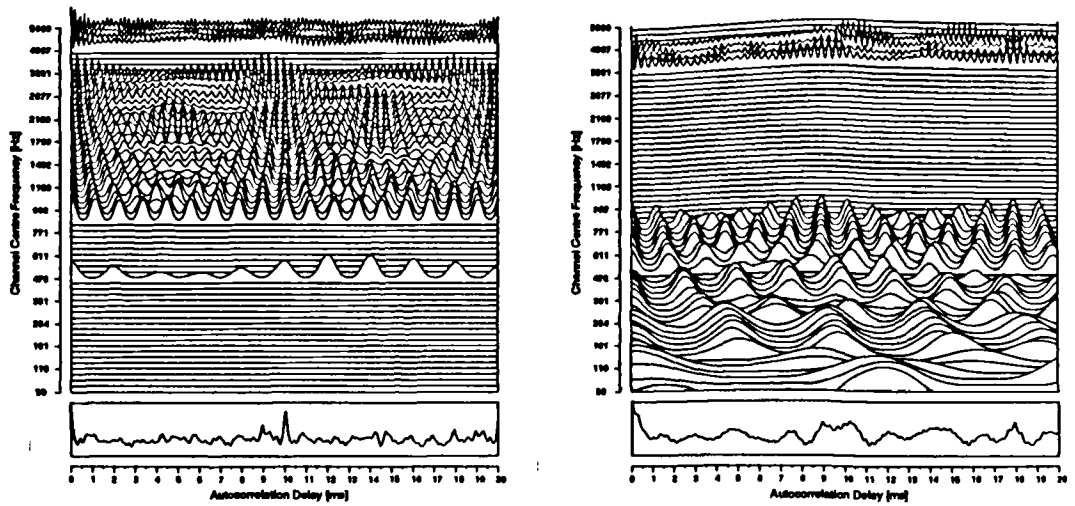

Figure 5.5: Segregation of the double vowel by the Meddis and Hewitt strategy. The group of channels on the left corresponds to the /e/, and the group on the right corresponds to the /a/.

A more successful strategy, in terms of predicting human performance, has been proposed by Meddis and Hewitt (M&H). Given that there are two vowels present with different fundamental frequencies, the M&H scheme partitions the autocorrelation map into two mutually exclusive sets of channels. Initially, the largest peak in the summary autocorrelation is identified, and this is taken to be the pitch period of the dominant vowel. Channels with a peak in their autocorrelation functions at this delay are removed from the map, and matched with a template. The remaining channels are assumed to belong to the second vowel, and are matched with a template in a similar manner. Hence, only the pitch of the most dominant vowel is estimated. This is advantageous, since the second pitch is often weak, and may be an unreliable cue for segregation (Meddis and Hewitt [169]). The M&H scheme is able to model Scheffer's findings quite closely, and shows an improvement in performance when the difference in fundamental frequency between the two vowels is increased.

The M&H strategy is illustrated in figure 5.5, for the mixture of two vowels /e/ and /a/ shown in figure 5.4. The vowels have fundamental frequencies of 100 Hz and 112 Hz respectively, corresponding to a difference in fundamental of two semitones. In the mixture, the largest peak in the summary autocorrelation occurs at a delay of 10.0 ms, which is the pitch period of the /e/. The channels of the map with a peak at this delay are shown in the left panel of figure 5.5. Note that the peak at 10.0 ms in the summary autocorrelation of this map is now more clearly defined, indicating that some segregation has been achieved. The remaining channels are assumed to belong to the /a/, as shown in the right panel of figure 5.5. Here, the peak in the summary autocorrelation at a delay of 8.93 ms, corresponding to the pitch period of the /a/, has become relatively larger. However, there is still a significant peak at

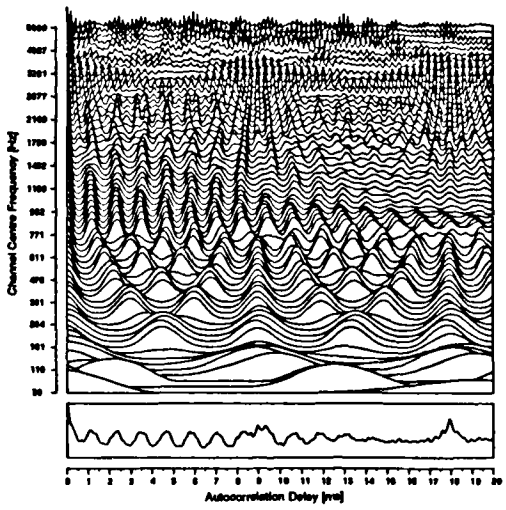

Figure 5.6: Autocorrelation map of the vowel /a/ (fundamental 112 Hz) and siren noise source. Note that the summary autocorrelation contains many peaks which might be incorrectly interpreted as the pitch period of a source.

10.0 ms delay, suggesting that the two vowels have not been completely segregated.

Although the A&S and M&H strategies are able to model the identification of double vowels quite closely, they suffer from a number of disadvantages. In particular, the schemes do not generalize in situations where several arbitrary sound sources are active at the same time. For example, both the A&S and M&H strategies require *a priori* knowledge of the number of sound sources that are present. Consider the A&S scheme, which attempts to find the pitch period of each source by identifying peaks in the summary autocorrelation function. For stimuli other than synthetic double vowels, this is a non-trivial problem. The point is illustrated in figure 5.6, which shows an autocorrelation map of the vowel /a/ mixed with the siren noise source. As before, the fundamental frequency of the vowel is 112 Hz. Although there are only two sources present, there is a multitude of peaks in the summary autocorrelation. Clearly, the A&S strategy would have great difficulty in identifying the number of sources present and assigning a pitch period to each one. The M&H scheme overcomes the problem of multiple peaks by identifying the pitch period of the dominant source, and partitioning the channels of the map into two mutually-exclusive sets. But what if there are more than two sound sources present? The M&H strategy does not generalize in this case.

A related criticism of the M&H scheme is that listeners can often hear both pitches in a double vowel, and are able to indicate which vowel has the higher pitch and which has the lower pitch (Summerfield *et al.* [266]). Similarly, Beerends and Houtsma [10] have found that listeners are often able to correctly identify the pitches of concurrent two-tone complexes, for differences in fundamental frequency of two semitones or more. It seems unlikely, therefore, that segregation is based only on the most dom-

inant pitch. Note also that the M&H scheme is quite sensitive to small variations in the channel autocorrelation functions of the map. This is apparent in figure 5.5. Several low-frequency channels which should belong to the /e/ have been incorrectly assigned to the /a/, because their peaks near to 10.0 ms are slightly displaced. A possible solution to this problem is to allow a tolerance in the position of the channel pitch period peak. However, Meddis and Hewitt [169] found that allowing a 1% tolerance in peak position did not improve the performance of their model. Another solution might be to allow a wider tolerance at lower characteristic frequencies, since there tends to be more variation in the broader peaks of low frequency channels.

Another point concerns the relationship between the pitch system and perceptual grouping mechanisms. The A&S and M&H segregation strategies assume that the pitch of a source is identified first, and then this pitch is used to group the components of the source together. However, Bregman [24] notes that this is unlikely to be the case:

> "The pitch system acts to group harmonically related partials. We might conclude that this grouping is then used to derive other properties of the now segregated partials. This description implies a one-way transaction, the pitch system influencing the grouping system and not vice versa. However, this appears not to be true. There is evidence that the pitch that is calculated can depend on cues other than harmonicity, cues that we might think of as operating outside the pitch system."(page 247)

This point has been demonstrated by McAdams [161], using a paradigm in which the odd and even harmonics of an oboe sound were separated and sent to different speakers. When the two sets of harmonics were coherently frequency modulated, a single source was heard with a single pitch. However, when the odd and even harmonics were incoherently modulated, two sounds were heard that had different pitches. Hence, it appears that perceptual grouping determines pitch, rather than vice versa. This conclusion may also be supported by the finding of Darwin and Ciocca [62] described in section 4.3.1, which indicates that a harmonic with a different onset time makes a reduced contribution to the pitch of a complex tone. Note, however, that Darwin and Ciocca's data could be explained by peripheral adaptation at the frequency of the leading harmonic, rather than by perceptual grouping mechanisms. An effect of grouping could be confirmed by a demonstration that differences in *offset* affect perceived pitch, for the reasons discussed in section 4.3.1. Clearly, this is a research issue.

Finally, the A&S and M&H segregation strategies both suffer from the problem of *overlapping harmonics* (Assmann and Summerfield [8]). Consider the autocorrelation map of the double vowel /a/ (fundamental 112 Hz) and /e/ (fundamental 100 Hz) shown in figure 5.4. The channel of this map with centre frequency 898 Hz is shown in figure 5.7. It is dominated by the eighth harmonic of the /a/, which has a frequency of 896 Hz. Peaks occur in the autocorrelation function at the period of this harmonic (1.12 ms) and at integer multiples of this period. A large peak

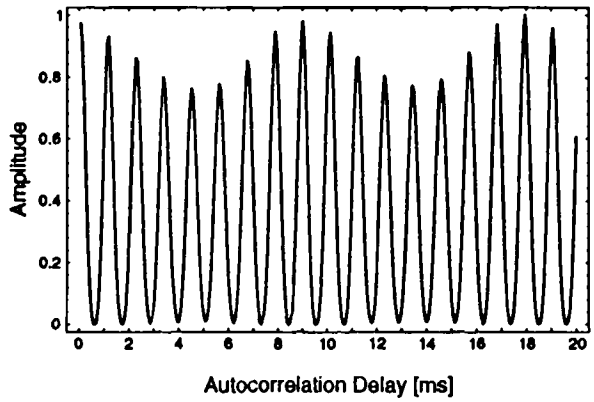

Figure 5.7: Autocorrelation function for the channel of the map with centre frequency 898 Hz. The channel is dominated by the eighth harmonic of the vowel /a/, which has a frequency of 896 Hz. A large peak occurs at the pitch period of the /a/ (8.93 ms), and a smaller peak occurs at 10.04 ms which is near to the pitch period of the /e/.

occurs at a delay of eight periods (8.93 ms), which corresponds to the pitch period of the /a/. However, there is also a smaller peak at a delay of nine periods (10.04 ms), which is close to the pitch period of the /e/ (10.0 ms). Since the /e/ is the dominant vowel in the mixture, the M&H strategy initially removes the channels of the map which have a peak at a delay of 10.0 ms. Consequently, the channels dominated by the 896 Hz harmonic of the /a/ are incorrectly assigned to the /e/ (see figure 5.5). A similar error is made by the A&S strategy, since the peak in the channel autocorrelation function at 10.04 ms is almost as large as the peak at 8.93 ms. Hence, the "synchrony spectrum" sampled at the pitch period of the /e/ contains spurious energy in the region of 896 Hz. This point is discussed further in the next section.

5.2.2 A New Strategy

In this section, a new autocorrelation-based segregation strategy is presented which avoids many of the limitations of the A&S and M&H schemes. Firstly, the basic principles of the new strategy are discussed. Following this, the application of the strategy to the grouping of auditory objects is described.

Principles of the Strategy

Recall from section 4.2.4 that the summary autocorrelation of a periodic sound has peaks at integer multiples of the pitch period, as well as a peak at the pitch period itself. In order to reduce the influence of these "false" pitch peaks on the segrega-

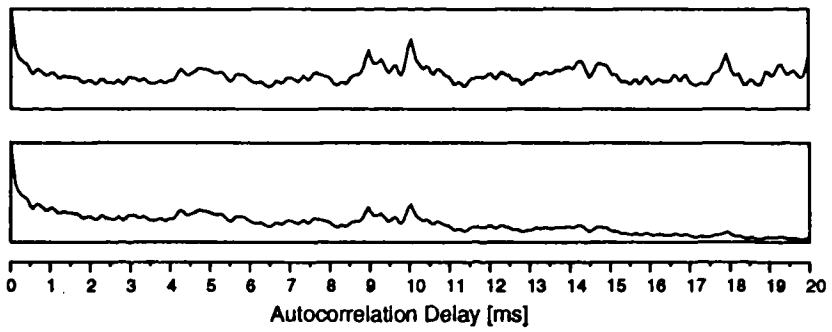

Figure 5.8: Summary autocorrelation functions before weighting (upper panel) and after weighting (lower panel). The "false" peak at twice the pitch period of the /a/ (17.86 ms) has been attenuated by the weighting.

tion strategy described here, a weighting is applied to the summary autocorrelation which attenuates peaks at longer delay times. Specifically, a modified summary autocorrelation

$$s_w[t, \Delta t] = \frac{w[\Delta t]}{M} \sum_{f=1}^{M} acm[t, f, \Delta t] \tag{5.1}$$

is computed, where the weighting function $w[\Delta t]$ is defined by

$$w[\Delta t] = 1.0 - 0.9 \frac{\Delta t}{\Delta t_{max}} \tag{5.2}$$

as suggested by Weintraub [277]. Here, Δt_{max} is the longest autocorrelation delay, and the other parameters are defined in section 4.2.3. The function $w[\Delta t]$ imposes a linear weighting on the summary autocorrelation, which varies from 1.0 at zero delay to 0.1 at the longest delay. This ensures that the peak at the pitch period of a source is larger than the peaks at integer multiples of the pitch period.

The weighted summary autocorrelation $s_w[t, \Delta t]$ is an average measure of the periodicities present in the autocorrelation map. As such, it indicates the likelihood of a pitch period Δt occurring in the map at time t. Similarly, the channel autocorrelation functions $acm[t, f, \Delta t]$ indicate the likelihood of a particular pitch period occurring in a channel of the map. Therefore, the product of these two quantities gives an estimate of the probability[†] that a channel f belongs on a pitch period Δt at time t,

$$Pr[t, f, \Delta t] = acm[t, f, \Delta t] s_w[t, \Delta t] \tag{5.3}$$

[†]Note that the term "probability" is used loosely. $Pr[t, f, \Delta t]$ is not a true probability since, in general, $\sum_{\Delta t} Pr[t, f, \Delta t] \neq 1$.

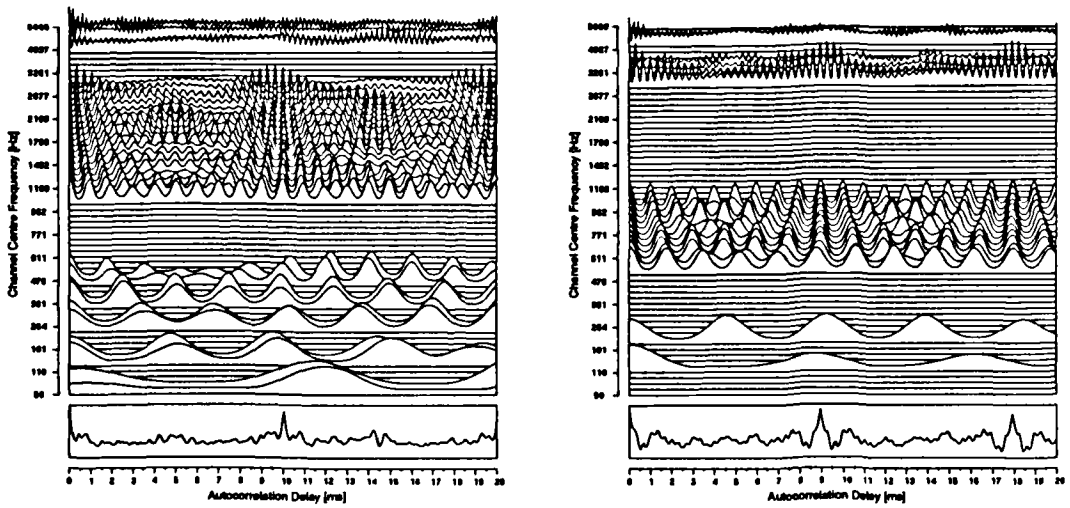

Figure 5.9: Segregation of the double vowel by the new strategy. The group of channels on the left corresponds to the /e/, and the group on the right corresponds to the /a/.

From equation 5.3, it is possible to predict the pitch period that a channel is most likely to belong on. Specifically, the predicted pitch period $p[t, f]$ is given by the autocorrelation delay at which $Pr[t, f, \Delta t]$ is highest,

$$p[t, f] = \max_{\Delta t} Pr[t, f, \Delta t] \tag{5.4}$$

Here, $p[t, f]$ is computed for values of Δt between 2 ms and 20 ms, corresponding to pitches in the range 50 Hz to 500 Hz (see section 4.2.3). Segregation can now be achieved by application of the following grouping principle:

> *Channels of the autocorrelation map are grouped together if they have the same predicted pitch period* $p[t, f]$.

This strategy is illustrated in figures 5.8 and 5.9, for the double vowel /a/ and /e/ shown in figure 5.4. Initially, the weighted summary autocorrelation is computed, which is shown together with the conventional summary autocorrelation is figure 5.8. The spurious peak at twice the pitch period of the /a/ (17.86 ms) has been attenuated by the weighting, as required. Subsequently, the pitch period of each channel is predicted, and channels with the same pitch period are grouped. The two largest groups found by this process, which account for 80% of the channels in the map, are shown in figure 5.9. The group on the left of the figure has a pitch period of 10.0 ms, and corresponds to the /e/. Similarly, the group on the right has a pitch period of 8.93 ms, and corresponds to the /a/. The remaining channels of the map form small groups, or fail to group at all.

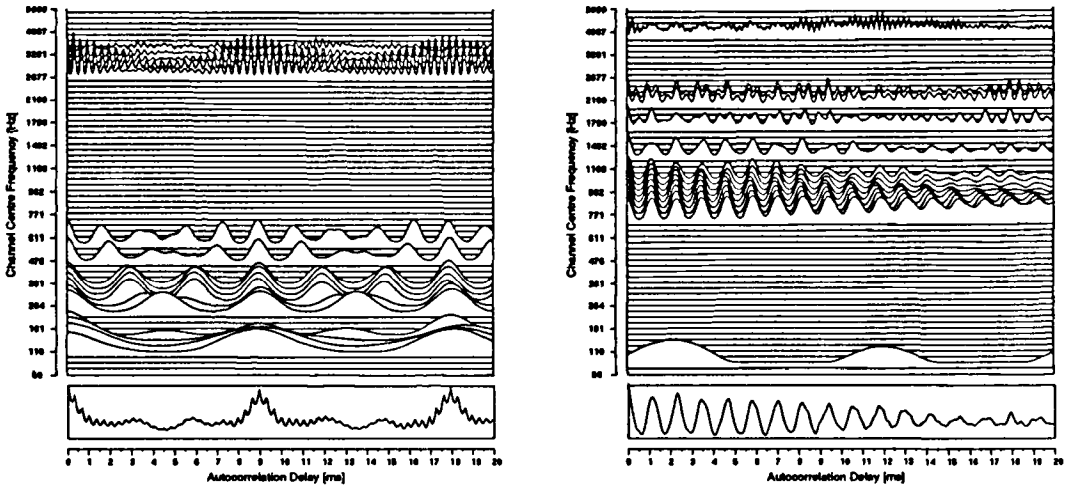

Figure 5.10: Segregation of the vowel /a/ from the siren noise source by the new strategy. The group of channels on the left corresponds to the /a/, and the group on the right corresponds to the siren.

This approach has a number of advantages when compared with the A&S and M&H strategies. Firstly, no prior knowledge of the number of sources present in the stimulus is required. Rather, the number of groups that are formed is determined by the number of different predicted pitch periods. Secondly, because the strategy determines the *most likely* pitch period for each channel, it tolerates small irregularities in the channel autocorrelation functions. For example, a comparison of figures 5.5 and 5.9 shows that the new strategy has correctly assigned several low-frequency channels to the /e/ that were incorrectly grouped with the /a/ by the M&H scheme. Thirdly, the new strategy does not attempt to identify a *global* pitch for each source. Rather, it predicts a *local* pitch for every channel in the map, and groups channels with the same local pitch. This approach is consistent with the view that grouping determines the perceived pitch of a source, rather than vice versa. Additionally, the strategy is robust in situations where there are many spurious peaks in the summary autocorrelation function. For example, the mixture of the vowel /a/ and siren noise source, shown in figure 5.6, was previously considered as a difficult stimulus for the A&S and M&H segregation strategies. The two largest groups found by the new scheme for this mixture are shown in figure 5.10. Inspection of the summary autocorrelation functions suggests that the vowel and siren have been segregated very effectively.

It is also apparent from figure 5.9 that the new strategy is able to solve the problem of overlapping harmonics. The channels in the region of 896 Hz have been assigned to the /a/, as required. Note that this result is not due to any change in the dominance of the two vowels caused by the weighting of the summary autocorrelation function. If the M&H scheme were to use the weighted summary autocorrelation, it would still produce the groups shown in figure 5.5, since the largest peak still occurs at

the 10.0 ms pitch period of the /e/ (see figure 5.8). Rather, the new strategy is able to solve the problem of overlapping harmonics because of two factors. Firstly, channels of the autocorrelation map are allocated exclusively to one source. This point is discussed further in section 5.5. Secondly, the strategy uses information about the height of the pitch period peak in the summary autocorrelation *and* in the channel autocorrelation. The A&S and M&H schemes do not take both of these factors into account.

Although the new strategy solves the problem of overlapping harmonics in many situations where the A&S and M&H schemes cannot, it is not guaranteed to do so in every case. Again, consider the double vowel /a/ and /e/. The channels near to 896 Hz are correctly assigned by the new strategy because the product of the summary and channel autocorrelation functions at the pitch period of the /a/ is larger than the product at the pitch period of the /e/ (see figures 5.7 and 5.8). However, if the pitch period peak of the /a/ in the weighted summary autocorrelation was much smaller than the pitch period peak of the /e/, the strategy would fail and the channels near to 896 Hz would be incorrectly grouped with the /e/. This problem could be minimized by exaggerating the differences in peak height in the channel autocorrelation functions. One way of achieving this would be to square the autocorrelation function in each channel of the map.

Another strategy for solving the problem of overlapping harmonics has been proposed by Summerfield *et al.* [266]. They attempt to identify local pitches in an autocorrelation map by convolving adjacent channels with Gabor functions. However, this approach is computationally expensive and fails at low frequencies where harmonics are resolved.

Grouping Auditory Objects by Common Periodicity

The implementation of the new segregation strategy described in this section exploits the fact that temporal continuity has been made explicit in the auditory object representation. Rather than comparing predicted pitch periods at each time frame, a temporally-extensive *pitch contour* is computed for each object in the auditory scene. Subsequently, objects are grouped if their pitch contours are similar.

As before, the probability of each pitch period is predicted by computing the product of the channel and summary autocorrelation functions. However, auditory objects generally occupy more than one channel of the autocorrelation map at each time frame (see section 5.1.2). Therefore, a *local summary autocorrelation* is computed, which averages the channel autocorrelation functions over the frequency spread of the object. For an auditory object which occupies channels f_1 to f_2 of the autocorrelation map at time t, the local summary autocorrelation $l[t, f_1, f_2, \Delta t]$ is given

by

$$l[t, f_1, f_2, \Delta t] = \frac{1}{f_2 - f_1 + 1} \sum_{f=f_1}^{f_2} acm[t, f, \Delta t] \tag{5.5}$$

where $acm[t, f, \Delta t]$ is the channel autocorrelation function defined in equation 4.11. Note that the effect of this averaging will be small, since the channels occupied by an object are, by definition, very similar. Now, using the same rationale as for equation 5.3, the probability of the object belonging on a particular pitch period Δt at time t is given by

$$Pr[t, f_1, f_2, \Delta t] = l[t, f_1, f_2, \Delta t] s_w[t, \Delta t] \tag{5.6}$$

Here, $s_w[t, \Delta t]$ is the weighted summary autocorrelation previously defined in equations 5.1 and 5.2.

As described in the last section, the most likely pitch period could be estimated from equation 5.6 in a frame-by-frame manner. However, this approach does not take advantage of temporal continuity. Instead, $Pr[t, f_1, f_2, \Delta t]$ is computed at every time frame occupied by the auditory object, and the best path through this series of functions is found by a *dynamic programming* algorithm. Dynamic programming (Cooper and Cooper [54]) is a mathematical technique in which a stepwise decision-making process is used to find an optimal solution (in this case, the most likely pitch contour for an auditory object).

Since the optimum pitch contour passes through peaks in $Pr[t, f_1, f_2, \Delta t]$, the dynamic programming algorithm actually finds the best path through the series of functions

$$m[t, \Delta t] = \begin{cases} Pr[t, f_1, f_2, \Delta t] & \text{if } \frac{\partial}{\partial \Delta t} Pr[t, f_1, f_2, \Delta t] = 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.7}$$

Here, $m[t, \Delta t]$ is zero except at delays where a local maximum in the pitch probability occurs. Equation 5.7 can be computed by using a finite difference approximation to the differential, checking the sign of zero crossings to ensure that a maximum has been found rather than a minimum.

Now, the dynamic programming algorithm proceeds as follows. The dynamic programming score $ds[t, \Delta t]$ for a pitch period Δt at time frame t is defined as the dynamic programming score at the previous frame, plus the transition score gained by moving to the current pitch period. Formally, the recursive relation

$$ds[t, \Delta t] = \begin{cases} ds[t-1, \Delta t_{prev}] + \max_{\Delta t} ts[\Delta t_{prev}, \Delta t, t] & \text{if } t_s < t \leq t_e \\ 0 & \text{if } t = t_s \end{cases} \tag{5.8}$$

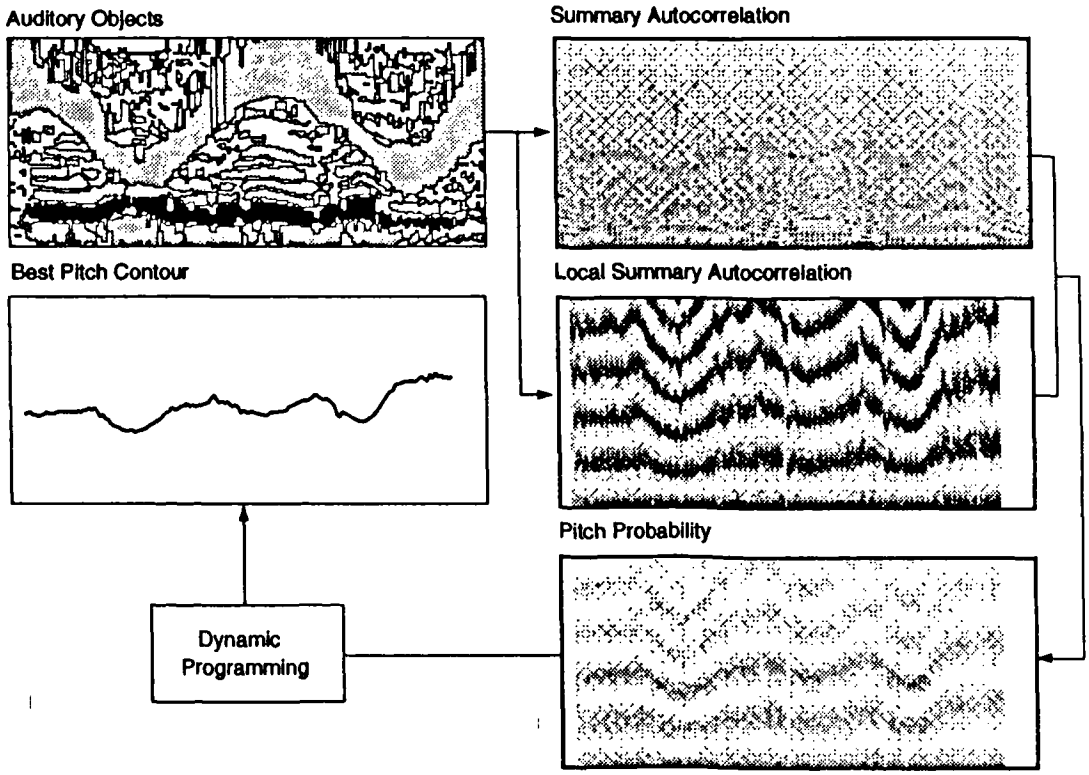

Figure 5.11: Strategy for finding the most likely pitch contour for an auditory object (highlighted in black). The product of the summary autocorrelation and local summary autocorrelation gives the pitch probability. The best path through the pitch probability corresponds to the pitch contour of the object.

is calculated for each time frame t between the start time t_s and end time t_e of the auditory object. The transition score $ts[\Delta t_{prev}, \Delta t, t]$ quantifies the cost of moving from a pitch period Δt_{prev} in the previous frame to a pitch period Δt in the current frame, and is given by

$$ts[\Delta t_{prev}, \Delta t, t] = m[t, \Delta t] \exp\left[-\frac{(\Delta t - \Delta t_{prev})^2}{2\delta_t^2}\right] \tag{5.9}$$

Hence, the transition score for a new pitch period depends upon its probability, and its distance from the previous pitch period. The exponential term in equation 5.9 applies a Gaussian weighting to the pitch period difference, so that smaller changes in pitch period give a higher transition score. In the absence of any experimental data, the standard deviation δ_t of the Gaussian was derived empirically. A value of 0.6 ms was found to give good results.

A dynamic programming score $ds[t_s, \Delta t_s]$ is computed for each initial pitch period Δt_s, and the pitch period with the highest score is taken to be the start of the best path. Subsequently, the best path is retraced through the series of functions

$m[t, \Delta t]$ in order to determine the pitch contour. This process is repeated for each object in the auditory scene.

Figure 5.11 illustrates the procedure for one object in a mixture of voiced speech and the siren noise source. The local summary autocorrelation $l[t, f_1, f_2, \Delta t]$ for the object highlighted in black is shown in the middle right panel of the figure. Multiplying $l[t, f_1, f_2, \Delta t]$ with the summary autocorrelation $s_w[t, \Delta t]$ gives the pitch probability $Pr[t, f_1, f_2, \Delta t]$, illustrated in the bottom right panel. Note that the pitch contour of the object is clearly defined in this representation. The dynamic programming algorithm finds the best path through the pitch probability function, giving the pitch contour shown on the left of the figure. Pitch contours for the remaining objects in this example are illustrated in figure 5.12. Two distinct groups are visible, corresponding to the pitches of the speech and siren. Additionally, a small number of contours occur at twice the pitch period of the speech, which are due to sub-octave errors in the tracking procedure.

Given a predicted pitch contour for each object in the auditory scene, segregation can now be achieved by application of the following grouping principle:

> *Auditory objects which overlap in time are grouped together if their predicted pitch contours are sufficiently similar.*

For two objects that overlap in time, the similarity of their pitch contours $p_1[t]$ and $p_2[t]$ can be quantified by the metric

$$sim[p_1, p_2] = \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} \exp\left[\frac{-(p_1[t] - p_2[t])^2}{2\delta_p^2}\right] \tag{5.10}$$

Here, t_1 and t_2 define the first and last time frames at which the two objects overlap. This similarity metric computes the average Gaussian-weighted difference between the two pitch contours. As such, $sim[p_1, p_2]$ varies between unity (indentical pitch contours) and zero (very different pitch contours). The standard deviation δ_p of the Gaussian determines the amount of tolerance in the comparison. Here, δ_p was set to 0.3 ms by inspection.

Finally, two objects are allowed to form a group if their $sim[p_1, p_2]$ score exceeds a threshold value. In practice, the pitch contours of objects that belong to the same source tend to be very similar, so the threshold can be set quite high. A value of 0.9 is used here. Clearly, this process groups auditory objects in a pairwise manner. Section 5.4 describes a strategy for searching the auditory scene which forms larger groups from these pairwise comparisons.

Superficially, the technique presented in this section appears to be similar to the pitch tracking procedure described by Weintraub [277]. However, there are important differences between the two approaches. Firstly, Weintraub uses dynamic programming to track a *global* pitch contour for each source. Here, we compute a

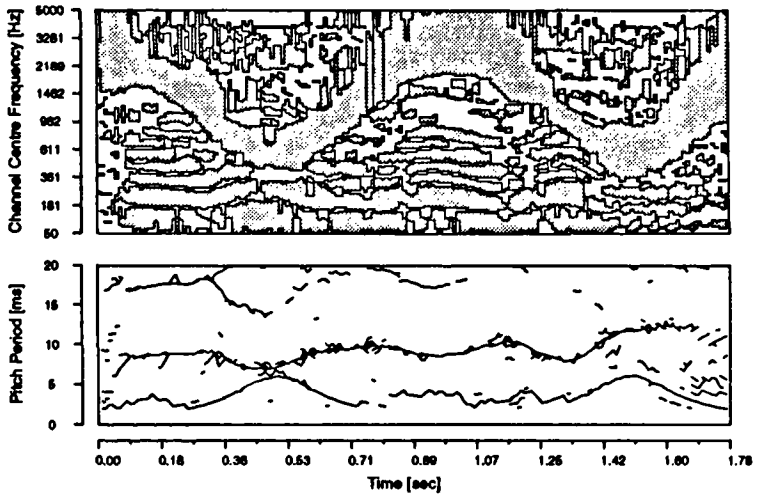

Figure 5.12: Auditory object representation of speech and siren intrusion (upper panel), and the pitch contours for each object (lower panel). Note that the pitch contours cluster into two groups, corresponding to the pitch of each source.

local pitch contour for each object, and group objects which are likely to belong to the same source. The advantages of the latter approach have already been discussed in section 5.2.1. Secondly, Weintraub's system attempts to track the pitch period of each source using a representation which is equivalent to the summary autocorrelation function. Here, the pitch contour of an object is derived from its pitch probability function. It is apparent from figure 5.11 that the summary autocorrelation contains many peaks which can disrupt the tracking process. Conversely, the pitch contour of an object is clearly defined in its pitch probability function, so that tracking is relatively straightforward. Weintraub's segregation system is discussed in detail in chapter 7.

5.3 Grouping by Common Onset and Common Offset

Recall from section 4.3.1 that the auditory system tends to group acoustic components which have the same onset and offset time. A simple way of implementing this process in the model would be to group auditory objects which start and end synchronously. However, the birth-death tracking strategy described in section 5.1.2 tends to break an object in situations where a tracking error is likely to occur. Therefore, the start and end times of an object do not necessarily correspond to the appearance and disappearance of an acoustic event.

The onset and offset maps described in section 4.3 provide a solution to this problem, since the presence of activity in the maps verifies that an onset or offset of an acoustic event has occurred. Therefore, the following principle can be applied to

group objects with a common onset or offset time:

> *Auditory objects which start or end synchronously are more likely to form a group, providing that there is sufficient activity in the onset or offset map at the appropriate time.*

In practice, objects tend not to be exactly synchronous, so it is desirable to allow a tolerance in the comparison of onset and offset times. Darwin [60] and Roberts and Moore [222] find onset and offset segregation effects at disparities of 30 ms, so the tolerance should clearly be less than this (see section 4.3.1). Here, objects are judged to be synchronous if the difference between their start or end times is not more than two time frames, corresponding to a tolerance of 20 ms.

Given the start or end time of an object, the onset or offset map is checked to ensure than an acoustic event has actually started or stopped. Again, it is desirable to allow a tolerance when comparing the start/end time of an object with the time of activity in the onset/offset map. This is because auditory filters tend to ring at their centre frequencies for a few milliseconds after an abrupt onset, which delays the formation of periodicity groups. Similarly, periodicity groups may extend for a few milliseconds after a sudden offset, because the filters continue to ring at the frequency of the stimulus. Therefore, the activity $act[t]$ in the onset or offset map $o[t, f]$ at the start/end time t of an auditory object is quantified by

$$act[t] = \sum_{f=f_1}^{f_2} \sum_{\tau=-2}^{2} o[t+\tau, f] \tag{5.11}$$

Here, f_1 and f_2 define the range of channels in the filterbank occupied by the object during its first (in the case of onset) or last (in the case of offset) time frame. As before, a two frame (20 ms) tolerance is allowed either side of the start/end time t. An onset or offset is indicated when the activity in the map $act[t]$ exceeds a threshold value. Here, onset and offset maps with a 1 ms delayed input are used, which respond to abrupt changes in stimulus amplitude (see section 4.3.3). In practice, *any* activity in these maps is a reliable indication than an onset or offset has occurred, so the threshold can be set to zero.

When two auditory objects start or end at the same time, and an onset or offset is indicated by the maps, the objects are more likely to form a group. In the model, the tendency of two objects to group is increased by adding a constant weighting to the similarity score $sim[p_1, p_2]$ defined in equation 5.10. The weightings for common onset and common offset are both set to 0.5. Recall from section 5.2.2 that objects are allowed to fuse if their $sim[p_1, p_2]$ score exceeds a threshold value of 0.9. Therefore, objects which have a common onset *and* a common offset will form a group regardless of their pitch contour similarity, since their onset and offset score (1.0) exceeds the threshold. However, objects with a common onset *or* a common offset must also have a pitch contour similarity of at least 0.4, in order to exceed the

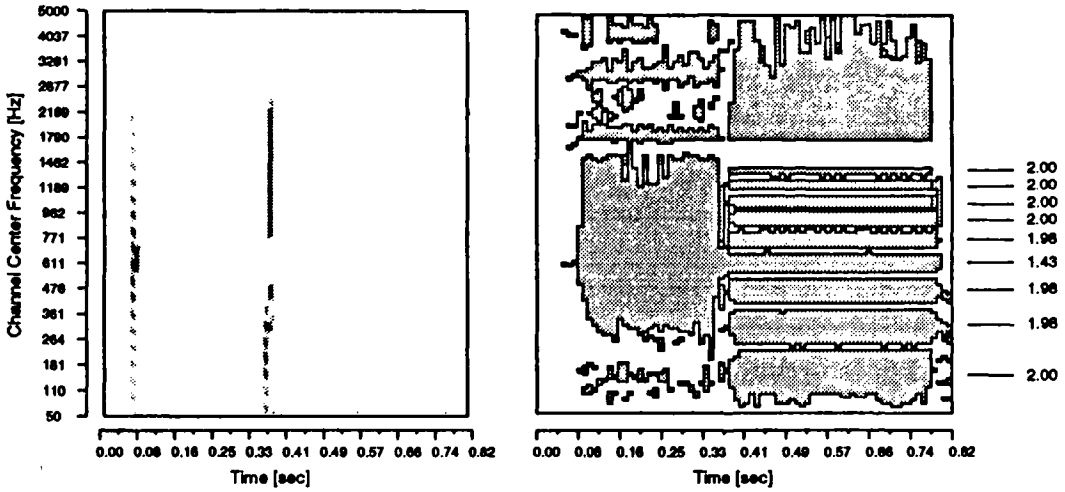

Figure 5.13: Onset map (left) and auditory objects (right) for the harmonic complex used by Darwin and Ciocca. The figures on the right indicate the similarity between the lowest harmonic and the other harmonics. Note that the fourth harmonic, which starts 30 ms before the other harmonics, has a low similarity score.

threshold and form a group. The reasons for imposing this constraint are discussed in the next section.

The principles of this approach are illustrated in figure 5.13, which shows onset map and auditory object representations of the 12-component harmonic complex used by Darwin and Ciocca [62] (see section 4.3.1). The fourth harmonic of this complex leads the remaining components by 30 ms. On the right of the figure, $sim[p_1, p_2]$ scores are given for the similarity between the first harmonic and the other harmonics. The first harmonic has a maximum score of 2.0, since it is effectively compared with itself. Additionally, the other harmonics (except the fourth) have a similar onset time, offset time and pitch contour as the first harmonic, so their scores are near to 2.0. The fourth harmonic has a lower $sim[p_1, p_2]$ value, since it does not gain any score from common onset.

Note that in order to segregate the fourth harmonic from this complex, the grouping threshold in the model would have to be raised above 0.9. Consequently, the model cannot explain Darwin and Ciocca's results in its current form. If the threshold was raised, objects would only group if they had very similar pitch contours *and* a common onset or offset. This would be undesirable, since the frequency components of many environmental sounds (such as speech) start and stop at different times. See section 5.5 for further discussion of this point.

In the model, the same weighting is applied to objects in the case of common onset and common offset. Effectively, this assumes that perceptual grouping mechanisms treat onsets and offsets as equally important events. For short stimuli, this is probably the case (see section 4.3.1). Darwin [60] and Roberts and Moore [222] find offset effects that approach the magnitude of onset effects for short stimuli, and

suggest that the remaining difference is due to the contribution of peripheral adaptation to onset asynchrony. However, the difference between onset and offset effects is greater when longer stimuli are used. Whether this result reflects an increased effect of adaptation, or a preference of perceptual grouping mechanisms for using onsets rather than offsets, is a research issue.

Finally, the model presented in this section implies that there is an equivalence between synchrony and periodicity grouping cues. The validity of this hypothesis could be investigated psychophysically. For example, it may be possible to quantify the amount of onset asynchrony that is equivalent to a particular amount of mistuning using the paradigm described by Darwin and Ciocca [62].

5.4 Searching the Auditory Scene

The process described in this section aims to partition the auditory scene into groups of objects that are likely to belong together. An algorithmic *search strategy* is employed, which takes advantage of the time-frequency object representation described in section 5.1.2. Similar schemes have recently been proposed by Cooke [52] and Mellinger [170].

Currently, the strategy groups objects according to their periodicity, onset time, and offset time. As such, the algorithm models *primitive* auditory scene analysis, and does not attempt to use learned (schema-driven) grouping principles. Additionally, the strategy is limited to searching for simultaneous organization in the auditory scene, and is not able to group objects that are widely separated in time. However, the time-frequency nature of the representation used here does allow the sequential propagation of groups in situations where objects overlap.

5.4.1 Motivation

The issues that arise in formulating a strategy for searching the auditory scene have been comprehensively discussed by Cooke [52]. Here, a new strategy is proposed that is motivated by several of Cooke's observations, although it is substantially different from the algorithm described in his thesis. Cooke's model is discussed in detail in chapter 7.

Firstly, the strategy employed here assumes that every object in the auditory scene must be allocated to a group. Hence, the search terminates when all of the objects in the scene have been accounted for. In some cases, a group may consist of a single object.

A second point concerns the allocation of objects between groups. Recall from section 5.2.2 that channels of the autocorrelation map are allocated exclusively to one source by the segregation strategy. Consequently, an auditory object cannot belong to more than one group, and once it has been assigned to a group it is

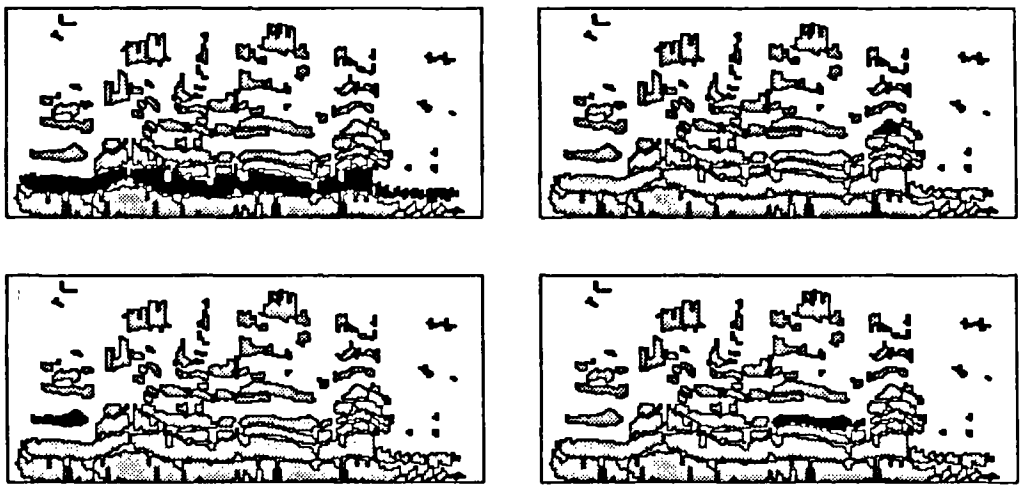

Figure 5.14: Rigid application of exclusive allocation in the model does not prevent the same groups from being found by searches that start from different objects. Here, the same group in a natural speech signal has been found by searches starting from the four objects highlighted in black.

effectively "removed" from the auditory scene. The theoretical implications of this approach are considered in section 5.5.

One potential problem in rigidly applying a principle of exclusive allocation is that the search strategy may find different organizations in the auditory scene if it starts from different objects. For example, consider a situation in which a frequency component can belong to two harmonic series, so that different groups are competing for the same object. If components were grouped according to their harmonic relations, the object would be assigned arbitrarily to the harmonic series that was identified first by the search strategy. However, the algorithm proposed here does not suffer from this problem, for two reasons. Firstly, objects are grouped according to the similarity of their predicted pitch contours, rather than by harmonicity *per se*. It is very unlikely that the pitch contours of two groups will be so similar that they will compete for the same objects. Secondly, exclusive allocation is not imposed at the level of the search strategy. Rather, it emerges as a consequence of the fact that objects are assigned to a single predicted pitch contour. This point is illustrated by figure 5.14, which shows that the same groups are found by searches that start from different objects.

In practice, the search time can be reduced by starting from "dominant" objects in the auditory scene, as suggested by Cooke [52]. Here, the length of an object is taken as an indication of its dominance, although other metrics (such as time-frequency area) could also be used. Long objects generally give rise to large groups, and are likely to have a significant acoustic correlate. Therefore, the search for a new group starts with the longest object in the auditory scene, and long objects are recruited to groups before shorter objects.

5.4.2 A Search Strategy

The algorithm used to search the auditory scene is shown in figure 5.15, in the form of a flow diagram. Since exclusive allocation applies, objects are essentially "removed" from the scene when they are assigned to a group. The algorithm iterates until there are no ungrouped objects left.

Initially, the longest object in the auditory scene is selected as the start of a new group. Then, every object remaining in the scene is considered as a possible match (*focus*) to the group. A similarity score $sim[p_1, p_2]$ is calculated between the pitch contour of the focus object and every object in the group that it overlaps, as described in section 5.2.2. Subsequently, the score is adjusted if the objects being compared have a common onset or a common offset (see section 5.3). If the focus object has a $sim[p_1, p_2]$ score greater than 0.9 for every object in the group that it overlaps, it is added to the group. This process iterates until the group cannot recruit any more objects. Then, a new group is started if there are any objects remaining in the auditory scene.

Note that objects are recruited to groups under very tight constraints. Specifically, a focus object can only be recruited to a group if it is sufficiently similar to *all* the members of the group that it overlaps in time. This constraint is imposed to prevent small objects in a group from acting as a "bridge" to dissimilar objects. For example, a focus object which generally has a different pitch contour to a group, but is similar for a short time, could be recruited by an object in the group that spans the period when the pitch tracks are similar. Checking that the focus object is consistent with every member of the group alleviates this problem.

An example of the search strategy is shown in figure 5.16, for the grouping of a voiced speech signal. The search starts from the longest object in the auditory scene, which in this case corresponds to the fundamental (recall from figure 5.14 that the search could be initiated from another object with the same results). Subsequently, objects with a similar pitch contour (lower panel in each diagram) are recruited to the group, longest first. The algorithm proceeds very efficiently, and identifies the majority of the objects which belong to the group by iteration 36. The group is terminated at iteration 84, when no more objects can be recruited. Note that although this strategy only seeks simultaneous organization, the nature of the auditory object representation allows a group to extend across time as well as across frequency.

5.4.3 Examples of Grouping

Several examples of grouping by the model are illustrated in figures 5.17 to 5.20, for stimuli in which an intrusive noise has been added to voiced speech. In each figure, the upper left panel shows the auditory objects for the speech and the upper right panel shows the objects for the noise. The middle panel illustrates the resulting mixture, and the largest group found by the search strategy is shown at the bottom of the figure.

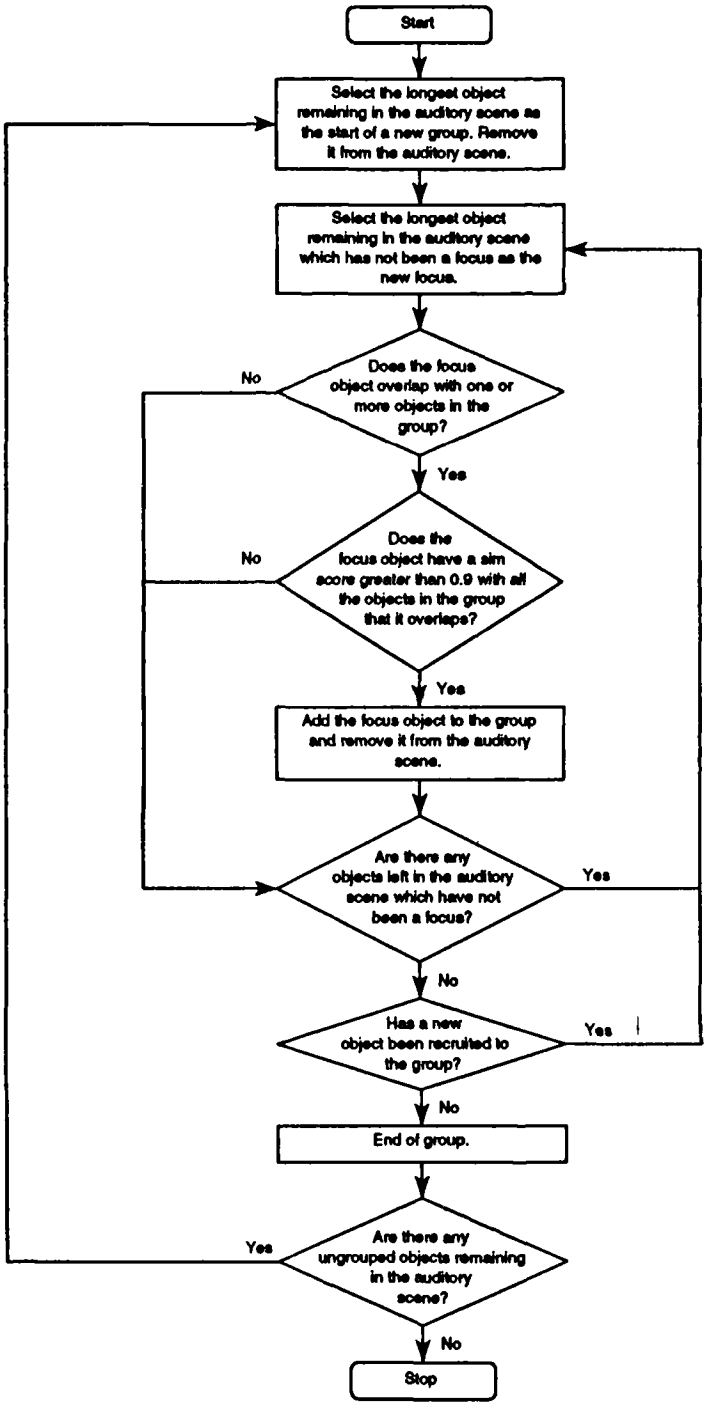

Figure 5.15: Flow diagram for the search strategy. See text for details.

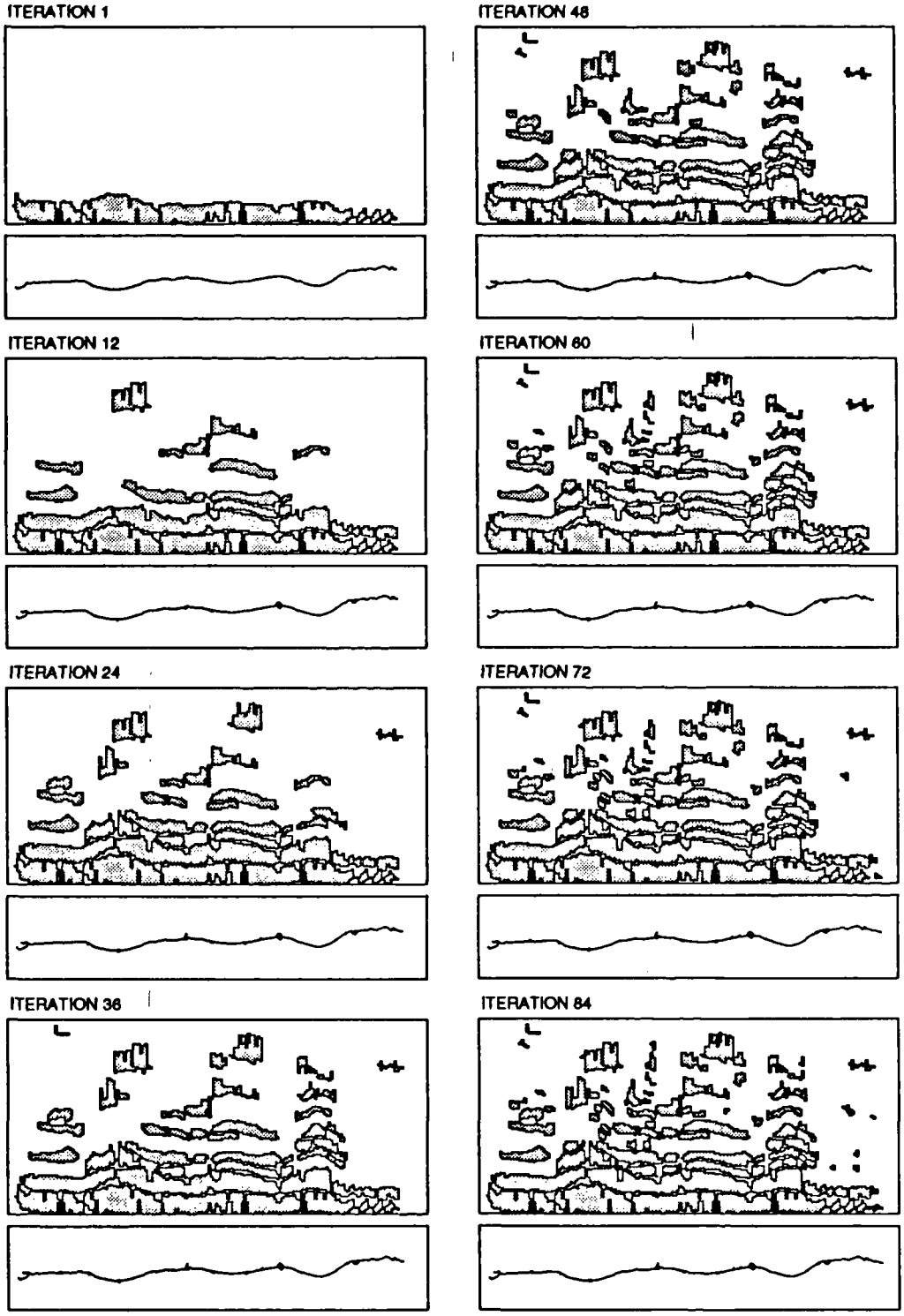

Figure 5.16: Example of the search strategy for voiced speech. Auditory objects are shown in the upper panel of each diagram, and the pitch contours of the objects are shown in the lower panel. Time is represented on the abscissa, and frequency (50Hz-5kHz) is represented on the ordinate.

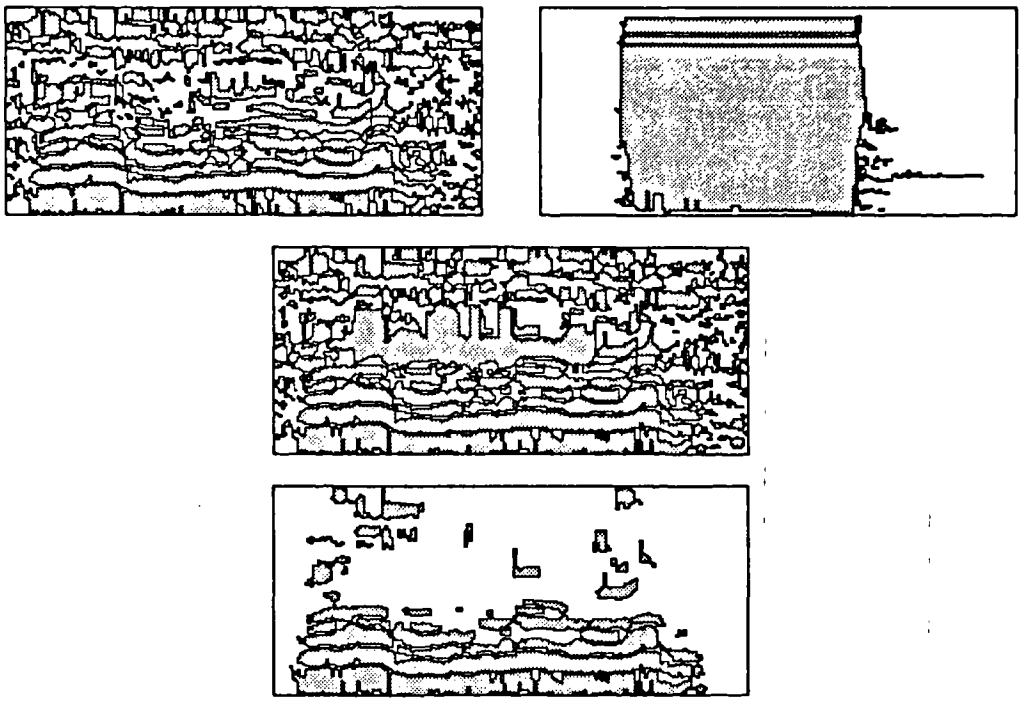

Figure 5.17: Segregation of voiced speech (v4) from a 1kHz tone intrusion (n0).

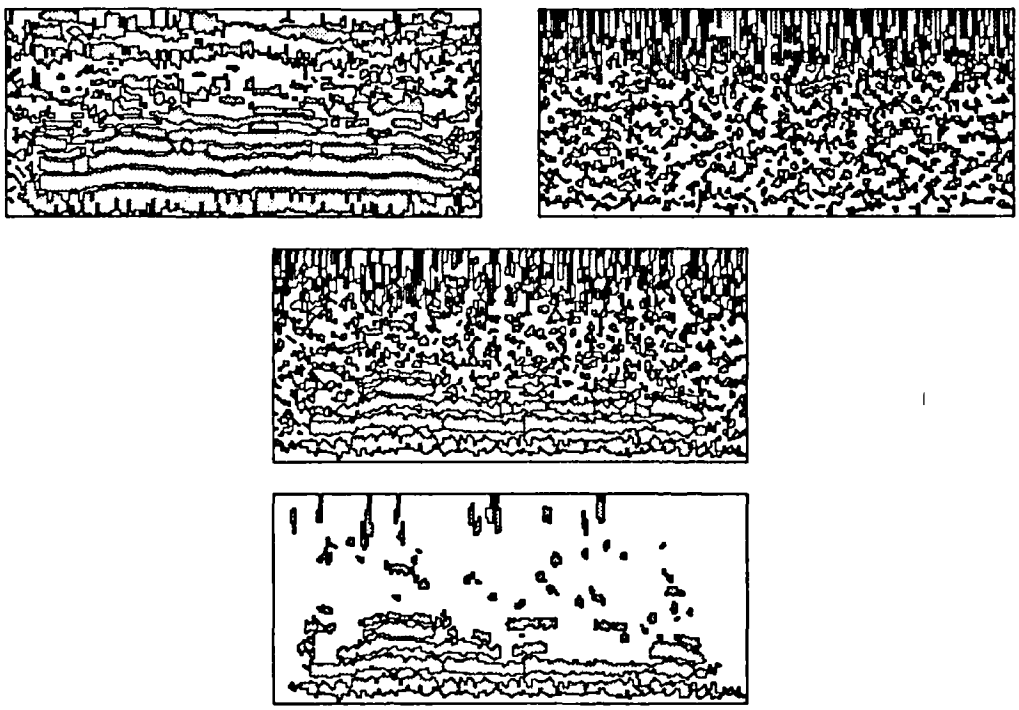

Figure 5.18: Segregation of voiced speech (v9) from a random noise intrusion (n1).

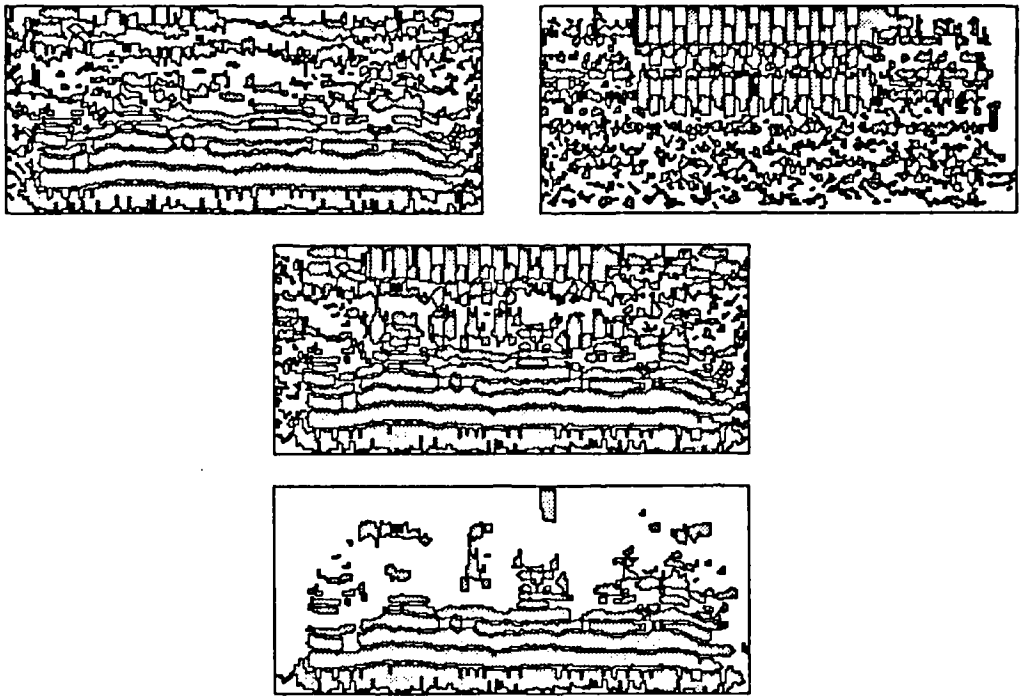

Figure 5.19: Segregation of voiced speech (v9) from a telephone intrusion (n6).

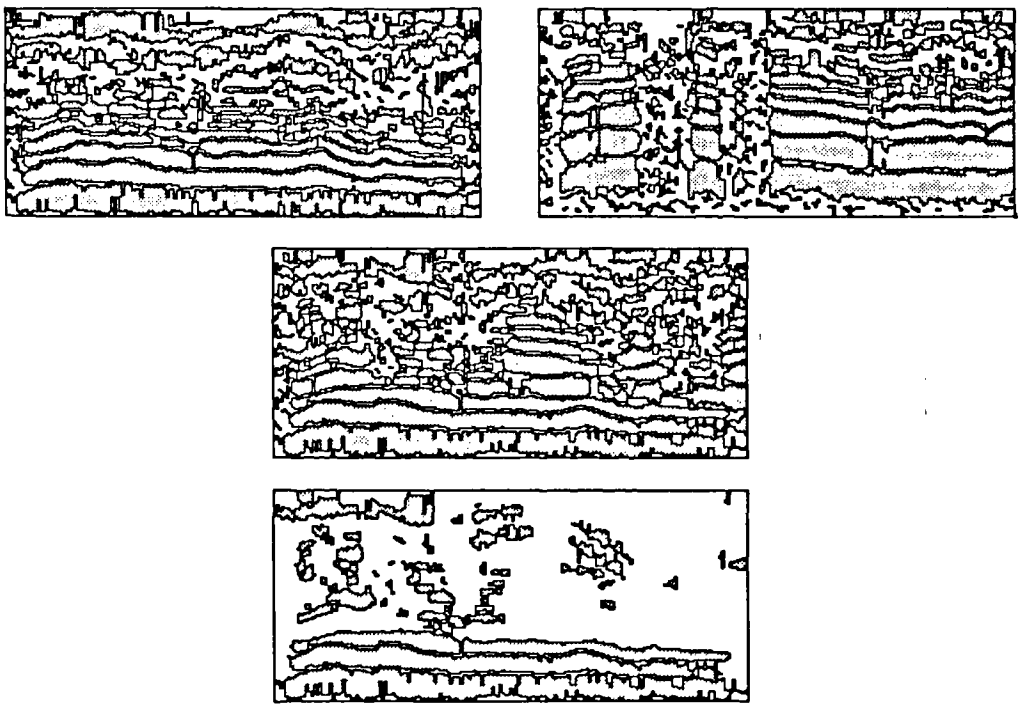

Figure 5.20: Segregation of voiced speech (v6) from a speech intrusion (n7).

In figure 5.17, the intrusion is a 1 kHz tone. The largest group found in the mixture corresponds to the speech, and the tone has been almost completely removed. Note that a conventional autocorrelation scheme would have difficulty in segregating this mixture, since tonal stimuli give rise to many spurious peaks in the summary autocorrelation function (see section 5.2.1 and figure 5.10). The segregation of speech from random noise, shown in figure 5.18, has been less successful. However, the grouping strategy has recovered a number of harmonics that are visible in the mixture. Figure 5.19 illustrates the segregation of speech from a "trill" telephone. The intrusion has significant energy in the frequency regions of speech formants, so that only the low harmonics of the speech are clearly defined in the mixture. This group of harmonics has been successfully recovered by the search strategy. Finally, figure 5.20 shows a mixture of two speech signals. Again, most of the objects belonging to the voiced speech have been recovered.

5.4.4 Derivation of Global Properties from Groups

Many properties of an acoustic source, such as pitch, timbre, loudness and direction, are derived from groups of frequency components rather than individual components. Bregman [24] calls these *global properties.* In this section, the derivation of pitch and timbre from groups of auditory objects is considered.

Derivation of pitch from a group of objects can be achieved in two ways. Firstly, a summary autocorrelation can be computed at each time frame by averaging the autocorrelation functions of channels that are included in the group. Formally, the summary autocorrelation at time delay Δt for a group of channels $G[t]$ is given by

$$s_g[t, \Delta t] = \frac{1}{M} \sum_{f=1}^{M} g[t, f, \Delta t] \tag{5.12}$$

where

$$g[t, f, \Delta t] = \begin{cases} acm[t, f, \Delta t] & \text{if } f \in G[t] \\ 0 & \text{otherwise} \end{cases} \tag{5.13}$$

Here, M is the number of channels occupied by the group at time t, and $acm[t, f, \Delta t]$ is the autocorrelation function of channel f. Summary autocorrelation functions of this form are illustrated in the lower panels of figures 5.9 and 5.10. If the segregation has been successful, $s_g[t, \Delta t]$ contains a large peak at the pitch period of the group. Again, note that derivation of global pitch in this manner is consistent with the view that grouping determines perceived pitch, rather than vice versa (see page 110).

Alternatively, the global pitch of a group can be computed by averaging the local pitch contours of every object in the group. In practice, a weighted mean $p_g[t]$ of

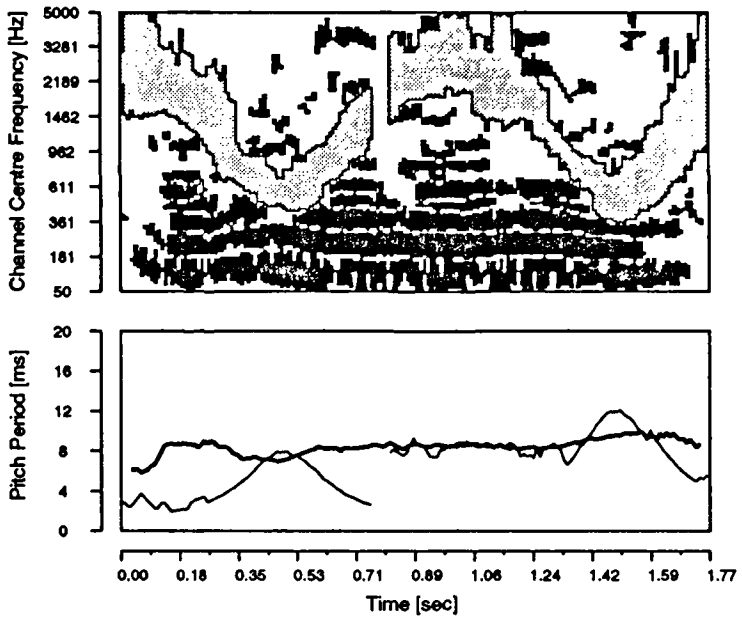

Figure 5.21: Derivation of global pitch contours from groups of objects. Three groups are shown (one is highlighted in black), which consist of objects from a mixture of speech and a siren intrusion. Note that the pitch contours cross in several places.

the local pitch contours is used, where

$$p_g[t] = \frac{\sum_{i=1}^{N} p[i,t]w[i,t]}{\sum_{i=1}^{N} w[i,t]} \tag{5.14}$$

and

$$w[i,t] = f_2[i,t] - f_1[i,t] + 1 \tag{5.15}$$

Here, N is the number of objects in the group, $p[i,t]$ is the pitch contour of object i at time t, and $f_1[i,t]$ and $f_2[i,t]$ define the range of frequency channels occupied by an object. The weight $w[i,t]$ ensures that objects with a wide frequency spread make a larger contribution to the mean than objects with a narrow frequency spread. Figure 5.21 shows global pitch contours derived in this way for three groups in a mixture of speech and a siren intrusion. Note that a conventional autocorrelation scheme would have difficulty in segregating this mixture, since the pitch contours of the two sources cross in several places. However, the time-frequency representation employed here allows the mixture to be separated quite effectively (a similar observation has been made by Cooke [52]). Although a subharmonic of the pitch has been incorrectly tracked for the second half of the siren, this has not affected the ability of the search algorithm to segregate the two sources.

Parameter	Description	Value	Units
$w[f]$	width of birth-death acceptance region	0.6	ERB
δ_t	standard deviation of transition score Gaussian	0.6	ms
δ_p	standard deviation of similarity score Gaussian	0.3	ms
	threshold similarity score for group formation	0.9	
	score weighting for common onset	0.5	
	score weighting for common offset	0.5	

Table 5.1: Parameter settings for the pitch tracking strategy and search algorithm.

Meddis and Hewitt [169] suggest a method for deriving timbre from an autocorrelation map. The summary autocorrelation function defined in equation 4.12 is partitioned into two sections, corresponding to a "pitch region" at longer delays and a "timbre region" at shorter delays. Hence, the "timbre region" of the summary autocorrelation function contains information about the higher-frequency components of the group. However, it is not clear where the boundary between the two regions should be placed, and the regions may overlap in certain situations. Further research is needed to resolve these problems.

5.5 Summary and Discussion

This chapter has presented a representation of objects in the auditory scene, and a search strategy which groups objects with a similar pitch contour, onset time and offset time. The parameter settings used in this part of the model are summarized in table 5.1. On the following pages, a number of theoretical and computational issues are discussed.

Exclusive Allocation of Objects

The model assumes that the principle of exclusive allocation is rigidly applied, so that an object can only belong to one group. This assumption is supported by several experiments, such as the one by Bregman and Pinker [30] discussed in section 4.3.1. Here, a tone A was alternated with a pair of tones B and C, as shown in figure 4.11. Recall that B either grouped sequentially with A, or formed a simultaneous organization with C. The principle of exclusive allocation clearly applies in this example, since B cannot belong to both groups at the same time.

Nonetheless, it is clear that there are many situations in which the principle of exclusive allocation is violated. An example occurs in the "duplex" perception of speech, described by Liberman *et al.* [151] and Rand [210]. Liberman [148] presented listeners with a synthetic three-formant syllable at one ear, from which the third formant transition was removed. When the missing transition was played simultaneously to

the opposite ear, it contributed to the percept of the syllable but was also heard as an isolated "chirp". Clearly, the principle of exclusive allocation is violated in this case, since the formant transition is heard as a part of two sounds. Several other examples of "duplex" perception have already been reviewed, such as the findings of Moore *et al.* [185] described in section 4.2.1. Here, a component of a harmonic complex that was mistuned by 2-3% contributed fully to the pitch of the complex, but was also heard as a separate source. Hence, the mistuned harmonic was shared by two organizations at the same time. Similarly, Gardner *et al.* [90] found that a formant of a synthetic syllable which was incoherently frequency modulated still contributed to the syllable percept, but was heard as a separate sound (see section 4.5.2).

These results, and those of similar experiments, have led Bregman [24] to conclude that

> "The principle of exclusive allocation does not describe the activity of primitive scene analysis. The latter involves competitions between groupings, but such competitions do not have to be resolved in an all-or-nothing manner."(page 637)

Although this is probably the case, rigid application of the principle of exclusive allocation in the model has a number of advantages. Firstly, exclusive allocation of channels by the autocorrelation map segregation strategy allows the problem of overlapping harmonics to be solved in many situations. Secondly, assigning objects to a single group simplifies the search algorithm. For example, there is no redundancy in the groups that are found which needs to be resolved at a later stage (see the discussion of Cooke's thesis in chapter 7).

It should be stressed that an object which could belong to two groups is not assigned arbitrarily. Rather, the object is assigned to the group that it is *most likely* to belong to, on the basis of its predicted pitch contour. Whether this is a serious limitation of the model is a question that requires further research. Certainly, relaxation of the exclusive allocation constraint would demand a new strategy for solving the problem of overlapping harmonics, which was able to share channels of the autocorrelation map between groups.

Default Condition: Separation or Fusion?

In the model, it is assumed that objects in the auditory scene are segregated unless there is evidence to group them together. However, segregation may not be the default condition of organization. Rather, the auditory system may prefer to *fuse* all the components in the auditory scene, so that objects are only segregated when there is evidence for doing so. Bregman [24] makes an observation that may distinguish between these two alternatives. He notes that although white noise contains random fusion and segregation cues, it is heard as a coherent sound rather than a multitude

of components in different frequency regions. This suggests that fusion is the default condition, rather than segregation.

It would be relatively straightforward to make fusion the default condition in the model. For example, an object could be rejected from a group if its $sim[p_1, p_2]$ score with the members of the group was sufficiently low. Whether this approach would have any advantages over the algorithm presented here is an issue for further research.

Divisibility of Auditory Objects

Once formed, auditory objects are not subjected to any further modification in the model. For example, an object cannot be split across time or frequency. A possible limitation of this approach is suggested by an experiment by Darwin and Sutherland [65], reviewed in section 4.3.1. They measured the changes in vowel percept that were caused by adding a tone to the first formant region of a vowel. In one condition, the tone started 30 ms before the vowel, so that the stimulus was similar to the one shown in figure 5.13. When a harmonic of the leading tone was added which stopped as the vowel started, listeners were more likely to hear a change in the vowel colour. This suggests that the two leading tones formed a separate perceptual group, which ended at the start of the vowel.

Currently, the model cannot reproduce this result, since it requires the object representing the leading tone to be broken at the point where the vowel starts. This action would be a violation of Marr's "principle of least commitment", since it demands that an earlier decision (the formation of an object) should be undone at a later stage of processing. Clearly, this limitation questions the validity of the time-frequency object representation used here. Further research is required to address this issue.

Sensitivity to Similarity Threshold

Hard thresholding is generally avoided until the last stage of the model, where a similarity threshold is applied to determine whether objects should form a group. Clearly, the similarity threshold is an important parameter of the model, but its value has been chosen somewhat arbitrarily.

In practice, this is not a serious limitation. The search strategy tends to find similar groups over a wide range of $sim[p_1, p_2]$ thresholds, since the pitch contours of separate sources are generally very different. Additionally, the autocorrelation map segregation strategy usually tracks the pitch of an object very reliably, so that objects belonging to the same source have very close pitch contours (see figure 5.16). This point is illustrated in figure 5.22, which shows the longest group found in a mixture of two speakers for a range of similarity thresholds. A qualitatively similar organization has been identified in each case. However, a high threshold is preferred

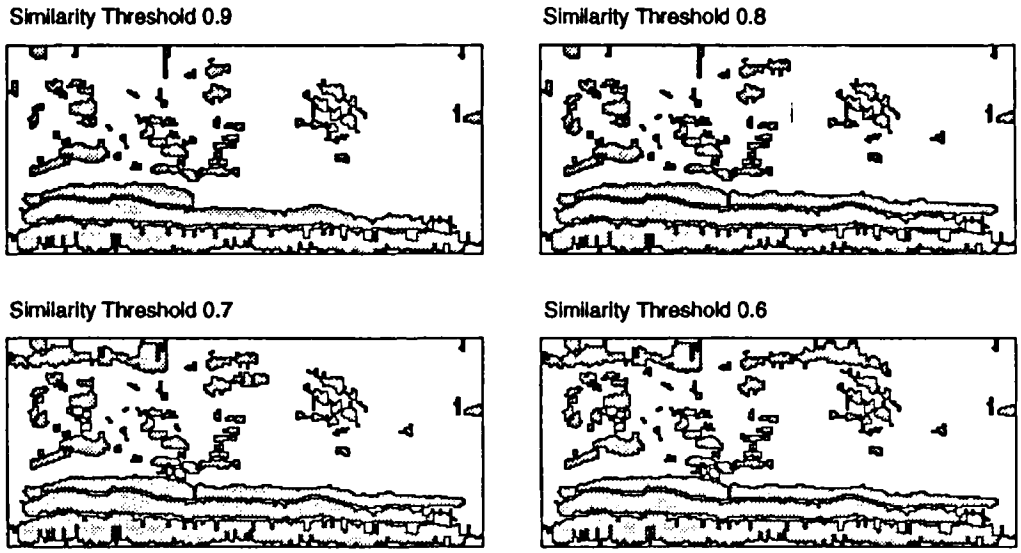

Figure 5.22: Largest group found in a mixture of two speakers for four different threshold similarity scores. The group is qualitatively similar in each case.

here, since it reduces the chance of inappropriately grouping objects in cases where the pitch contours are close.

Sequential Grouping

Currently, the search strategy only seeks simultaneous organization in the auditory scene. However, it is clear that components which are widely separated in time may belong to the same source, such as a sequence of speech sounds from a single speaker. Many properties influence the sequential grouping of sounds, such as timbre, spatial location, temporal proximity, fundamental frequency, intensity and spectral shape (Bregman [24]). Incorporating these grouping principles into the model is a challenging task for future research.

Schema-Based Grouping

Recall from section 3.4 that listeners are able to use learned (*schema-based*) principles to segregate concurrent sounds. A schema for a particular sound appears to become active when the components that it requires are present in the auditory scene, even when no cues for primitive grouping are available. For example, Scheffers [233] found that listeners could identify the constituents of a double vowel with a performance that was above chance, even when the vowels started and stopped at the same time and had the same fundamental frequency (see sections 1.4.3 and 4.2.1).

Currently, only primitive (data-driven) grouping principles are employed in the

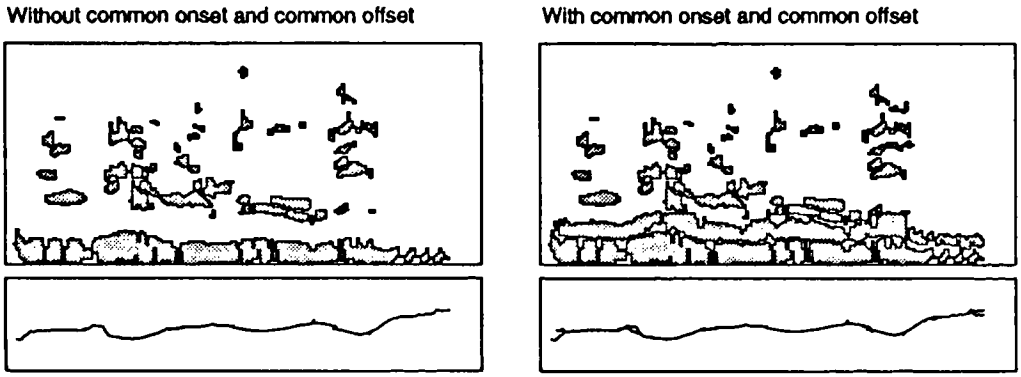

Figure 5.23: Grouping with and without common onset and common offset cues. Synchrony cues allow objects with small irregularities in their pitch contours to be recruited to a group.

model. However, the auditory object representation is equally suited to the application of schema-driven grouping principles, and the search strategy is flexible enough to allow top-down information to influence the groups that are formed. For example, a schema which requires a group of objects could increase their similarity scores, so that the objects are more likely to fuse. A possible role for auditory maps in the formation and application of schemas is discussed in chapter 7.

Retroactive Effects in Grouping

In the search algorithm, objects at a particular time can be recruited to a group that starts at a later time. This assumes that perceptual grouping mechanisms are able to operate retroactively. There is good evidence that this is the case. For example, Darwin [60] has shown that a harmonic which stops after the other components of a vowel can be excluded from the vowel percept (see section 4.3.1). Additionally, Ciocca and Bregman [49] have demonstrated that the perceived continuity of a gliding tone through a band of noise is affected by the characteristics of the post-noise glide (see section 4.4.1). Here, no limit is set on how far the scene analysis strategy can search back through time. In practice, however, perceptual grouping mechanisms may operate over a temporal window of a few hundred milliseconds.

Role of Common Onset and Common Offset

Grouping by common onset and common offset is subject to tight constraints in the model. Generally, objects must have some similarity in their pitch contours (a $sim[p_1, p_2]$ score of at least 0.4) in order for common onset and common offset to be effective. If synchrony was a *sufficient* condition for grouping, onset and offset groups could be propagated inappropriately. Specifically, the search algorithm could form a "staircase" of objects in which adjacent components had similar onset/offset

times, but the onset/offset times of non-adjacent components were widely separated.

Figure 5.23 illustrates a typical situation in which grouping by common onset confers an advantage. Without onset cues, a large component of the group has been excluded because its pitch contour contain small irregularities (left panel). Grouping by common onset allows the object to exceed the similarity threshold and fuse with the group (right panel).

Note that this approach is consistent with the suggestion of Darwin and Sutherland [65] that common onset and common offset are neither necessary nor sufficient conditions for grouping the components of speech (see section 4.3.1). Additionally, the form of the model is compatible with Darwin and Ciocca's [62] suggestion that different perceptual mechanisms may interpret onset time differences in different ways. For example, mechanisms of pitch perception appear to be more tolerant of onset time differences than mechanisms of vowel perception. In the model, this could be simulated by weighting the $sim[p_1, p_2]$ score of two synchronous objects differently for different perceptual mechanisms.

Chapter 6

Evaluation of the Model

Two methods for evaluating sound source segregation in the model are considered here. Firstly, a technique for resynthesizing a waveform from a group of objects in the auditory scene is described, which allows segregation performance to be qualitatively assessed in informal listening tests. Secondly, the resynthesis technique allows a comparison of the relative levels of the signal and noise waveforms before and after segregation, so that performance can be quantified as an improvement in signal-to-noise ratio.

6.1 Resynthesis

A number of workers have used resynthesis to determine whether an auditory representation preserves perceptually important features of the acoustic input (Cooke [52], Ghitza [92], Heinbach [112], Hukin and Damper [122]). For example, Hukin and Damper assess the adequacy of their auditory model by comparing the perceived phonetic categories in original and resynthesized speech signals. The non-auditory work of McAulay and Quatieri [163] and Lienard [153] adopts a similar approach. Resynthesis also provides a convenient means of assessing the performance of systems that attempt to segregate concurrent sounds (Parsons [198], Weintraub [277], Boll [14], Naylor and Boll [191], Cooke [52], Denbigh and Zhao [71]). By listening to the segregated output, it is possible to estimate the amount of signal that has been retained, and the amount of noise intrusion that has been rejected.

It should be noted that validation-by-resynthesis suffers from a number of potential problems. Firstly, perceptual restoration may disguise inadequacies in the resynthesized waveform by "filling in" missing parts of the signal. Secondly, Cooke [52] points out that the resynthesized waveform will be subjected to any nonlinearities twice, once in the model and once in the auditory system of the listener. Finally,

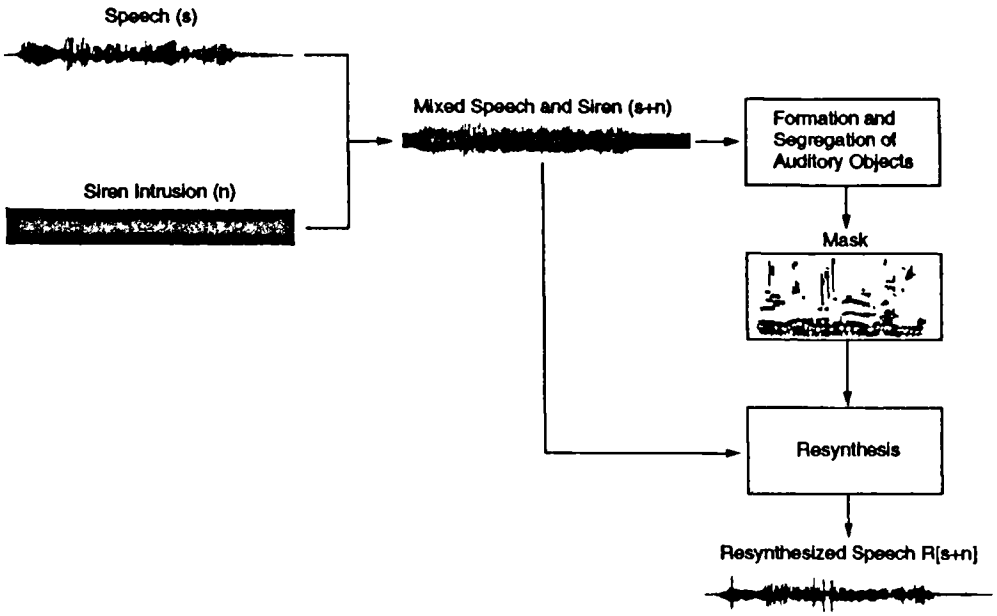

Figure 6.1: Resynthesis of a waveform from auditory objects. A mask is derived from a group of objects, which indicates the time-frequency regions of the gammatone filter output that belong to the group. These regions of the filter output are summed to produce the resynthesized waveform.

assessment of segregation performance by resynthesis may be influenced by the perceptual grouping mechanisms of the listener. Nonetheless, resynthesis is a useful technique for rapid evaluation of the model's performance.

6.1.1 Resynthesis From Auditory Objects

The resynthesis technique employed here is similar to the scheme proposed by Weintraub [277]. Figure 6.1 illustrates the process for a mixture of speech and a siren intrusion, although the technique can be applied to any arbitrary input.

Segregation in the model produces a number of groups of objects, as described in section 5.4. The first stage in resynthesizing a waveform for a group is to form a *mask* (see the figure). If a channel of the auditory filterbank is occupied by an object in the group at a particular time frame, the value of the mask at that time and channel is unity. Otherwise, the value of the mask is zero. Hence, the mask consists of an array of binary weights, that indicate which frequency channels of the filterbank belong to the group at each time frame.

Subsequently, a resynthesized waveform is constructed from the gammatone filter output (see section 4.1.2). In order to remove any across-channel phase differences, the output of each filter is time-reversed, filtered a second time, and time-reversed again. Then, each time-frequency region of the phase-corrected filter output is

multiplied by the corresponding weight in the mask. The weights are applied to 20 ms segments of the filter output, which overlap by 10 ms and are windowed with a raised cosine. Finally, the resynthesized waveform is obtained by summing the weighted filter outputs across all channels of the filterbank.

6.1.2 Informal Listening Tests

The validity of the resynthesis technique has been confirmed by resynthesizing a signal when every element in the mask is unity, so that all of the time-frequency regions of the filterbank output are included (see the appendix). Speech resynthesized in this way is completely indistinguishable from the original utterance. Additionally, segregated speech has been resynthesized from each of the 100 mixtures described in section 6.2.3. Generally, the resynthesized speech is highly intelligible and quite natural. The best exemplars occur when the noise intrusion is narrowband (1 kHz tone, siren), and the worst occur when the noise is random and wideband (laboratory noise, random noise).

Clearly, it is difficult to describe the quality of resynthesized speech in a written document. Hence, an audio cassette is included with this thesis, which contains recordings of a variety of resynthesized signals. The contents of the tape are catalogued in the appendix.

Essentially, listening to the resynthesized waveform of a group is equivalent to "attending" to a particular stream, so that the stream becomes more prominent than the background. This is the auditory equivalent of the Gestalt "figure-ground phenomenon", which was discussed in section 3.1. It is clear, however, that the auditory system does not completely remove unattended streams. For example, when attending to a particular conversation in a "cocktail party" situation, we are still aware of the other conversations in the background. This effect can be simulated in the model by attenuating the time-frequency regions of the gammatone filterbank that do not belong to a group, rather than completely removing them[†]. Several examples are given on the audio tape, where the outputs of unattended channels have been reduced to 5% of their normal amplitude.

6.2 Quantitative Evaluation

A new method for quantitative evaluation of the model is presented in this section, which allows signal-to-noise ratios to be compared before and after segregation. First, some previous approaches to quantitative evaluation are discussed.

[†]The author is grateful to Ray Meddis for this suggestion.

6.2.1 Previous Work

Four methods of evaluation are discussed in this section, with regard to their comparability, ease of interpretation and speed of execution.

Modelling Human Performance on a Specific Task

One method of evaluating a model is to compare its performance on a particular task with the performance of human listeners on the same task. For example, Scheffers [233], Assmann and Summerfield [8] and Meddis and Hewitt [169] have investigated the ability of a model to predict the performance of human listeners in identifying the constituents of a double vowel. This work has already been discussed in sections 5.2.1 and 1.4.3.

This approach suffers from a number of potential problems. Firstly, a model which can predict human performance is not necessarily correct. For example, the A&S and M&H schemes both come close to predicting the overall accuracy of listeners responses, but the two models differ in many important aspects. Secondly, close quantitative agreement between a model and human performance data can be due to "fine tuning" of the model on the test stimuli. Meddis and Hewitt [169] note that

> "The overall level of correct responses is broadly comparable for the model and human listeners. This correspondence should not be overemphasized, however, because of the many opportunities which the modeller has to optimize performance on a small data set." (page 239)

Thirdly, a model which accurately reproduces overall human performance may give highly deviant predictions of the performance on individual exemplars of the test set (Assmann and Summerfield [8]). Finally, although it is possible to model early auditory processing quite closely, higher-level perceptual mechanisms are usually modelled very crudely. For example, the M&H segregation scheme uses *ad hoc* decision rules to determine how many vowels are present.

Automatic Music Transcription

Potentially, a segregation system could be evaluated by determining its ability to accurately transcribe a piece of music. Mellinger [170] describes a system for segregating musical sounds, but he evaluates the model visually rather than quantitatively. As a result, it is difficult to assess the performance of his model. Mellinger's system is discussed in detail in chapter 7.

Intelligibility Tests

If a resynthesis path is available from a model, human listeners can assess the intelligibility of the segregated output in formal listening tests (Hanson and Wong [109], Hanson *et al.* [110]). However, this approach may be time consuming, and subjects require training in order to perform the task.

Alternatively, listeners can be replaced in intelligibility tests by an automatic recognizer. This allows comparatively rapid processing of large test sets. Generally, the test stimulus is speech (Weintraub [277], Gramss and Strube [99]), although the same principle can be applied to other stimuli (Varga and Moore [272]). A potential problem with this approach is that interpretation of results may be difficult if an auditory representation is used as input to the recognizer. For example, Beet [11] has demonstrated that a mismatch can exist between an auditory front-end and a conventional speech recognition system. One solution to this problem is to resynthesize a waveform for a segregated source, and present this directly to an unmodified recognition architecture (Weintraub [277]).

Mixture Component Identification

Cooke [52] describes a scheme that determines which components in a mixture are likely to belong to one source, and which components are likely to belong to another. This methodology allows intuitive metrics to be computed which indicate, for example, the percentage of a segregated mixture that belongs to one source. However, the technique is rather dependent on the specific representation used in Cooke's model, so it is difficult to compare his results with those of other workers. Cooke's system is discussed in detail in section 7.4.

6.2.2 Comparison of Signal-to-Noise Ratios

In this section, a new technique is described which allows a signal-to-noise ratio (SNR) to be computed before and after segregation by the model. This evaluation methodology is fast, simple to implement and leads to an easily interpreted metric. Additionally, quoting the performance of the model in terms of an improvement in SNR allows the results here to be compared with those of other workers.

Generally, the sounds in the test set of mixtures used here are non-stationary (see section 6.2.3). Therefore, a running short-term SNR is computed, which takes the form

$$snr[t] = \frac{2}{\pi} \arctan \left[\frac{\sum_{i=0}^{w-1} s^2[t+i]}{\sum_{i=0}^{w-1} n^2[t+i]} \right] \tag{6.1}$$

where s and n are the speech and noise waveforms respectively. Here, a 10 ms non-overlapping window of size w is used, and results are expressed as the mean $snr[t]$

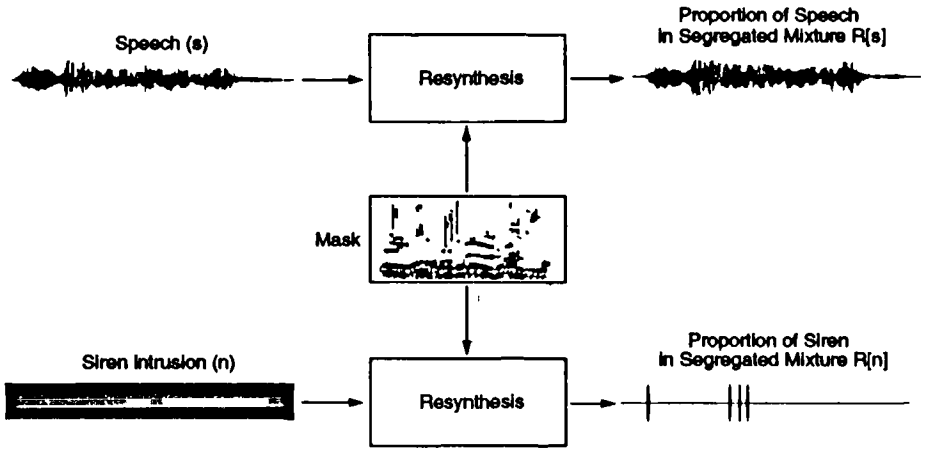

Figure 6.2: Derivation of the proportions of signal and noise in a segregated waveform. Note that in this example, nearly all of the siren intrusion has been removed from the speech.

over every time frame t in the mixture.

Following segregation by the model, all of the noise intrusion n may have been removed within a particular time window (an example is shown in figure 6.2). Clearly, this is a very good result, but it gives rise to an infinite SNR. Hence, an arctangent compression is applied in equation 6.1, which ensures that $snr[t]$ is always finite. In practice, this leads to a highly intuitive metric. When there is no signal in the mixture, $snr[t]$ is zero. Similarly, when there is no noise in the mixture, $snr[t]$ is unity. A $snr[t]$ of 0.5 indicates that the levels of signal and noise are equal.

In order to express the performance of the model as an improvement in SNR, it must be possible to obtain separate signal and noise waveforms *after* segregation. A property of the resynthesis procedure described in section 6.1.1 allows this to be achieved. Recall from section 4.1.2 that the gammatone filter is *linear*. Consequently, the resynthesis process is also linear, since it essentially consists of two passes of gammatone filtering and an across-channel summation. Linear systems satisfy the property of *superposition*, which states that the response of the system to two inputs presented simultaneously is equal to the sum of the responses when the two inputs are presented individually. Formally, for a linear system R,

$$R[s+n] = R[s] + R[n] \tag{6.2}$$

Now, consider the case where the system R represents the resynthesis of a waveform from a particular mask, and s and n represent the signal and noise respectively. Equation 6.2 implies that the proportion of signal in a segregated mixture can be obtained by resynthesizing the signal waveform from the mask, and that the proportion of noise can be obtained by resynthesizing the noise waveform from the

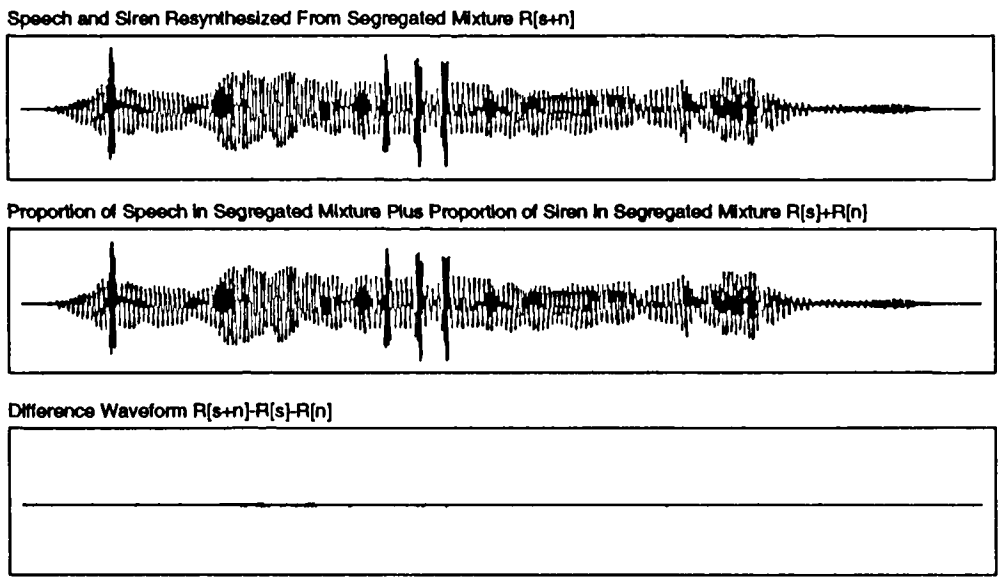

Figure 6.3: Demonstration that the resynthesis process is linear. The two waveforms $R[s+n]$ and $R[s]+R[n]$ are almost identical, as indicated by the difference waveform in the bottom panel.

mask. Hence, separate signal and noise waveforms can be obtained from a segregated mixture. Furthermore, this technique can be applied to any representation from which a linear resynthesis path is available.

This process is illustrated in figure 6.2, for the segregated mixture of speech and siren shown at the bottom of figure 6.1. Resynthesis of the speech waveform from the mask gives $R[s]$, the proportion of speech in the segregated mixture. The proportion of siren $R[n]$ in the segregated mixture is obtained in a similar manner. This approach has a number of useful properties. Firstly, it is possible to compute $snr[t]$ after segregation. Secondly, visual examination of the resynthesized waveforms indicates how much of the signal has been retained, and how much of the noise has been removed. For example, it is clear from figure 6.2 that nearly all of the siren intrusion has been removed by the model. Finally, it is possible to listen separately to the proportion of signal and proportion of noise in the segregated output. Hence, the degradation of the signal and noise waveforms after segregation can be assessed in informal listening tests.

If the resynthesis path is linear, equation 6.2 implies that the resynthesized waveform in figure 6.1 can be obtained by summing the two resynthesized waveforms in figure 6.2. Comparison of $R[s+n]$ and $R[s]+R[n]$ in figure 6.3 confirms that this is indeed the case. Equation 6.2 also implies that

$$R[s+n] - R[s] - R[n] = 0 \tag{6.3}$$

This difference waveform is shown in the bottom panel of figure 6.3. It is nearly

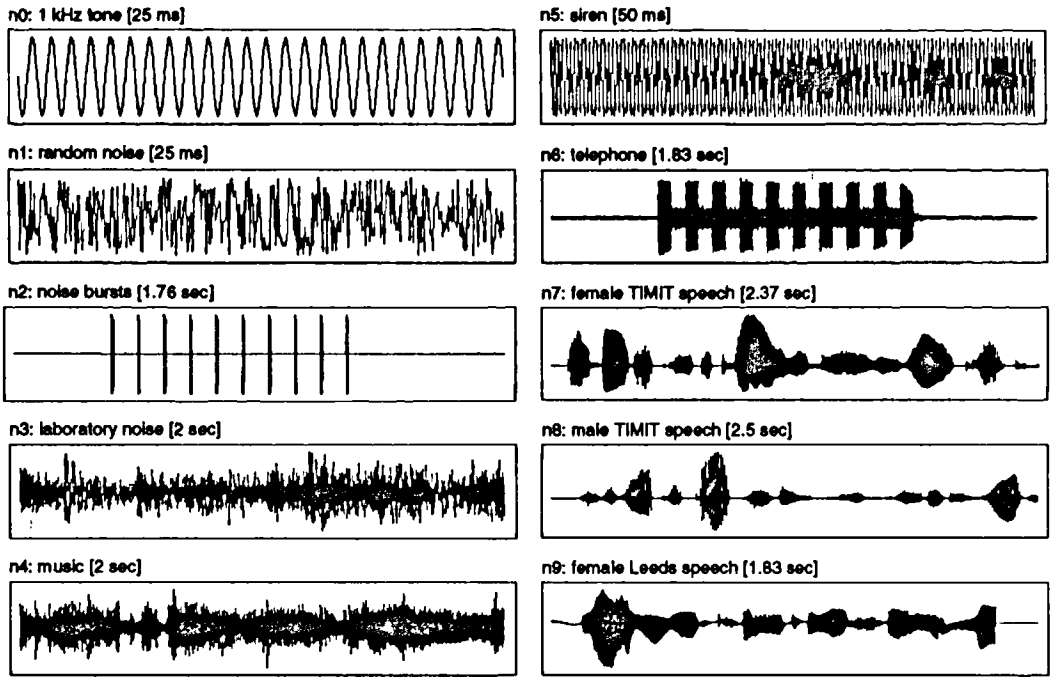

Figure 6.4: Waveforms of the ten noise sources. From Cooke [52], with permission.

always zero, as required. In practice, the energy in the difference waveform is always less than 1% of the energy in the resynthesized mixture. This amount of error was considered to be quite acceptable, and can mainly be attributed to rounding errors in the digital implementation.

6.2.3 The Mixture Test Set

In recent years, psychological studies of auditory scene analysis have prompted a number of computer models. With the arrival of a coherent text from Bregman [24], this trend seems likely to continue. Clearly, standardized evaluation techniques are required in order to compare the strengths and weaknesses of different models. The SNR metric described in the last section allows the results here to be compared easily with those of other workers. Similarly, the quantitative evaluation in this chapter uses the database of speech and noise mixtures employed by Cooke [52].

Although the majority of segregation systems have been evaluated using the task of separating speech from other interfering speech (Weintraub [277], Parsons [198], Naylor and Boll [191], Hanson and Wong [109]), it is clear that a wide variety of noise intrusions occur in natural listening environments. Hence, Cooke's test set contains a range of different noise sources, including synthetic stimuli (1 kHz tone, random noise) and environmental sounds (music and "office" noise). The waveforms of these intrusions are shown in figure 6.4. Additionally, various auditory representations of the intrusions have already been illustrated in chapters 4 and 5.

ID	Speaker	Utterance
v0	speaker 1	I'll willingly marry Marilyn
v1	"	Why were you away a year, Roy?
v2	"	Why were you weary?
v3	"	Why were you all weary?
v4	"	Our lawyer will allow you a rule
v5	speaker 2	I'll willingly marry Marilyn
v6	"	Why were you away a year, Roy?
v7	"	Why were you weary?
v8	"	Why were you all weary?
v9	"	Our lawyer will allow you a rule

Table 6.1: Voiced utterances used in the mixture test set. From Cooke [52], with permission.

Here, the model is evaluated on a set of 100 mixtures, obtained by adding the waveforms of each of the 10 intrusions to each of the 10 utterances listed in table 6.1. The sentences were spoken by two male speakers. Note that fully voiced utterances have been used, since the model (and Cooke's model) is not able to sequentially group a stream of voiced-unvoiced speech sounds (see section 5.5).

Combining separate speech and noise waveforms has a number of advantages over recording speech in the presence of environmental noise. Firstly, the SNR can be computed before and after segregation. Secondly, speakers tend to compensate for the presence of an interfering noise, so that their speech level becomes louder and their voice characteristics change (Lombard [154]). This could affect the evaluation of a model by intelligibility testing, particularly if an automatic recognizer is used.

6.2.4 Results

Each of the 100 mixtures of speech and noise in the test set were processed by the model. The groups corresponding to the speech were identified visually from the auditory object representation, or by listening to the resynthesized waveforms of each group. In every case, the speech corresponded to the longest group of objects that the model identified in the mixture.

Two performance metrics are used here. Firstly, the proportions of speech and noise in each group have been derived, allowing the mean $snr[t]$ to be computed after segregation as described in section 6.2.2. Similarly, the mean $snr[t]$ has been computed for the original mixture, so that performance can be quantified as an improvement in SNR. Secondly, the number of non-zero time-frequency regions (TFR) in the mask is determined for each group. This gives an estimate of how much of the auditory scene has been recovered by the grouping process.

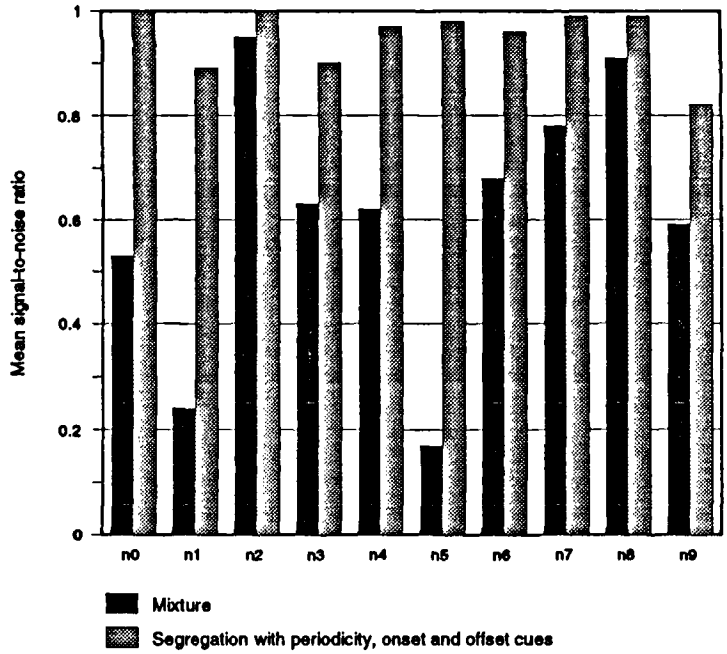

Figure 6.5: Comparison of mean $snr[t]$ before and after segregation using periodicity, onset and offset grouping cues.

In each of the noise conditions the voiced utterances v0-v9 gave similar results using the SNR and TFR metrics. Hence, the results for each intrusion n0-n9 have been averaged over the 10 utterances. Where grouping has been evaluated with no intrusions, results are expressed individually for v0-v9.

Segregation With Periodicity, Onset and Offset Cues

Figure 6.6 shows the distribution of $snr[t]$ values in the original mixtures of speech and noise. The $snr[t]$ values for each intrusion have been averaged over the 10 voiced utterances, and partitioned into 13 bins. Recall that a low $snr[t]$ indicates that there is less signal than noise in the mixture. Hence, many of the histograms in figure 6.6 have significant activity in the lower bins. In particular, the leftwards skew of the random noise (n1) and siren (n5) histograms indicates that the SNR in these mixtures is very unfavourable.

The $snr[t]$ distributions after segregation by the model using periodicity, onset and offset grouping cues are shown in figure 6.7. A pronounced shift to the right is visible in all of the histograms, indicating that the majority of the noise has been rejected. This trend is especially noticeable in the random noise and siren histograms. Additionally, note that the largest bin in the siren histogram now occurs at an $snr[t]$ near to 1, indicating that the intrusion has been almost completely removed. This result would be anticipated from figure 6.2.

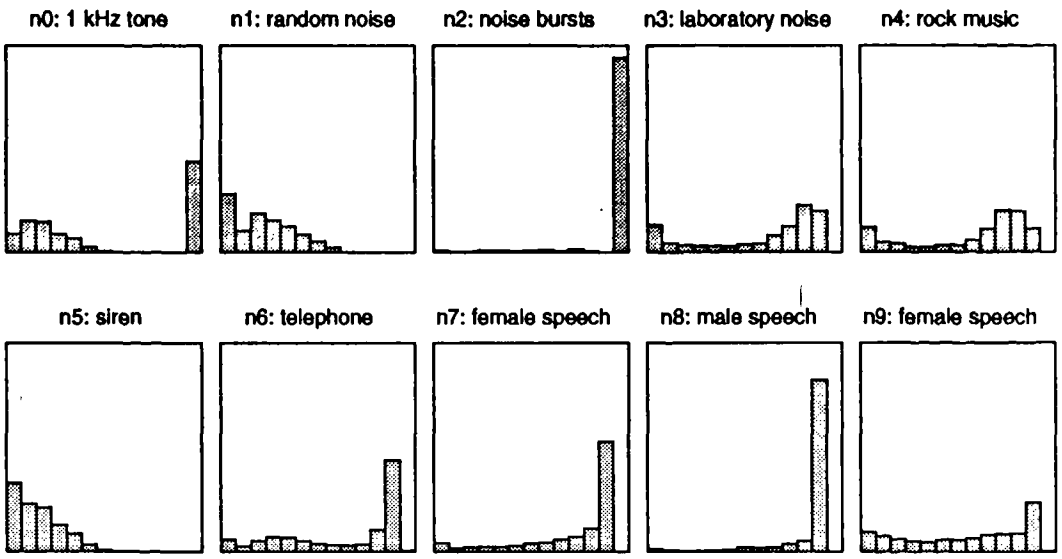

Figure 6.6: Distribution of signal-to-noise ratios in the original mixtures. The $snr[t]$ is represented on the abscissa, between values of 0 (no signal) and 1 (no intrusion).

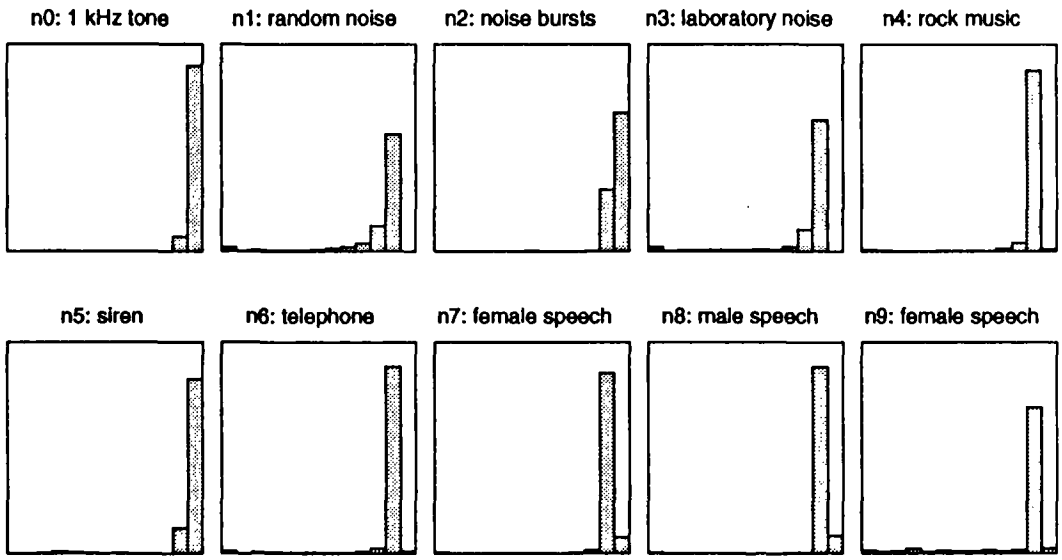

Figure 6.7: Distribution of signal-to-noise ratios after segregation by the model using periodicity, onset and offset grouping cues. The $snr[t]$ is represented on the abscissa, between values of 0 (no signal) and 1 (no intrusion).

A convenient way of summarizing these results is to plot the mean $snr[t]$ for each noise source, as shown in figure 6.5. It is apparent that segregation by the model has improved the mean $snr[t]$ in each case. For some intrusions (n0, n1 and n5) the improvement is very significant.

Segregation With Periodicity Cues Only

The manner in which grouping principles have been implemented in chapter 5 allows the contribution of different cues to be determined. For example, figure 6.8 compares the mean $snr[t]$ before and after segregation for grouping with and without onset and offset cues. In four cases (n0, n1, n4 and n7) grouping by common onset and common offset gives a small advantage compared to grouping only by common periodicity. For the remaining intrusions, segregation performance with and without onset/offset cues is the same. Note that in one condition (n2), maximum performance is obtained using periodicity cues alone, so it is not possible to assess whether grouping by onset and offset would have anything further to contribute.

Figure 6.9 shows the percentage of TFRs that have been included in the mask of each group, with and without onset/offset cues. As expected, more TFRs are included in the masks when onset and offset cues are used. This confirms that grouping by common onset and common offset is effective in recruiting more objects to a group than would be recruited by periodicity cues alone. However, the benefits are small, as would be expected from the discussion in section 5.5.

Random Grouping

The significance of the previous results can be assessed by determining how well a system would perform if it grouped frequency channels randomly at each time frame. Figure 6.10 shows the mean values of $snr[t]$ before and after random grouping. As might be expected, the proportions of signal and noise in a random group are approximately the same as they are in the original mixture. However, small increases in $snr[t]$ occur with some intrusions (n0, n1, n5 and n9). Comparison with figure 6.5 confirms that the performance of the model is significantly better than random grouping for every condition.

Grouping With No Intrusion

Figure 6.11 illustrates the percentage of TFRs that are recruited when no intrusion is present, for each of the utterances v0-v9. Typically, about 35% of the TFRs are recruited. Although this figure seems low, it should be noted that not every TFR in the auditory scene will be occupied by an object (see figure 5.3). Also, errors in the pitch contours of objects, or inadequacies in the grouping rules, may mean that objects are excluded from a group. Resynthesis of the groups evaluated in figure

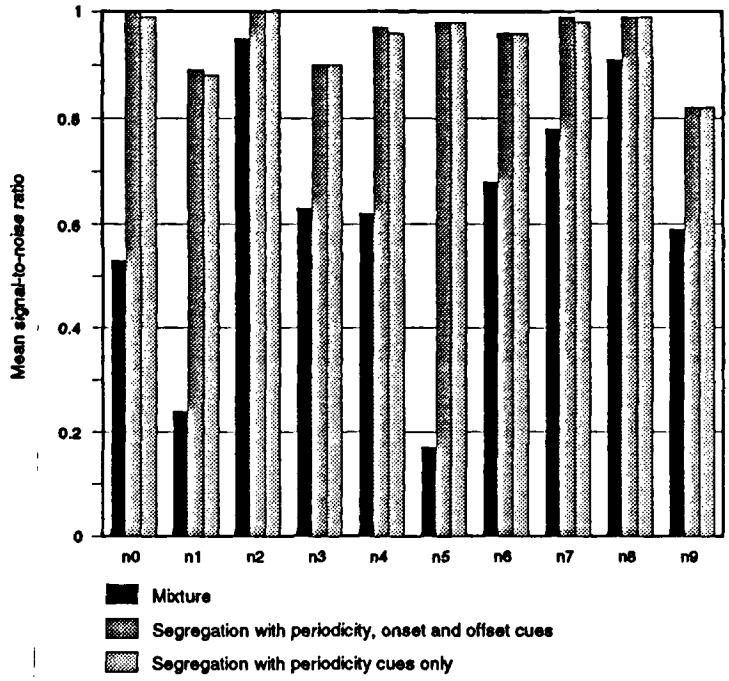

Figure 6.8: Comparison of SNR before and after segregation by the model, with and without onset/offset cues.

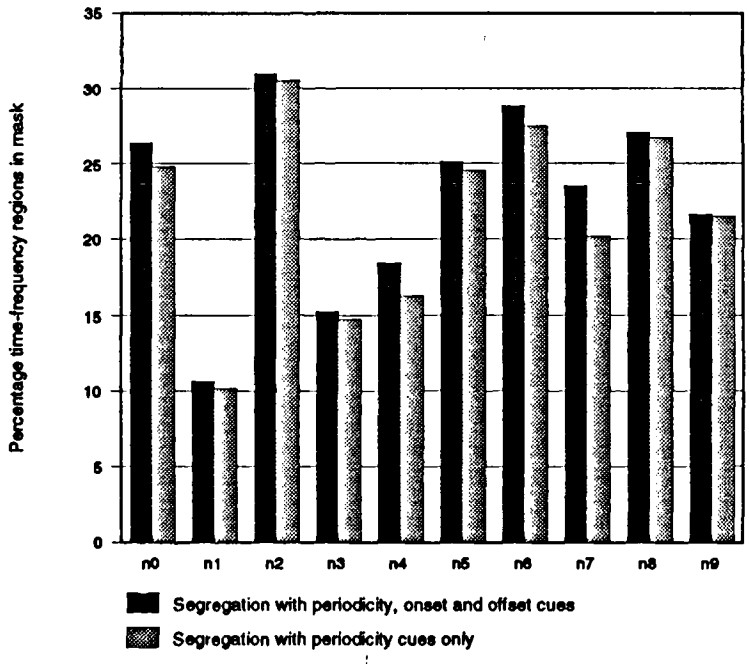

Figure 6.9: Comparison of percentage TFRs recruited by the model, with and without onset/offset cues.

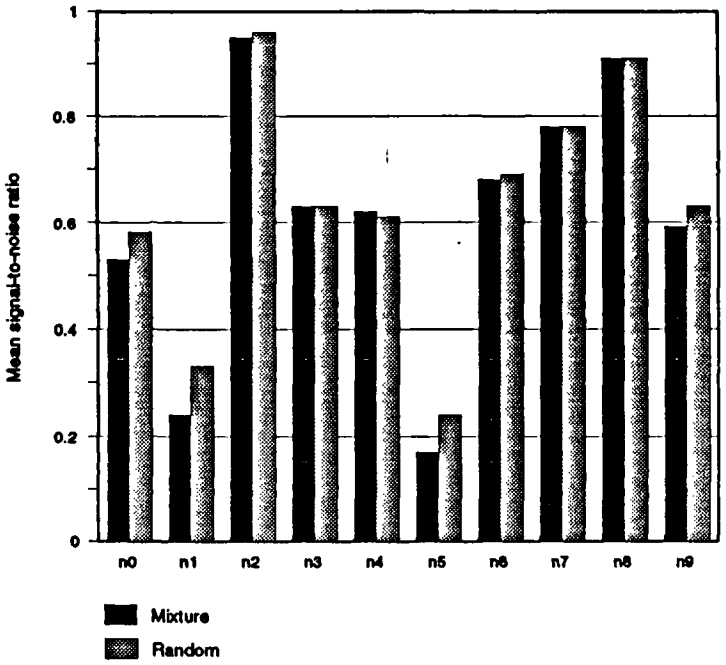

Figure 6.10: Comparion of SNR before and after random grouping.

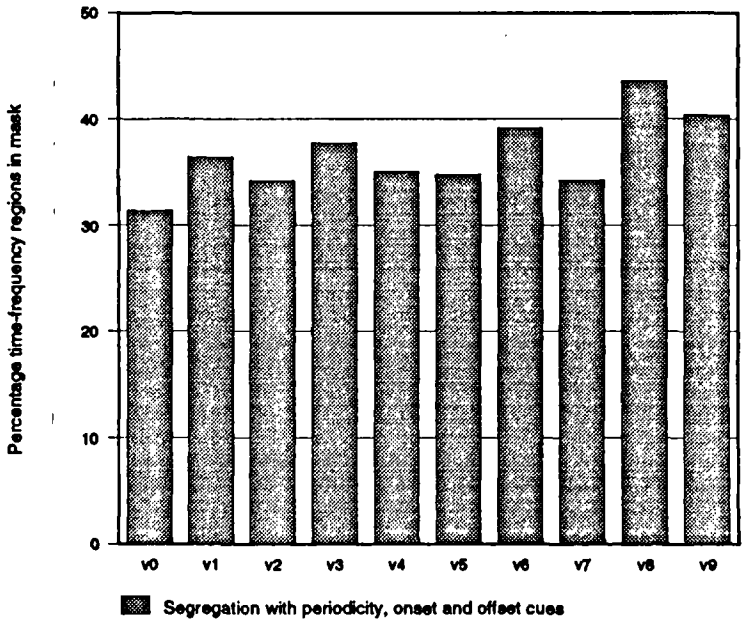

Figure 6.11: Percentage of TFRs recruited by the model when no intrusions are present.

6.11 confirms that the utterances are highly intelligible and quite natural (see the appendix).

Comparison With A Conventional Autocorrelation Strategy

It is instructive to compare the performance of the model with that of a conventional frame-based autocorrelation segregation strategy. Here, a strategy similar to the one proposed by Meddis and Hewitt [169] has been used. Initially, pitch contours were derived for each of the 10 voiced utterances. This was achieved by computing a summary autocorrelation representation for the clean speech, and identifying the location of the largest peak in each time frame (see figure 4.5). Where necessary, sub-octave errors were manually corrected. Subsequently, these pitch contours were used to inform the segregation of the utterances from the noise intrusions. As suggested by Meddis and Hewitt, an autocorrelation map was computed at each time frame, and channels of the map which had a peak at the given pitch period were allocated to the speech source.

Clearly, this approach gives the autocorrelation strategy an unfair advantage in the comparison, since it has *a priori* knowledge of the pitch period of the speech at each time frame. Normally, the pitch periods of the two sources would have to be estimated from the summary autocorrelation function of the mixture. The results here assume that this difficult task has been performed without any errors. As such, the results represent the *optimum* performance of a frame-based autocorrelation segregation strategy on the test set.

Figure 6.12 shows the mean value of $snr[t]$ after segregation, for the model and the autocorrelation strategy. The performance of the model is better for every intrusion except n9, for which it is the same. In the majority of conditions, the model also recruits more TFRs than the autocorrelation strategy (see figure 6.13). Generally, therefore, the model is able to recover more of the speech source from the auditory scene. Undoubtedly, the poorer performance of the autocorrelation strategy arises from the fact that frame-based schemes do not exploit temporal continuity.

Note that in two conditions (n1 and n8), the autocorrelation strategy *degrades* the mean $snr[t]$ after segregation. This might be expected for the random noise intrusion (n1), since it causes many peaks to occur in the channel autocorrelation functions. If a spurious peak in a channel dominated by the noise intrusion coincides with the pitch period of the speech source, the channel will be inappropriately grouped. The problem of overlapping harmonics may contribute to the poor performance on n8, the male speech intrusion (see page 110). From the pitch tracks of the speech intrusions and the voiced utterances, the frequencies of the first 10 harmonics were calculated at each time frame and compared for overlap. This informal analysis suggests that, on average, overlapping harmonics occur more frequently for condition n8 (5.3% of time frames) than for n7 (1.3%) or n9 (2.5%).

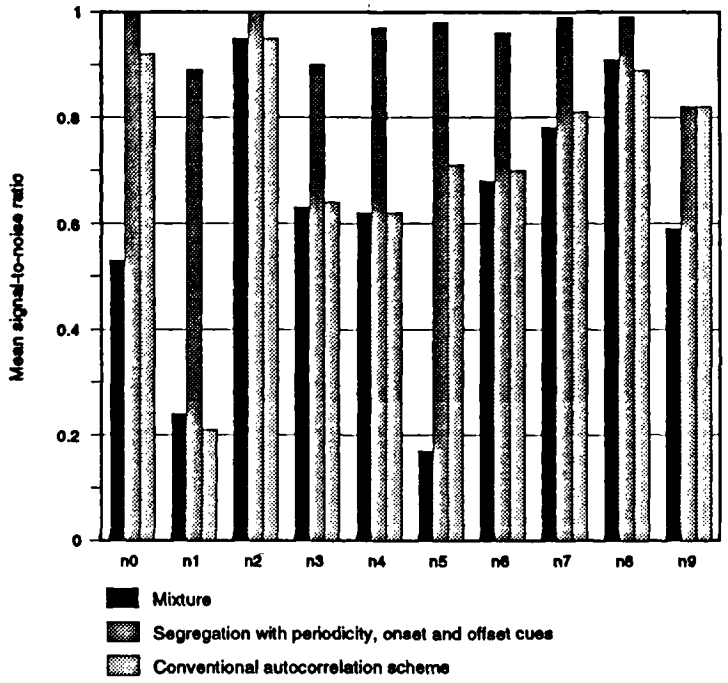

Figure 6.12: Comparison of SNR before and after segregation by the model using periodicity, onset and offset grouping cues, and by a conventional autocorrelation scheme.

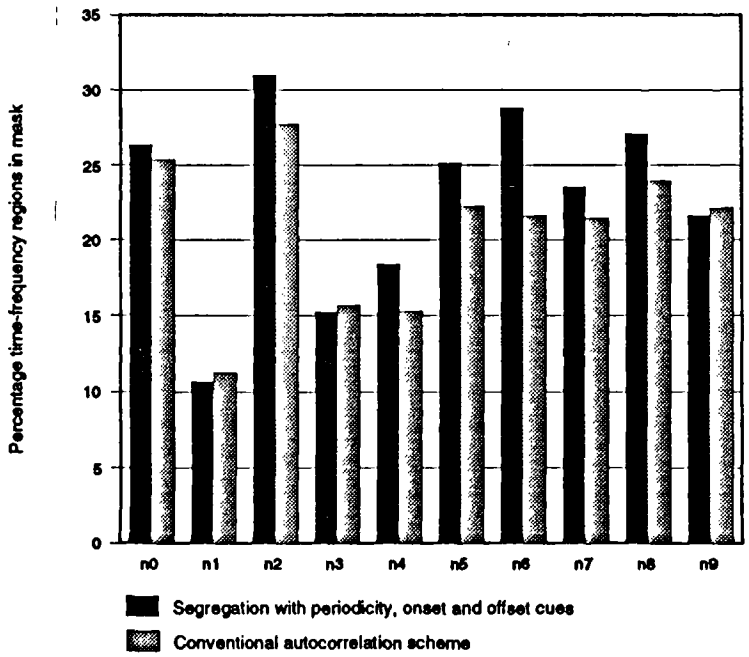

Figure 6.13: Comparison of TFRs recruited by the model using periodicity, onset and offset grouping cues, and by a conventional autocorrelation scheme.

6.3 Summary and Discussion

This chapter has presented methodologies for qualitative and quantitative evaluation of sound source segregation in the model. A novel resynthesis technique has been described, which allows the proportions of signal and noise in a mixture to be quantified before and after segregation. The results of an evaluation on a test set of 100 mixtures are very encouraging. In every condition, segregation by the model improves the SNR of the original mixture, and performance is significantly better than random. Additionally, the performance of the model is at least as good as a conventional autocorrelation segregation strategy, and is substantially better for some intrusions.

A possible limitation of the model arises from the fact that objects are exclusively allocated to one group. When an intrusion is removed from the auditory scene, it leaves a "gap" in the spectrum which can often be heard in the resynthesized waveform. One solution to this problem may be to share the energy in an auditory filter channel between sources (Weintraub [277]). Alternatively, an extrapolation technique could be used to complete the missing parts of the spectrum. For speech, a model of the vocal tract could provide estimates of the energy in spectral gaps.

The evaluation in this chapter uses the same database of 100 mixtures employed by Cooke [52]. Although no formal comparison between his results and those presented here is attempted, some informal observations can be made. His "results for the positive evidence metric" (page 114) are roughly equivalent to the SNR evaluation in figure 6.7. Generally, the two figures show a similar pattern of performance, although the model here may have an advantage for the music (n4) and speech (n7 and n8) conditions. A formal comparison would be required to confirm this. Since Cooke's model uses the same linear filterbank employed here, the resynthesis technique described in section 6.2.2 could be applied to his auditory representation. Hence, a direct comparison of the two models may be possible.

The performance of the model is generally much better than that of a conventional autocorrelation segregation scheme. Several factors may contribute to this result. Firstly, temporal continuity is made explicit in the model, whereas conventional autocorrelation strategies operate on a frame-by-frame basis. Secondly, the strategy used here is tolerant of variations in the position of peaks in the channel autocorrelation functions (see page 113). Hence, the groups found by the model tend to be larger (see figure 6.13). Finally, the model is able to solve the problem of overlapping harmonics on many occasions. The results in figure 6.12 suggest that this may give the model a significant advantage over other autocorrelation-based approaches.

It may be possible to isolate some of these factors, in order to assess their contribution to the overall performance of the model. For example, the segregation strategy described in section 5.2.2 could be implemented in a frame-based manner. This would allow the contribution of temporal continuity constraints to be quantified.

Given that the model only implements three grouping principles (common peri-

odicity, common onset and common offset), the results of the evaluation are very encouraging. The inclusion of other grouping cues may provide further increases in performance. However, it is likely that large improvements in performance will only be obtained by incorporating schema-based mechanisms (see section 7.5).

Chapter 7

Summary and Conclusions

7.1 Summary of the Model

A model of auditory processing has been described, which uses information derived from physiologically-motivated representations to group acoustic components that are likely to belong to the same sound source. The model consists of four processing stages. Firstly, the auditory periphery is simulated by a bank of bandpass filters and a model of inner hair cell function. In the second stage, a number of *auditory map* representations make particular properties of auditory nerve firing patterns explicit. Periodicities in the auditory nerve are identified by an *autocorrelation map*. Channels of the autocorrelation map which are responding to the same spectral dominance, and therefore have a similar pattern of response, are combined into *periodicity groups*. The times at which acoustic components start and stop are identified by an *onset map* and an *offset map*. Information about spectral continuity is extracted by a *frequency transition map*, which measures the orientation of spectral dominances. In the third processing stage, periodicity groups are tracked across time, using frequency transition information, and concatenated to form *auditory objects*. Finally, auditory objects which have similar properties are grouped together. Specifically, objects are likely to form a group if they have a similar pitch contour, onset time or offset time.

The model has been evaluated using the task of segregating speech from a variety of different noise intrusions. Performance has been assessed qualitatively by resynthesis, and quantitatively by comparison of the signal-to-noise ratio (SNR) before and after segregation. An improvement in SNR is obtained after segregation for each noise condition. Additionally, the model performs significantly better than a conventional frame-based autocorrelation segregation strategy.

Summaries of the model can also be found in [36] and [37].

7.2 Original Contributions

The title of this thesis emphasizes a *representational* approach to modelling auditory processing. Perhaps the most important characteristic of the model is its physiologically-principled, multi-representational view of auditory function. Physiological studies indicate that important acoustic parameters appear to be place-coded in the higher auditory system, within orderly arrays of neurons called *auditory maps.* Here, computational models of auditory maps have been employed to provide a rich representational description of the auditory scene. Specifically, the maps extract information about onsets, offsets, frequency transitions and periodicities in different spectral regions.

This approach is similar in concept to the computational approach to vision described by Marr (see section 1.2.1). Marr [159] suggested that the first stage in the description of a visual image should be a rich representation of intensity-level changes, which he called the *primal sketch.* In subsequent stages, a number of processes operate on the primal sketch to identify more abstract levels of structure. Similarly, the auditory maps employed here provide a primitive, but rich, representation of the auditory scene. These primitives form the basis for deriving abstract time-frequency objects, which can be searched rapidly and effectively. Hence, auditory maps play a central role in bridging the gap between an acoustic waveform and its description as a collection of symbolic auditory objects.

Although auditory maps have been specifically applied to scene analysis in this thesis, they undoubtedly have a more general utility. For example, parameters such as onset time and frequency transition are important cues for distinguishing speech sounds (Stevens [259]). Hence, auditory maps could provide useful primitives for an automatic speech recognition system. Indeed, the maps could be employed directly in the two-stage system described by Green *et al.* [102], which uses knowledge of acoustic features (for example, formant transitions) to refine the results of a conventional hidden Markov model recognizer. Additionally, auditory maps might provide a novel means of visualizing sound. In particular, an animated display of activity in the frequency transition map could provide a useful visual representation of spectral variation, as shown in figure 4.30. Similar animations of autocorrelation maps have recently been described by Slaney and Lyon [252].

Section 6.2.2 presents a new scheme for quantitative evaluation of sound segregation systems, which allows a highly intuitive SNR metric to be computed before and after separation. The technique can be applied to any model from which a linear resynthesis path is available. Hence, the results presented here can be compared directly with those of other workers.

Another novel aspect of the model is the way in which it incorporates spectral continuity constraints into an autocorrelation-based segregation scheme. Additionally, the model exploits redundancy in the autocorrelation map at an early stage, by grouping adjacent channels that have a similar pattern of response. In contrast,

other autocorrelation-based schemes treat each channel independently throughout the grouping process. Further, the segregation strategy used here is able to solve the problem of overlapping harmonics in many situations.

The ability of a conventional frame-based autocorrelation strategy to segregate speech from a variety of noise intrusions has been quantified for the first time in chapter 6. Additionally, the performance of a conventional autocorrelation scheme has been compared with that of the model presented here. The results of this comparison suggest that the use of spectral continuity constraints in the model gives it a significant advantage over frame-based approaches.

Finally, a distinctive feature of the model is that it relies heavily on periodicity information. The auditory object representation does not contain any information about average firing rates in the auditory nerve. Similarly, objects are grouped principally by common periodicity, although average rate is used to identify objects with a common onset or offset. Hence, the fine time structure of auditory nerve firings appears to carry a wealth of information for source segregation. Note, however, that average rate must be used at frequencies above 4-5 kHz, since phase-locking in the auditory nerve is abolished.

7.3 Limitations of the Model

The model has a number of possible limitations, which have previously been discussed and are briefly reviewed below. Future developments of the model will address many of these issues (see section 7.5).

The gammatone filters employed in the peripheral auditory model are linear. In contrast, auditory filters are known to be nonlinear, with tuning curves that broaden at high intensities. Note that a nonlinear filterbank could be used in the model, but it would not allow a comparison of signal-to-noise ratios before and after segregation (see section 6.2.2).

Autocorrelation-based approaches to periodicity detection have been questioned on the grounds of their physiological plausibility. In particular, there is no direct evidence for the system of delay lines proposed by Licklider [152].

The movement of a spectral dominance can generate "phantom" activity in the onset and offset maps. This is a minor problem, which does not affect the scene analysis strategy.

In the model, auditory objects are segregated by default, and fused only when there is good evidence for doing so. However, there is some evidence that fusion is the default condition of organization, rather than segregation.

The principle of exclusive allocation is rigidly applied in the model, so that an object can only belong to one group. In fact, there are many situations in which perceptual grouping mechanisms violate the principle of exclusive allocation. Additionally, the

allocation of an object to a single group leaves a "gap" in the spectrum, which may be audible in a resynthesized waveform.

Once formed, auditory objects cannot be split across time or frequency. This is quite a serious limitation, which prevents the model from explaining certain experimental findings.

No sequential grouping is implemented in the model. For example, the model is unable to group a sequence of voiced and unvoiced speech sounds that have arisen from a single speaker.

Currently, only primitive (unlearned) grouping principles are employed in the model. However, the auditory system is also able to use schema-based (learned) grouping principles.

An arbitrary threshold is used to determine whether objects should form a group. This is not a serious limitation, since qualitatively similar groups are found by the model over a wide range of threshold values.

The auditory system uses binaural cues, such as timing and intensity differences between the two ears, to group sounds that originate from the same spatial location. No attempt has been made to model binaural auditory processing in the current system.

7.4 Other Models

Weintraub

Weintraub [277] describes a model of auditory processing which attempts to segregate the voices of two simultaneous speakers. His system consists of three main processing stages. Firstly, the pitch period of each voice is determined. Secondly, a Markov model is used to indicate how many voices are active, and whether they are periodic or nonperiodic. Finally, an iterative algorithm estimates the amplitude spectrum of each voice.

The first stage of processing employs the cochlear model proposed by Lyon [156]. A *coincidence function* is computed at the output of each auditory filter, giving a representation that is similar to the autocorrelation map used here. Subsequently, the coincidence functions are smoothed and averaged across all channels of the filterbank. The resulting representation is equivalent to the summary autocorrelation function described in section 4.2.4. Then, a dynamic programming algorithm is used to track the peaks in the average coincidence function across time. The dominant pitch period is tracked first, followed by the weaker pitch period.

In the second stage of Weintraub's system, the number of active sources and their characteristics are determined by a pair of Markov models. The Markov model for a particular voice can be in one of seven states, corresponding to silence, periodic, non-

periodic, onset, offset, increasing periodicity and decreasing periodicity. Transition probabilities were obtained by training on a database of hand-labelled utterances.

The last stage of processing estimates the amplitude spectrum of each source, given the current state of its Markov model. An initial spectral estimate is obtained from pre-computed histograms, which relate the height of the coincidence function at the pitch period of a voice to the voice's actual spectral amplitude. Subsequently, this estimate is iteratively refined using local spectral continuity constraints.

Weintraub evaluates his system very thoroughly. The accuracy of the pitch tracker, Markov model and spectral estimation algorithm are independently assessed. Additionally, he resynthesizes a waveform for each voice, and assess its intelligibility using an automatic speech recognizer. For comparison, the recognition rates for the original mixtures of two speakers are also determined. However, his results are rather inconclusive. The system gives a small improvement in recognition rate for male speakers, but *degrades* the recognition rate for female speakers.

There are some similarities between Weintraub's model and the one presented here. Both systems use autocorrelation-like techniques for periodicity detection, and employ dynamic programming to track pitch periods across time. However, the two approaches also have substantial differences, which are discussed below.

Firstly, Weintraub's model attempts to track the *global* pitch of each source (see section 5.2.1). Additionally, his system does not exploit spectral continuity constraints during the pitch tracking process, so that it faces the difficult problem of determining which pitch period belongs to which source at every time frame. The solution adopted in his model is to assign the dominant pitch period to the voice with the closest average pitch. Hence, Weintraub's system is limited to segregating speakers with different average pitch periods (a male and a female speaker are used in his evaluation). He acknowledges the problems associated with this approach:

> "Even if one could precisely determine both pitch periods, one still has to determine which sound stream these pitch periods belong to. When both talkers have the same average pitch period, determining which pitch period belongs to which talker is a difficult problem which has no straightforward solution." (page 139)

In contrast, the model described here computes a *local* pitch contour for each time-frequency object in the auditory scene, and groups objects which have a similar local pitch. Hence, spectral continuity constraints are exploited during the pitch tracking process, and the problem of allocating pitch periods to sources does not arise.

A second criticism of Weintraub's system is that it requires training on a particular task. For example, the Markov model must be trained to distinguish periodic and nonperiodic speech. Other sounds, such as a tonal intrusion, give rise to different patterns of periodicity and would require separate training (see figure 5.6). Similarly, the spectral estimation algorithm requires information that is pre-computed from the isolated utterances of each speaker. In contrast, the model described here finds

organization in the auditory scene independently of the number of sources present, and their characteristics.

To conclude, Weintraub's system has two main disadvantages with respect to the model described here. Firstly, it is principally a frame-based scheme which fails to make full use of temporal continuity constraints. Secondly, it has been designed to solve the task of segregating two concurrent speakers, and does not generalize easily to different types, or different numbers, of acoustic sources.

Cooke

Recently, Cooke [52] has described a model of auditory processing that is similar in concept to the one presented here. The model consists of two stages. Firstly, the auditory scene is characterized as a collection of time-frequency objects. Secondly, the objects are searched for coherent organization, and objects with similar properties are fused into a single group.

In the first stage of processing, mechanical filtering in the auditory periphery is simulated by a bank of gammatone filters. Subsequently, the frequency of the most dominant component in the output of each filter is computed, using median-smoothed instantaneous frequency. Regions of the filterbank that are responding to the same dominant frequency are combined into *place groups*. Then, place groups are aggregated over time, using a technique that is similar to the birth-death process described in section 5.1.2. This aggregation process produces an explicit time-frequency representation of synchrony in the auditory filterbank, which Cooke calls *synchrony strands*.

The second stage of Cooke's model employs a two-pass strategy for searching the auditory scene. In the first pass, a synchrony strand is selected as the seed for a search. Subsequently, interleaved stages of simultaneous and sequential propagation recruit similar strands to the group containing the seed. The similarity of strands is assessed by their harmonicity (using a harmonic sieve) and common amplitude modulation (using envelope repetition rate and instantaneous amplitude fluctuations). This process continues with a new seed until every synchrony strand in the auditory scene has been allocated to a group. A second pass then combines related groups of strands. Small groups that form a subset of larger groups are subsumed, and groups with a similar derived pitch contour are fused.

Cooke evaluates his model on the test set of 100 mixtures of speech and noise that are used here. He employs a technique that determines which synchrony strands in a mixture are likely to belong to one source, and which are likely to belong to another. This approach allows a number of intuitive metrics to be computed, such as the percentage of intrusion that remains in a particular group. Cooke concludes that his grouping rules have a significant effect in accurately combining the strands which belong to the speech source.

The similarities and differences between Cooke's model and the one presented here

are now considered. Firstly, both systems remove redundancy in the auditory nerve response at an early stage of processing. In Cooke's model, auditory filters that are responding to the same spectral dominance are identified by comparison of their instantaneous frequencies. Here, channels of the autocorrelation map which have a similar pattern of response are identified by a cross-correlation algorithm. The latter approach has the advantage of being physiologically plausible (see section 4.2.8).

In both models, regions of spectral dominance are tracked across time using a trajectory principle. However, the two systems obtain trajectory information by different methods. Cooke's model estimates the trajectory of a synchrony strand by measuring the derivative of its frequency. Here, a map of frequency transition provides information about the orientation of spectral peaks.

Another difference concerns the grouping cues that are used by the two models. Cooke's system employs harmonicity and common amplitude modulation to identify objects that belong together. In contrast, the model described here uses common periodicity, common onset and common offset. Additionally, the second pass of Cooke's scene analysis strategy combines groups that have a common pitch contour. Here, common pitch contour is used to combine single objects during a one-pass grouping process. Also, note that pitch contours derived from the autocorrelation map are generally very smooth (see figure 5.16). In comparison, the pitch contours that Cooke's model derives from amplitude modulation information tend to be quite erratic (figure 5.8 of his thesis).

A related point is that Cooke's model uses different cues to group low and high frequency components. Resolved components in low frequency regions are grouped by a harmonic sieve, whereas unresolved components in high frequency regions are grouped by common amplitude modulation. Cooke's scene analysis strategy has to perform a second pass in order to combine these two types of organization. In contrast, the model described here groups acoustic components by common periodicity, regardless of whether they are resolved or unresolved. Hence, the auditory scene can be searched efficiently in a single pass.

Auditory objects in the model described here are assigned exclusively to a single group. In contrast, Cooke's system allows the same object to belong to many groups. Whether the latter approach has any advantages is a question for future research. Certainly, rigid application of the principle of exclusive allocation allows the auditory scene to be searched very efficiently. In Cooke's model, multiple allocation of objects produces many redundant groups, which need to be rationalized by a further stage of processing.

Cooke evaluates his system thoroughly, but it is difficult to compare his results with those of other workers. Although his performance metrics are intuitive, they are rather specific to the synchrony strand representation. Here, a technique for quantitative evaluation has been described, which can be applied to any model that has a linear resynthesis path.

Finally, both systems are influenced by the computational approach to vision de-

scribed by Marr. However, the model proposed here builds a multi-representational description of the auditory scene, which is closer to Marr's idea of a "primal sketch" than Cooke's synchrony strands. Effectively, Cooke's model takes a single leap between peripheral auditory activity and symbolic objects. Here, physiologically-principled auditory maps act as intermediate representations, which provide a rich database of information for the subsequent formation of auditory objects.

Mellinger

In a recent thesis, Mellinger [170] describes an auditory model that is tuned specifically for the segregation of musical sounds. Initially, Mellinger's system uses Lyon's [156] cochlear model to provide a simulation of auditory nerve firing activity. Subsequently, a number of *feature maps* operate on the auditory nerve representation to extract information about onsets and frequency variation. Finally, an algorithm tracks spectral peaks across time, and groups acoustic events that have a common onset time or common frequency modulation.

The feature maps that Mellinger employs are similar in concept to the auditory maps used here. Onsets are identified by convolving each auditory filter channel with a bipolar operator, which effectively smooths and differentiates the filter output. A technique for identifying offsets is also described, although Mellinger's segregation algorithm does not use offset information. The frequency variation of spectral dominances is determined by two methods. Firstly, a map in which each receptive field is tuned to a particular orientation measures the direction of movement of spectral peaks. This technique is similar to the frequency transition map described in section 4.4.3, although the form of the receptive fields is different. A second method of extracting frequency variation information attempts to track the movement of peaks in an autocorrelation map. Sounds that vary in frequency produce vertical and horizontal motion in the autocorrelation map, corresponding to changes in frequency and pitch period respectively. Mellinger describes an algorithm for following these changes, although this method of detecting frequency variation is not used in his segregation strategy.

In the last stage of Mellinger's model, peaks in the auditory nerve representation are tracked across time. Spectral components that generate activity in the onset map at the same time are allocated to the same source. Additionally, spectral components may be grouped if their pattern of frequency variation is similar.

Mellinger evaluates his model in a rather informal manner, by visually examining the groups that are found in short pieces of music. The model appears to group the frequency components of a single instrument reasonably well. However, it performs poorly when two or more instruments are active at the same time. Mellinger concludes that

> "The model presented here fails to work in some musically significant situations." (page 183)

Although the overall structure of Mellinger's system is similar to the model presented here, the two approaches differ in many important respects. Firstly, the cues used in Mellinger's system are common onset and common frequency modulation. Here, they are common periodicity, common onset and common offset. Curiously, Mellinger observes that

> "harmonicity...is vitally important in musical sound, since most music consists of pitched notes." (page 190)

and yet his system makes no use of harmonic relations. As a result, his model is unable to group frequency components that are slightly asynchronous and do not exhibit frequency modulation. This problem occurs with piano music, for example.

Another difference between the two models concerns the use of frequency transition information. The approach described here employs a frequency transition map to solve the temporal correspondence problem (see section 4.4.1). In contrast, Mellinger uses a similar map to provide a cue for grouping. Additionally, note that the use of common frequency modulation as a grouping cue is unsupported by psychoacoustical evidence (see section 4.5.2). A related point is that spectral peaks are tracked across time using a proximity principle in Mellinger's model, whereas a trajectory principle is used here.

Similarly, onset information is used differently in the two systems. In Mellinger's model, activity in the onset map indicates that a new acoustic event has started. Here, the start and end times of an auditory object are determined by a birth-death peak tracking process, and the onset map is used to identify objects that start synchronously. The latter approach has some advantages. For example, the onsets of some sounds (such as a bowed string) are quite gradual, and may be difficult to detect. In Mellinger's scheme, an undetected onset causes an entire event to be missed by the tracking procedure.

Finally, the performance of the model described here has been quantified using an intuitive SNR metric. In contrast, Mellinger does not attempt any quantitative evaluation. Hence, it is difficult to assess the performance of his system. Additionally, Mellinger's model incorporates eight different thresholds, which are fine-tuned for a particular test sound. In the model proposed here, the application of hard thresholds has been avoided until the final grouping stage, where a single (untuned) threshold is applied.

7.5 Future Research Directions

Physiological Findings

The model described here already has a strong physiological foundation, and further physiological data can be incorporated as it becomes available. In particular, it is

likely that other auditory maps will be identified by physiological studies. For example, the onset and offset maps hypothesized in section 4.3 have yet to be detected. In practice, maps at high levels of the auditory system may be difficult to locate. Generally, neurons have complex response properties, are tuned to a number of parameters and are arranged along irregular or convoluted axes (Knudsen *et al.* [138]). Additionally, the size of a map can be very small. Maps in the visual cortex occupy less than one square millimetre of the cortical surface (Hubel and Weisel [121]). Clearly, physiological investigation of computational maps is a challenging area for future research.

Equally, physiological data is available which has not been used in the current model. For instance, stellate cells of the cochlear nucleus are known to enhance particular rates of amplitude modulation (Kim *et al.* [132]). Hence, they could provide a more principled basis for a map of periodicity than Licklider's autocorrelation scheme. Modelling studies of stellate cells have recently been described by Banks and Sachs [17] and Hewitt *et al.* [114], and several other workers have proposed models of cochlear nucleus function (Pont and Damper [208, 209], Ainsworth and Meyer [4]). Also, the physiology of auditory space maps has been described in some detail (King and Hutchings [135], Knudsen [137]). Models of these maps could be used to group sounds with a common spatial location.

Finally, a number of workers have recently described nonlinear models of cochlear filtering (Deng and Kheurallah [76], Ambikairajah and Jones [6]). The use of a nonlinear filterbank in the model would give more realistic auditory nerve responses (particularly to speech sounds) and could improve the performance of the cross-correlation algorithm described in section 4.2.6.

The Representational Correspondence Problem

In the model described here, it is assumed that the auditory system builds multiple map representations of the acoustic input. An important question is how perceptual grouping mechanisms coordinate activity between these maps, so that the representation of an acoustic component in one map can be related to its representation in another map. In analogy with the temporal correspondence problem discussed in section 4.4.1, we might refer to this task as the *representational correspondence problem*.

Consideration of the topography of auditory maps suggests a partial solution to this problem. Specifically, auditory maps are arranged within a common cochleotopic framework, so that the values of different mapped parameters at the same characteristic frequency can be easily compared. However, it is not clear how the activity in different maps is correlated over *time*. In the computer model described here, this problem does not arise because time has been made explicit. In biological systems, time is continuously varying and there will be transmission delays between the various neural representations. A possible solution to this problem is to "time-stamp" information, a technique which is used in distributed computer systems (Lamport

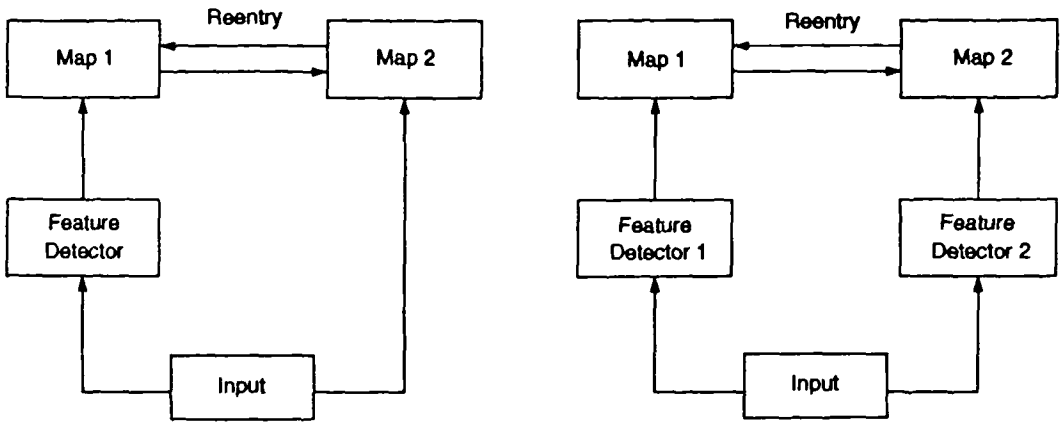

Figure 7.1: Reentry as a solution to the representational correspondence problem. On the left, the mapped output of a feature detector is continuously correlated with successive mapped inputs. On the right, the mapped outputs of two different feature detectors are continuously correlated.

[143]). However, this approach is unlikely to be used in the nervous system, since it would incur a massive informational burden and an accurate "biological clock" would be required.

An elegant solution to the representational correspondence problem has been proposed by Edelman [82, 83]. He suggests that the output of computational maps may be *reentered* onto other maps (for example, onto a map of the sensory input). This scheme allows the reentered signal to be correlated with the mapped features of the next input in a temporal sequence (see the left panel of figure 7.1). Effectively, this architecture allows a continuous spatiotemporal representation of an input object. Additionally, reentry provides a means of coordinating activity in different maps without the need for an elaborate time-stamping system (right panel of figure 7.1).

In the model described here, reentry could provide a means of combining information from different auditory maps. For example, a map could be constructed which receives reentrant inputs from periodicity and frequency transition maps. A representation of this kind could form the basis for identifying harmonically related components that are moving together in frequency.

Maps, Schemas and Learning

Currently, the model described here only employs unlearned (primitive) grouping principles. However, it is likely that a great deal of the auditory system's ability to segregate concurrent sounds is derived from learned (schema-based) processes.

Interestingly, Edelman [83] suggests that reentrant computational maps could play an important role in learning. For example, consider the network in the right panel of figure 7.1. If the two maps exhibit simultaneous activity, it is possible to link

them by increasing the strengths of their mutual reentrant connections. A linkage of this type could serve to associate certain simultaneously occurring features of a stimulus. Additionally, the redundancy of computational maps may act to reduce unreliability in this kind of distributed system (see section 2.2.3).

The above discussion suggests that auditory maps could play a role in the generation and application of learned grouping principles. Interestingly, neural networks that incorporate reentrant connections have already been applied to the problem of automatic speech recognition, with impressive results (Robinson *et al.* [223], Robinson [224]). It is tempting to suggest that reentrant networks could give similar benefits when applied to the problem of auditory scene analysis.

Other Grouping Principles

Currently, the model only implements a small number of grouping principles. In particular, the system is unable to correlate nonperiodic amplitude modulation in different spectral regions. A mechanism of this kind could underlie the phenomenon of co-modulation masking release (see section 4.5.1). Additionally, the implementation of common onset and common offset in the model is incompatible with certain experimental findings (see section 5.3). These points should be addressed during further development of the system.

Spatial location is undoubtedly an important grouping cue. A number of workers have described models of binaural auditory processing (Shamma *et al.* [248], Lyon [157]), which could be incorporated into the system proposed here. Additionally, segregation techniques which use information about spatial location have recently been described by Bodden [21] and Denbigh and Zhao [71].

The most important grouping principles that are missing from the current model are those for sequential integration, such as spectral shape, timbre, temporal proximity and fundamental frequency. Note that the map of frequency transition identified by Shamma and Chettiar [247] could provide a basis for grouping by spectral shape (see section 4.4.2).

Processing Speed

In its present form, the model requires approximately two hours of computation time for every second of acoustic input. Clearly, an enormous increase in processing speed is required before the system can be employed in practical applications, such as automatic speech recognizers and hearing prostheses.

The use of computational maps has some advantages with respect to this problem. Maps perform their computations in parallel, and could potentially run in real time on a parallel computer architecture. Additionally, the maps execute many simple, repeated operations, so it would be straightforward to implement them directly in silicon.

7.6 Summary and Conclusions

In this chapter, the model has been summarised and its limitations have been discussed. Additionally, some directions for future work have been suggested.

It is apparent from the review in section 7.4 that models of source segregation have, thus far, concentrated on unlearned (primitive) processing. Certainly, there is a limit to the performance that can be achieved by modelling only primitive mechanisms, and learned (schema-based) process must also be incorporated. Future research must address this issue if further progress is to be made in this challenging area of research.

Appendix A

Audio Demonstration Cassette

This appendix catalogues the accompanying audio tape. The tape contains five demonstrations. In the first, the segregation of speech from an interfering noise is demonstrated by resynthesis. The second demonstration compares the performance of the model and a frame-based autocorrelation scheme. The third and fourth demonstrations illustrate some properties of the resynthesis process that are discussed in section 6.2.2. Finally, a modified resynthesis scheme is demonstrated.

Demonstration One

You will hear

- Original mixture of speech and noise
- Resynthesized speech obtained after segregation by the model using periodicity, onset and offset grouping cues

for each of the speech and noise combinations v9n0, v8n1, v7n2, v8n3, v1n4, v9n5, v1n6, v3n7, v4n8 and v9n9. The characteristics of the speech and noise signals are summarized in table 6.1 and figure 6.4 respectively. Generally, the best exemplars occur with narrowband noise intrusions, such as the tone (n0) and siren (n5). The worst example occurs with the random noise intrusion (n1). However, note that the signal-to-noise ratio in this case is very low (see figure 6.6), and the speech in the original mixture is barely intelligible.

Demonstration Two

You will hear

- Original mixture of speech and noise
- Resynthesized speech obtained after segregation using a conventional frame-based autocorrelation scheme
- Resynthesized speech obtained after segregation by the model using periodicity, onset and offset grouping cues

for the mixtures of speech and noise listed above. In most cases, segregation by the scheme presented here gives greater rejection of the noise intrusion and better intelligibility of the resynthesized speech. This is particularly noticeable in examples where the temporal continuity of the noise intrusion has been exploited, such as the tone (n0) and siren (n5). See section 6.3 for a discussion of these results.

Demonstration Three

You will hear

- Original utterance v0
- Utterance v0 resynthesized with all of the time-frequency regions of the auditory filterbank included

This demonstration confirms that the resynthesis process reproduces the original signal perfectly when all of the time-frequency regions of the filterbank are included in the resynthesized waveform. See section 6.1.2 for details.

Demonstration Four

You will hear

- Resynthesized speech/residual noise obtained after segregation of mixture v0n5 using periodicity, onset and offset grouping cues. The waveform is shown in the top panel of figure 6.3.
- Sum of the proportion of speech and proportion of noise in the above example. The waveform is shown in the centre panel of figure 6.3.
- Proportion of noise in the resynthesized speech/residual noise obtained after segregation of mixture v0n5. The waveform is shown on the bottom right of figure 6.2.

- Proportion of speech in the resynthesized speech/residual noise obtained after segregation of mixture v0n5. The waveform is shown on the top right of figure 6.2.

The first two examples demonstrate that the resynthesis process is linear (see section 6.2.2). From the second two examples, it is clear that the segregation process has degraded the siren intrusion to a much greater extent than the speech signal.

Demonstration Five

This demonstration has the same format as demonstration one, except that non-selected channels of the auditory filterbank have been attenuated to 5% of their normal amplitude, rather than completely removed. As a result, the naturalness of the resynthesized speech is slightly improved. See section 6.1.2 for details.

Bibliography

[1] M. Abeles and M.H. Goldstein (1972) Responses of single units in the primary auditory cortex of the cat to tones and to tone pairs. *Brain Research*, **42**, 337-352

[2] J.C. Adams (1986) Neuronal morphology in the human cochlear nucleus. *Arch Otolaryngol Head Neck Surg*, **112**, 1253-1261

[3] A.M.J.H. Aertsen and P.I.M. Johannesma (1980) Spectro-temporal receptive fields of auditory neurons in the grassfrog. I. Characterisation of tonal and natural stimuli. *Biological Cybernetics*, **38**, 223-234

[4] W. Ainsworth and G. Meyer (1992) Speech analysis by means of a physiologically based model of the cochlear nerve and cochlear nucleus. *Proceedings of the ESCA workshop on comparing speech signal representations*, Sheffield, April 8-9th, 1-6

[5] L. Aitken (1988) The auditory midbrain. Humana Press, London

[6] E. Ambijairajah and E. Jones (1990) An active cochlear model for speech recognition. *Proceedings of the international conference on speech science and technology (SST)*, 130-135

[7] J.F. Ashmore (1987) A fast motile response in guinea-pig outer hair cells: The cellular basis of the cochlear amplifier. *Journal of Physiology (London)*, **388**, 323-347

[8] P.F. Assmann and Q. Summerfield (1990) Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies. *Journal of the Acoustical Society of America*, **88** (2), 680-697

[9] H.B. Barlow (1972) Single units and sensation: A neuron doctrine for perceptual psychology. *Perception*, **1**, 371-394

[10] J.G. Beerends and A.J.M. Houtsma (1989) Pitch identification of simultaneous diotic and dichotic two-tone complexes. *Journal of the Acoustical Society of America*, **85** (2), 813-819

[11] S.W. Beet (1990) Automatic speech recognition using a reduced auditory representation and position-tolerant discrimination. *Computer Speech and Language*, **4**, 17-33

[12] J.P.L. Blokx and S.G. Nooteboon (1982) Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, **10**, 23-36

[13] G.R. Bock, W.R. Webster and L.M. Aitkin (1972) Discharge patterns of single units in inferior colliculus of the alert cat. *Journal of Neurophysiology*, **35**, 265-277

[14] S.F. Boll (1979) Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **ASSP-27** (2), 113-120

[15] E. de Boer and H.R. de Jongh (1978) On cochlear encoding: potentialities and limitations of the reverse-correlation technique. *Journal of the Acoustical Society of America*, **63** (1), 115-135

[16] E. de Boer and P. Kuyper (1968) Triggered correlation. *IEEE Transactions on Biomedical Engineering*, **15** (3), 169-179

[17] M.I. Banks and M.B. Sachs (1991) Regularity analysis in a compartmental model of chopper units in the anteroventral cochlear nucleus. *Journal of Neurophysiology*, **65**, 606-629

[18] M.W. Beauvois and R. Meddis (1991) A computer model of auditory stream segregation. *Quarterly Journal of Experimental Psychology*, **43A** (3), 517-541

[19] C.C. Blackburn and M.B. Sachs (1989) Classification of unit types in the anteroventral cochlear nucleus: PST histograms and regularity analysis. *Journal of Neurophysiology*, **62** (6), 1303-1329

[20] C.C. Blackburn and M.B. Sachs (1990) The representations of the steady-state vowel sound /e/ in the discharge patterns of cat anteroventral cochlear nucleus neurons. *Journal of Neurophysiology*, **63** (5), 1191-1212

[21] M. Bodden (1990) A concept for a cocktail-party processor. *Proceedings of the international conference on spoken language processing (ICSLP)*, 285-288

[22] T. Bojanowski and D.W.F. Schwarz (1989) Analogue signal representation in the medial superior olive of the cat. *Journal of Otolaryngology*, **18** (1), 3-9

[23] A.S. Bregman (1987) The meaning of duplex perception: Sounds as transparent objects. In *The Psychophysics of Speech Perception*, edited by M.E.H. Schouten, Martinus Nijhoff

[24] A.S. Bregman (1990) Auditory scene analysis: The perceptual organization of sound. *The MIT Press*, London

[25] A.S. Bregman (1991) Oral presentation at the Institute of Acoustics Speech Group Meeting, Sussex University, 27th February

[26] A.S. Bregman, J. Abramson, P. Doehring and C.J. Darwin (1985) Spectral integration based on common amplitude modulation. *Perception and Psychophysics*, **37**, 483-493

[27] A.S. Bregman and J. Campbell (1971) Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*, **89** (2), 244-249

[28] A.S. Bregman and G.L. Dannenbring (1973) The effect of continuity on auditory stream segregation. *Perception and Psychophysics*, **13** (2), 308-312

[29] A.S. Bregman, R. Levitan and C. Liao (1990) Fusion of auditory components: Effects of the frequency of amplitude modulation. *Perception and Psychophysics*, **47** (1), 68-73

[30] A.S. Bregman and S. Pinker (1978) Auditory streaming and the building of timbre. *Canadian Journal of Psychology*, **32** (1), 19-31

[31] A.S. Bregman and A. Rudnicky (1975) Auditory segregation: Stream or streams? *Journal of Experimental Psychology: Human Perception and Performance*, **1**, 263-267

[32] A.S. Bregman and Y. Tougas (1989) Propagation of constraints in auditory organization. *Perception and Psychophysics*, **46**, 395-396

[33] R. Britt and A. Starr (1976) Synaptic events and discharge patterns of cochlear nucleus cells. II. Frequency-modulated tones. *Journal of Neurophysiology*, **39**, 179-194

[34] D.E. Broadbent and P. Ladefoged (1957) On the fusion of sounds reaching different sense organs. *Journal of the Acoustical Society of America*, **29** (6), 708-710

[35] G.J. Brown and M.P. Cooke (1990) Modelling modulation maps in the higher auditory system. *British Journal of Audiology*, **24** (3), 196

[36] G.J. Brown and M.P. Cooke (1992) Modelling sound source segregation: A representational approach. *Canadian Acoustics*, in press

[37] G.J. Brown and M.P. Cooke (1992) A computational model of auditory scene analysis. *Proceedings of the international conference on spoken language processing (ICSLP)*, in press

[38] S. Buus (1985) Release from masking caused by envelope fluctuations. *Journal of the Acoustical Society of America*, **78** (6), 1958-1965

[39] R. Carhart, T.W. Tillman and E.S. Greetis (1968) Release from multiple maskers: Effects of interaural time disparities. *Journal of the Acoustical Society of America*, **45** (2), 411-418

[40] R. Carlson and B. Granström (1982) Towards an auditory spectrograph. In *The Representation of Speech in the Peripheral Auditory System*, edited by R. Carlson and B. Granström, 109-114

[41] R.P. Carlyon (1991) Discriminating between coherent and incoherent frequency modulation of complex tones. *Journal of the Acoustical Society of America*, **89** (1), 329-340

[42] R.P. Carlyon, L. Demany and C. Semal (1992) Detection of across-frequency differences in fundamental frequency. *Journal of the Acoustical Society of America*, **91** (1), 279-292

[43] L.H. Carney (1990) Sensitivities of cells in anteroventral cochlear nucleus of cat to spatiotemporal discharge patterns across primary afferents. *Journal of Neurophysiology*, **64** (2), 437-456

[44] L.H. Carney and T.C.T. Yin (1988) Temporal coding of resonances by low-frequency auditory nerve fibres: single-fibre responses and a population model. *Journal of Neurophysiology*, **60** (5), 1653-1677

[45] C. Chafe and D. Jaffe (1986) Source separation and note identification in polyphonic music. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1289-1292

[46] C. Chafe, D. Jaffe, K. Kashima, B. Mont-Reynaud and J. Smith (1985) Techniques for note identification in polyphonic music. *Proceedings of the International Conference on Computer Music (ICMC)*, 399-405

[47] M.H. Chalikia and A.S. Bregman (1989) The perceptual segregation of simultaneous auditory signals: Pulse train segregation and vowel segregation. *Perception and Psychophysics*, **46** (5), 487-496

[48] E.C. Cherry (1953) Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, **25** (5), 975-979

[49] V. Ciocca and A.S. Bregman (1987) Perceived continuity of gliding and steady-state tones through interrupting noise. *Perception and Psychophysics*, **42** (5), 476-484

[50] B.M. Clopton, J.A. Winfield and F.J. Flammino (1974) Tonotopic organization: Review and analysis. *Brain Research*, **76**, 1-20

[51] R.A. Cole and B. Scott (1973) Perception of temporal order in speech: The role of vowel transitions, *Canadian Journal of Psychology*, **27** (4), 441-449

[52] M.P. Cooke (1991) Modelling auditory processing and organisation. *Ph.D. Thesis*, University of Sheffield. To be published by Cambridge University Press

[53] M.P. Cooke and P.D. Green (1990) The auditory speech sketch. *Proceedings of the Institute of Acoustics*, **12** (10), 355-362

[54] L.L. Cooper and M.W. Cooper (1981) Introduction to dynamic programming. *Pergamon Press*, Elmsford, New York

[55] J.E. Cutting and B.S. Rosner (1974) Categories and boundaries in speech and music. *Perception and Psychophysics*, **16** (3), 564-570

[56] G.L. Dannenbring (1976) Perceived auditory continuity with alternately rising and falling frequency transitions. *Canadian Journal of Psychology*, **30** (2), 99-114

[57] G.L. Dannenbring and A.S. Bregman (1978) Streaming vs. fusion of sinusoidal components of complex tones. *Perception and Psychophysics*, **24** (4), 369-376

[58] C.J. Darwin (1981) Perceptual grouping of speech components differing in fundamental frequency and onset-time. *Quarterly Journal of Experimental Psychology*, **33A**, 185-207

[59] C.J. Darwin (1983) Auditory processing and speech perception. In *Attention and Performance X*, edited by D.G. Bouwhuis, Erlbaum, Hillsdale, NJ

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

[61] C.J. Darwin and C.E. Bethell-Fox (1977) Pitch continuity and speech source attribution. *Journal of Experimental Psychology: Human Perception and Performance*, **3**(4), 665-672

[62] C.J. Darwin and V. Ciocca (1992) Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component. *Journal of the Acoustical Society of America*, **91** (6), 3381-3390

[63] C.J. Darwin and R.B. Gardner (1986) Mistuning a harmonic of a vowel: Grouping and phase effects on vowel quality. *Journal of the Acoustical Society of America*, **79** (3), 838-845

[64] C.J. Darwin and R.B. Gardner (1988) Perceptual segregation of speech from concurrent sounds. In *The Psychophysics of Speech Perception*, edited by M.E.H. Schouten, Martinus Nijhoff, 112-124

[65] C.J. Darwin and N.S. Sutherland (1984) Grouping frequency components of vowels: When is a harmonic not a harmonic? *Quarterly Journal of Experimental Psychology*, **36A**, 193-208

[66] B. Delgutte (1984) Speech coding in the auditory nerve: II. Processing schemes for vowel-like sounds. *Journal of the Acoustical Society of America*, **75** (3), 879-886

[67] B. Delgutte and N.Y.S. Kiang (1984) Speech coding in the auditory nerve: I. Vowel-like sounds. *Journal of the Acoustical Society of America*, **75** (3), 866-878

[68] B. Delgutte and N.Y.S. Kiang (1984) Speech coding in the auditory nerve: III. Voiceless fricative consonants. *Journal of the Acoustical Society of America*, **75** (3), 887-896

[69] B. Delgutte and N.Y.S. Kiang (1984) Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics. *Journal of the Acoustical Society of America*, **75** (3), 897-907

[70] B. Delgutte and N.Y.S. Kiang (1984) Speech coding in the auditory nerve: V. Vowels in background noise. *Journal of the Acoustical Society of America*, **75** (3), 908-918

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

[72] L. Deng and C.D. Geisler (1987) A composite auditory model for processing speech sounds. *Journal of the Acoustical Society of America*, **82** (6), 2001-2012

[73] L. Deng and C.D. Geisler (1987) Responses of auditory-nerve fibres to nasal consonant-vowel syllables. *Journal of the Acoustical Society of America*, **82** (6), 1977-1988

[74] L. Deng, C.D. Geisler and S. Greenberg (1987) Responses of auditory-nerve fibres to multiple-tone complexes. *Journal of the Acoustical Society of America*, **82** (6), 1989-2000

[75] L. Deng, C.D. Geisler and S. Greenberg (1988) A composite model of the auditory periphery for the processing of speech. *Journal of Phonetics*, **16**, 93-108

[76] L. Deng and I. Kheurallah (1991) Dynamic formant tracking of noisy speech using temporal analysis on outputs from a nonlinear cochlear model. *IEEE Transactions on Biomedical Engineering*, submitted

[77] D.D. Dirks and D. Bower (1969) Effect of forward and backward masking on speech intelligibility. *Journal of the Acoustical Society of America*, **47** (2), 1003-1008

[78] B. Dodd (1980) Interaction of auditory and visual information in speech perception. *British Journal of Psychology*, **71**, 541-549

[79] J.M. Dolmazon (1982) Representation of speech-like sounds in the peripheral auditory system in light of a model. In *The Representation of Speech in the Peripheral Auditory System*, edited by R. Carlson and B. Granström, 151-163

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

[81] A.J. Duquesnoy and R. Plomp (1983) The effect of a hearing aid on the speech-reception thresholds of hearing-impaired listeners in quiet and in noise. *Journal of the Acoustical Society of America*, **73** (6), 2166-2173

[82] G.M. Edelman (1978) Group selection and phasic reentrant signalling: A theory of higher brain function. In *The Mindful Brain: Cortical organization and the group-selective theory of higher brain function*, by G.M. Edelman and V.B. Mountcastle, MIT Press, 51-100

[83] G.M. Edelman (1989) Neural Darwinism: The theory of neuronal group selection. *Oxford University Press*

[84] E.F. Evans and P.G. Nelson (1973) The responses of single neurons in the cochlear nucleus of the cat as a function of their location and anaesthetic state. *Experimental Brain Research*, **17**, 402-427

[85] J.M. Festen and R. Plomp (1983) Relations between auditory functions in impaired hearing. *Journal of the Acoustical Society of America*, **73** (2), 652-662

[86] R.D. Frisina, R.L. Smith and S.C. Chamberlain (1990) Encoding of amplitude modulation in the gerbil cochlear nucleus: I. A hierarchy of enhancement. *Hearing Research*, **44**, 99-122

[87] D. Gabor (1946) Theory of communication. *Journal of the Institute of Electrical Engineers*, **93**, 429-457

[88] R.B. Gardner (1989) An algorithm for separating simultaneous vowels. *British Journal of Audiology*, **23**, 170-171

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

[90] R.B. Gardner, S.A. Gaskill and C.J. Darwin (1989) Perceptual grouping of formants with static and dynamic differences in fundamental frequency. *Journal of the Acoustical Society of America*, **85** (3), 1329-1337

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

[92] O. Ghitza (1986) Speech analysis/synthesis based on matching the synthesized and the original representations in the auditory nerve level. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1995-1998

[93] O. Ghitza (1988) Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment. *Journal of Phonetics*, **16**, 109-123

[94] J.R. Glass and V.W. Zue (1988) Multi-level acoustic segmentation of continuous speech. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 429-432

[95] D.A. Godfrey, N.Y. Kiang and B.E. Norris (1975) Single unit activity in the posteroventral cochlear nucleus of the cat. *Journal of Comparitive Neurology*, **162**, 247-268

[96] J.M. Goldberg and P.B. Brown (1969) Response of binaural neurons in the dog superior olivary complex to dichotic tonal stimuli: some physiological mechanisms of sound localization. *Journal of Neurophysiology*, 613-636

[97] R. Goldhor (1983) A speech signal processing system based on a peripheral auditory model. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1368-1371

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

[99] T. Gramss and H.W. Strube (1990) Recognition of isolated words based on psychoacoustics and neurobiology. *Speech Communication*, **9**, 35-40

[100] P.D. Green, G.J. Brown, M.P. Cooke, M.D. Crawford and A.J.H. Simons (1990) Bridging the gap between signals and symbols in speech recognition. In *Advances in Speech, Hearing and Language Processing Volume 1*, edited by W. Ainsworth, 149-192

[101] G.G.R. Green and R.H. Kay (1973) The adequate stimuli for channels in the human auditory pathways concerned with the modulation present in frequency-modulated tones. *Journal of Physiology (London)*, **234**, 50-52P

[102] P.D. Green, A.J.H. Simons and P.J. Roach (1990) The SYLK project: Foundations and overview. *Proceedings of the Institute of Acoustics*, **12** (10), 249-258

[103] P.D. Green and A.R. Wood (1986) A representational approach to knowledge-based acoustic-phonetic processing in speech recognition. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1205-1208

[104] J.H. Grose and J.W. Hall (1989) Comodulation masking release using SAM tonal complex maskers: Effects of modulation depth and signal position. *Journal of the Acoustical Society of America*, **85** (3), 1276-1284

[105] C.M. Hackney (1987) Anatomical features of the auditory pathway from cochlea to cortex. *British Medical Bulletin*, **43** (4), 780-801

[106] J.W. Hall and J.H. Grose (1988) Comodulation masking release: Evidence for multiple cues. *Journal of the Acoustical Society of America*, **84** (5), 1669-1675

[107] J.W. Hall, J.H. Grose and M.P. Haggard (1988) Comodulation masking release for multicomponent signals. *Journal of the Acoustical Society of America*, **83** (2), 677-686

[108] J.W. Hall, M.P. Haggard and M.A. Fernandes (1984) Detection in noise by spectro-temporal pattern analysis. *Journal of the Acoustical Society of America*, **76** (1), 50-56

[109] B.A. Hanson and D.Y. Wong (1984) The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence of interfering speech. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 18A.5.1-18A.5.4

[110] B.A. Hanson, D.Y. Wong and B.H. Juang (1983) Speech enhancement with harmonic synthesis. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1122-1125

[111] W.A. Harris (1986) Learned topography: the eye instructs the ear. *Trends in the Neurosciences (TINS)*, March, 97-99

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

[113] M.J. Hewitt and R. Meddis (1991) An evaluation of eight computer models of mammalian inner hair-cell function. *Journal of the Acoustical Society of America*, **90** (2), 904-917

[114] M.J. Hewitt, R. Meddis and T.M. Shackleton (1992) A computer model of a cochlear-nucleus stellate cell: Responses to amplitude-modulated and pure-tone stimuli. *Journal of the Acoustical Society of America*, **91** (4), 2096-2109

[115] J.A. Hirsch, J.C.K. Chan, T.C.T. Yin (1985) Responses of neurons in the cat's superior colliculus to acoustic stimuli. I. Monaural and binaural response properties. *Journal of Neurophysiology*, **53**, 726-745

[116] J. Holdsworth, I. Nimmo-Smith, R. Patterson and P. Rice (1988) Implementing a gammatone filterbank. In *Auditory/Connectionist Techniques for Speech*, Applied Psychology Unit, Cambridge University

[117] J. Holdsworth, J.L. Schwartz, F. Berthommier and R.D. Patterson (1992) A multi-representation model for auditory processing of sounds. In *Advances in the Biosciences Volume 83*, edited by Y. Cazals, L. Demany and K. Homer, Pergamon Press, 447-453

[118] J. Horikawa and K. Murata (1988) Spatial distribution of response latencies in the rat inferior colliculus. *Proceedings of the Japan Academy Series B*, **64**, 181-184

[119] A.J.M. Houtsma and J. Smurzynski (1990) Pitch identification and discrimination for complex tones with many harmonics. *Journal of the Acoustical Society of America*, **87** (1), 304-310

[120] D.H. Hubel and T.N. Wiesel (1968) Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology (London)*, **195**, 215-243

[121] D.H. Hubel and T.N. Wiesel (1977) Functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London Series B*, **198**, 1-59

[122] R.W. Hukin and R.I. Damper (1989) Testing an auditory model by resynthesis. *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, **1**, 243-246

[123] M.J. Hunt and C. Lefebvre (1988) Speaker dependent and independent speech recognition with an auditory model. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 215-218

[124] D.R.F. Irvine (1986) Progress in sensory physiology 7: The auditory brainstem. Springer-Verlag, Berlin

[125] R.L. Jenison, S. Greenberg, K.R. Kluender and W.S. Rhode (1991) A composite model of the auditory periphery for the processing of speech based on the filter response functions of single auditory nerve fibres. *Journal of the Acoustical Society of America*, **90** (2), 773-786

[126] G. Johansson (1964) Perception of motion and changing form. *Scandinavian Journal of Psychology*, **5**, 181-208

[127] M.R. Jones (1976) Time, our lost dimension: Toward a new theory of perception, attention, and memory. *Psychological Review*, **83** (5), 323-355

[128] R.H. Kay and D.R. Matthews (1972) On the existence in human auditory pathways of channels selectively tuned to the modulation present in frequency-modulated tones. *Journal of Physiology (London)*, **225**, 657-677

[129] G.S. Kendall (1991) Visualization by ear: Auditory imagery for scientific visualization and virtual reality. *Computer Music Journal*, **15** (4), 70-73

[130] N.Y.S. Kiang, T. Watanabe, E.C. Thomas and L.F. Clark (1965) Discharge patterns of single fibres in the cat's auditory nerve. MIT Press, Cambridge

[131] D.O. Kim and G. Leonard (1988) Pitch-period following response of cat cochlear nucleus neurons to speech sounds. In *Basic Issues in Hearing*, edited by H. Duifhuis, J.W. Horst and H.P. Wit, 252-260

[132] D.O. Kim, J.G. Sirianni and S.O. Chang (1990) Responses of DCN-PVCN neurons and auditory nerve fibres in unanesthetized decerebrate cats to AM and pure tones: Analysis with autocorrelation/power-spectrum. *Hearing Research*, **45**, 95-113

[133] A.J. King and D.R. Moore (1991) Plasticity of auditory maps in the brain. *Trends in the Neurosciences (TINS)*, **14** (1), 31-37

[134] A.J. King and A.R. Palmer (1983) Cells responsive to free field auditory stimuli in guinea-pig superior colliculus: Distribution and response properties. *Journal of Physiology (London)*, **342**, 361-381

[135] A.J. King and M.E. Hutchings (1987) Spatial response properties of acoustically responsive neurons in the superior colliculus of the ferret: A map of auditory space. *Journal of Neurophysiology*, **57** (2), 596-624

[136] K.R. Kluender and R.L. Jenison (1992) Effects of glide slope, noise intensity, and noise duration on the extrapolation of FM glides through noise. *Perception and Psychophysics*, **51** (3), 231-238

[137] E.I. Knudsen (1981) The hearing of the barn owl. *Scientific American*, **245** (6), 82-91

[138] E.I. Knudsen, S. duLac and S.D. Esterly (1982) Computational maps in the brain. *Annual Review of Neuroscience*, **10**, 41-65

[139] E.I. Knudsen and M. Konishi (1978) A neural map of auditory space in the owl. *Science*, **200**, 795-797

[140] K. Koffka (1936) Principles of Gestalt psychology. *Harcout and Brace*, New York

[141] M. Kubovy, J.E. Cutting and R.M. McGuire (1974) Hearing with the third ear: Dichotic perception of a melody without monaural familiarity cues. *Science*, **186**, 272-274

[142] G.M. Kuhn (1975) On the front cavity resonance and its possible role in speech perception. *Journal of the Acoustical Society of America*, **58** (2), 428-433

[143] L. Lamport (1978) Time, clocks, and the ordering of events in a distributed system. *Communications of the the ACM*, **21** (7), 558-565

[144] G. Langner, C. Schreiner and M. Albert (1992) Tonotopy and periodotopy in the auditory midbrain of cat and Guinea fowl. In *Advances in the Biosciences Volume 83*, edited by Y. Cazals, L. Demany and K. Homer, Pergamon Press, 241-248

[145] G. Langner, C. Schreiner and M.M. Merzenich (1987) Covariation of latency and temporal resolution in the inferior colliculus of the cat. *Hearing Research*, **31**, 197-202

[146] I. Lehiste and G.E. Peterson (1961) Transitions, glides and diphthongs. *Journal of the Acoustical Society of America*, **33** (3), 268-277

[147] M.C. Liberman (1982) Single-neuron labeling in the cat auditory nerve. *Science*, **216**, 1239-1240

[148] A.M. Liberman (1982) On finding that speech is special. *American Psychologist*, **37** (2), 148-167

[149] A.M. Liberman, P.C. Delattre, L.J. Gerstman and F.S. Cooper (1956) Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology*, **52** (2), 127-137

[150] A.M. Liberman, F.S. Cooper, D.P. Shankweiler and M. Studdert-Kennedy (1967) Perception of the speech code. *Psychological Review*, **74** (6), 431-461

[151] A.M. Liberman, D. Isenberg, B. Rakerd (1981) Duplex perception of cues for stop consonants: Evidence for a phonetic mode. *Perception and Psychophysics*, **30** (2), 133-143

[152] J.C.R. Licklider (1951) A duplex theory of pitch perception. *Experientia*, **7** (4), 128-134

[153] J.S. Lienard (1987) Speech analysis and reconstruction using short-time, elementary waveforms. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 948-951

[154] E. Lombard (1911) Le signe de l'elevation de la voix. *Ann. Maladies Oreille, Larnx, Nez, Pharynx*, **37**, 101-119

[155] R.A. Lutfi and R.D. Patterson (1984) On the growth of masking asymmetry with stimulus intensity. *Journal of the Acoustical Society of America*, **76** (3), 739-745

[156] R.F. Lyon (1982) A computational model of filtering, detection and compression in the cochlea. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1282-1285

[157] R.F. Lyon (1988) A computational model of binaural localization and separation. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1148-1151

[158] C. von der Malsberg and J.D. Cowan (1982) Outline of a theory for the ontogenesis of iso-orientation domains in visual cortex. *Biological Cybernetics*, **45**, 49-56

[159] D. Marr (1976) Early processing of visual information. *Proceedings of the Royal Society of London Series B*, **275**, 483-519

[160] D. Marr (1982) Vision. W.H. Freeman and Company, San Francisco

[161] S. McAdams (1984) Spectral fusion, spectral parsing and the formation of auditory images. *Unpublished Ph.D. Thesis*, Stanford University

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

[163] R.J. McAulay and T.F. Quatieri (1986) Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **ASSP-34** (4), 744-754

[164] R. Meddis (1986) Simulation of mechanical to neural transduction in the auditory receptor. *Journal of the Acoustical Society of America*, **79** (3), 702-711

[165] R. Meddis (1988) Simulation of auditory-neural transduction: Further studies. *Journal of the Acoustical Society of America*, **83** (3), 1056-1063

[166] R. Meddis, M.J. Hewitt and T.M. Shackleton (1988) Implementation details of a computational model of the inner hair-cell/auditory-nerve synapse. *Journal of the Acoustical Society of America*, **87**, 1813-1816

[167] R. Meddis and M.J. Hewitt (1991) Virtual pitch and phase sensitivity of a computer model of the auditory periphery: I. phase sensitivity. *Journal of the Acoustical Society of America*, **89** (6), 2883-2894

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

[169] R. Meddis and M.J. Hewitt (1992) Modelling the identification of concurrent vowels with different fundamental frequencies. *Journal of the Acoustical Society of America*, **91** (1), 233-245

[170] D. Mellinger (1991) Event formation and separation in musical sound. *Unpublished Ph.D. Thesis*, Stanford University

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

[172] J.R. Mendelson, C.E. Schreiner, K. Grasse and M. Sutter (1988) Spatial distribution of responses to FM sweeps in cat primary auditory cortex. *Proceedings of the 11th A.R.O. Meeting*, 199-200

[173] J.C. Middlebrooks and E.I. Knudsen (1987) Changes in external ear position modify the spatial tuning of auditory units in the cat's superior colliculus. *Journal of Neurophysiology*, **57** (3), 672-687

[174] G.A. Miller and J.C.R. Licklider (1950) The intelligibility of interrupted speech. *Journal of the Acoustical Society of America*, **22** (1), 167-173

[175] G.A. Miller and W. Taylor (1948) The perception of repeated bursts of noise. *Journal of the Acoustical Society of America*, **20**, 171-182

[176] M.I. Miller and M.B. Sachs (1983) Auditory-nerve representation of voice pitch. In *Hearing - Physiological Bases and Psychophysics*, edited by R. Klinke and R. Hartmann, Springer-Verlag, Berlin

[177] A.R. Møller (1972) Coding of amplitude and frequency modulated sounds in the cochlear nucleus of the rat. *Acta Physiological Scandinavica*, **86**, 223-238

[178] D.B. Moody, D. Cole, L.M. Davidson and W.C. Stebbins (1984) Evidence for a reappraisal of the psychophysical selective adaptation paradigm. *Journal of the Acoustical Society of America*, **76** (4), 1076-1079

[179] B.C.J. Moore (1982) An introduction to the psychology of hearing. *Academic Press*, London

[180] B.C.J. Moore (1990) Co-modulation masking release: spectro-temporal pattern analysis in hearing. *British Journal of Audiology*, **24**, 131-137

[181] D.R. Moore (1987) Physiology of higher auditory system. *British Medical Bulletin*, **43** (4), 856-870

[182] J.K. Moore (1987) The human auditory brain stem: A comparative view. *Hearing Research*, **29**, 1-32

[183] B.C.J. Moore and B.R. Glasberg (1983) Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, **74** (3), 750-753

[184] B.C.J. Moore, B.R. Glasberg, T. Gaunt and T. Child (1991) Across-channel masking of changes in modulation depth for amplitude- and frequency-modulated signals. *Quarterly Journal of Experimental Psychology*, **43A** (3), 327-347

[185] B.C.J. Moore, B.R. Glasberg and R.W. Peters (1985) Relative dominance of individual partials in determining the pitch of complex tones. *Journal of the Acoustical Society of America*, **77** (5), 1853-1860

[186] B.C.J. Moore, B.R. Glasberg and G.P. Schooneveldt (1990) Across-channel masking and comodulation masking release. *Journal of the Acoustical Society of America*, **87** (4), 1683-1694

[187] B.C.J. Moore and S.M. Rosen (1979) Tune recognition with reduced pitch and interval information. *Quarterly Journal of Experimental Psychology*, **31**, 229-240

[188] R.K. Moore, A.P. Varga and M. Kadirkamanatha (1991) Automatic separation of speech and other complex sounds using hidden Markov model decomposition. *Institute of Acoustics Speech Group Meeting*, Sussex University, 27th February

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

[190] I. Nabelek and I.J. Hirsh (1969) On the discrimination of frequency transitions. *Journal of the Acoustical Society of America*, **45** (6), 1510-1519

[191] J.A. Naylor and S.F. Boll (1987) Techniques for suppression of an interfering talker in co-channel speech. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 205-208

[192] S. Nulsen, D. Landy, M. O'Kane, P.Kenne and S. Atkins (1990) Development of rules for automatic recognition of nasal consonants. *Proceedings of the international conference on speech science and technology (SST)*, 194-199

[193] L.P.A.S. van Noorden (1977) Minimum differences of level and frequency for perceptual fission of tone sequences ABAB. *Journal of the Acoustical Society of America*, **61** (4), 1041-1045

[194] D.L. Oliver and D.K. Morest (1984) The central nucleus of the inferior colliculus in the cat. *Journal of Comparitive Neurology*, **222**, 237-264

[195] Y. Oono and Y. Sujaku (1975) A model for automatic gain control observed in the firings of primary auditory neurons. *Abstracts of the Transactions of the Institute of Electronic and Communication Engineers Japan*, **58**, 61-62

[196] A.R. Palmer (1987) Physiology of the cochlear nerve and cochlear nucleus. *British Medical Bulletin*, **43** (4), 838-855

[197] A.R. Palmer and I.M. Winter (1992) Cochlear nerve and cochlear nucleus responses to the fundamental frequency of voiced speech sounds and harmonic complex tones. In *Advances in the Biosciences Volume 83*, edited by Y. Cazals, L. Demany and K. Homer, Pergamon Press, 231-239

[198] T.W. Parsons (1976) Separation of speech from interfering speech by means of harmonic selection. *Journal of the Acoustical Society of America*, **60** (4), 911-918

[199] R.D. Patterson (1987) A pulse ribbon model of monaural phase perception. *Journal of the Acoustical Society of America*, **82** (5), 1560-1586

[200] R.D. Patterson, I. Nimmo-Smith, J. Holdsworth and P. Rice (1987) An efficient auditory filterbank based on the gammatone function. *Institute of Acoustics Speech Group Meeting on Auditory Modelling*, RSRE, December 14-15

[201] R.D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang and M. Allerhand (1992) Complex sounds and auditory images. In *Advances in the Biosciences Volume 83*, edited by Y. Cazals, L. Demany and K. Homer, Pergamon Press, 429-446

[202] R. Patuzzi, P.M. Sellick and B.M. Johnstone (1984) The modulation of the sensitivity of the mammalian cochlea by low frequency tones. III. Basilar membrane motion. *Hearing Research*, **13**, 19-27

[203] R. Pfeiffer (1966) Classification of response patterns of spike discharges for units in the cochlear nucleus: Tone-burst stimulation. *Experimental Brain Research*, **1**, 220-235

[204] D.P. Phillips, S.S. Orman, A.D. Musicant and G.F. Wilson (1985) Neurons in the cat's primary auditory cortex distinguished by their responses to tones and wide-spectrum noise. *Hearing Research*, **18**, 73-86

[205] J.O. Pickles (1988) An introduction to the physiology of hearing (second edition). Academic Press, London

[206] C.J. Plack and B.C.J. Moore (1990) Temporal window shape as a function of frequency and level. *Journal of the Acoustical Society of America*, **87** (5), 2178-2187

[207] I. Pollack (1968) Detection of rate of change of auditory frequency. *Journal of Experimental Psychology*, **77** (4), 535-541

[208] M.J. Pont (1989) The role of the dorsal cochlear nucleus in the perception of voicing contrasts in initial English stop consonants: A computational modelling study. *Unpublished Ph.D. Thesis*, University of Southampton

[209] M.J. Pont and R.I. Damper (1991) A computational model of afferent neural activity from the cochlea to the dorsal acoustic stria. *Journal of the Acoustical Society of America*, **89**, 1213-1228

[210] T.C. Rand (1974) Dichotic release from masking for speech. *Journal of the Acoustical Society of America*, **55** (3), 678-680

[211] R.A. Rasch (1978) The perception of simultaneous notes such as in polyphonic music. *Acustica*, **40**, 21-33

[212] R.A. Rasch (1979) Synchronization in performed ensemble music. *Acustica*, **43**, 121-131

[213] A. Rees and R.H. Kay (1985) Delineation of FM rate channels in man by detectability of a three-component modulation waveform. *Hearing Research*, **18**, 211-221

[214] A. Rees and A.R. Møller (1987) Stimulus properties influencing the responses of inferior colliculus neurons to amplitude-modulated sounds. *Hearing Research*, **27**, 129-143

[215] D. Regan and B.W. Tansley (1979) Selective adaptation to frequency modulated tones: Evidence for an information-processing channel selectively sensitive to frequency changes. *Journal of the Acoustical Society of America*, **65** (5), 1249-1257

[216] H. Rheingold (1991) Virtual reality. *Secker and Walker*, London

[217] W.S. Rhode, D. Oertel and P.H. Smith (1983) Physiological response properties of cells labelled intracellularly with horseradish peroxidase in cat ventral cochlear nucleus. *Journal of Comparitive Neurology*, **213**, 448-463

[218] W.S. Rhode and P.H. Smith (1986) Encoding timing and intensity in the ventral cochlear nucleus of the cat. *Journal of Neurophysiology*, **56** (2), 261-286

[219] W.S. Rhode and P.H. Smith (1986) Physiological studies on neurons in the dorsal cochlear nucleus of cat. *Journal of Neurophysiology*, **56** (2), 287-307

[220] M.D. Riley (1989) Speech time-frequency representations. *Kluwer Academic Publishers*, Boston

[221] R.J. Ritsma (1967) Frequencies dominant in the perception of the pitch of complex sounds. *Journal of the Acoustical Society of America*, **42**, 191-198

[222] B. Roberts and B.C.J. Moore (1991) The influence of extraneous sounds on the perceptual estimation of first-formant frequency in vowels under conditions of asynchrony. *Journal of the Acoustical Society of America*, **89** (6), 2922-2932

[223] T. Robinson, J. Holdsworth, R. Patterson and F. Fallside (1990) A comparison of preprocessors for the cambridge recurrent error propagation network speech recognition system. *Proceedings of the international conference on spoken language processing (ICSLP)*, 1033-1036

[224] T. Robinson (1992) The state space and "ideal input" representations of recurrent networks. *Proceedings of the ESCA workshop on comparing speech signal representations*, Sheffield, April 8-9th, 361-368

[225] R. Romand (1978) Survey of intracellular recording in the cochlear nucleus of the cat. *Brain Research*, **148**, 43-65

[226] J.E. Rose, N.B. Gross, C.D. Geisler and J.E. Hind (1966) Some neural mechanisms in the inferior colliculus of the cat which may be relevant to localization of a sound source. *Journal of Neurophysiology*, **29**, 288-314

[227] J.E. Rose, J.E. Hind, D.J. Anderson and J.F. Brugge (1971) Some effects of stimulus intensity on response of auditory nerve fibres in the squirrel monkey. *Journal of Neurophysiology*, **34**, 685-699

[228] E.M. Rouiller and D.K. Ryugo (1984) Intracellular marking of physiologically characterized cells in the ventral cochlear nucleus of the cat. *Journal of Comparitive Neurology*, **225**, 167-186

[229] I.J. Russell (1987) The physiology of the organ of Corti. *British Medical Bulletin*, **43** (4), 802-820

[230] A.F. Ryan and J. Miller (1978) Single unit responses in the inferior colliculus of the awake and performing Rhesus monkey. *Experimental Brain Research*, **32**, 389-407

[231] M.B. Sachs and E.D. Young (1979) Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate. *Journal of the Acoustical Society of America*, **66** (2), 470-479

[232] M.B. Sachs and E.D. Young (1980) Effects of nonlinearities on speech encoding in the auditory nerve. *Journal of the Acoustical Society of America*, **68** (3), 858-875

[233] M.T.M. Scheffers (1983) Sifting vowels: Auditory pitch analysis and sound segregation. *Unpublished Ph.D. Thesis*, University of Gröningen

[234] M.T.M. Scheffers (1983) Simulation of auditory analysis of pitch: An elaboration on the DWS pitch meter. *Journal of the Acoustical Society of America*, **74** (6), 1716-1725

[235] P.H. Schiller, B.L. Finlay and S.F. Volman (1976) Quantitative studies of single-cell properties in monkey striate cortex. II. Orientation specificity and ocular dominance. *Journal of Neurophysiology*, **39** (6), 1320-1333

[236] G.P. Schooneveldt and B.C.J. Moore (1987) Comodulation masking release (CMR): Effects of signal frequency, flanking-band frequency, masker bandwidth, flanking-band level, and monotic versus dichotic presentation of the flanking band. *Journal of the Acoustical Society of America*, **82** (6), 1944-1956

[237] C.E. Schreiner and G. Langner (1988) Periodicity coding in the inferior colliculus of the cat. II. Topographical organization. *Journal of Neurophysiology*, **60** (6), 1823-1840

[238] C.E. Schreiner, J.R. Mendelson, K. Grasse and M. Sutter (1988) Spatial distribution of basic response properties in cat primary auditory cortex. *Proceedings of the 11th A.R.O. Meeting*, 198-199

[239] C.E. Schreiner and J.V. Urbas (1988) Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF). *Hearing Research*, **21**, 227-241

[240] M.R. Schroeder and J.L. Hall (1974) Model for mechanical to neural transduction in the auditory receptor. *Journal of the Acoustical Society of America*, **55** (5), 1055-1060

[241] H.A. Schwid and C.D. Geisler (1982) Multiple reservoir model of neurotransmitter release by a cochlear inner hair cell. *Journal of the Acoustical Society of America*, **72** (5), 1435-1440

[242] S. Seneff (1987) Vowel recognition based on line-formants derived from an auditory-based spectral representation. *DARPA review meeting*, San Diego

[243] S. Seneff (1988) A joint synchrony/mean-rate model of auditory speech processing. *Journal of Phonetics*, **16**, 55-76

[244] S. Shamma (1985) Speech processing in the auditory system I: The representation of speech sounds in the responses of the auditory nerve. *Journal of the Acoustical Society of America*, **78** (5), 1612-1621

[245] S. Shamma (1985) Speech processing in the auditory system II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve. *Journal of the Acoustical Society of America*, **78** (5), 1622-1632

[246] S. Shamma (1988) The acoustic features of speech sounds in a model of auditory processing: vowels and voiceless fricatives. *Journal of Phonetics*, **16**, 77-91

[247] S.A. Shamma and G.M. Chettiar (1991) A functional model of primary auditory cortex: spectral gradient columns. *Journal of the Acoustical Society of America*, submitted

[248] S.A. Shamma, N. Shen and P. Gopalaswamy (1989) Stereausis: Binaural processing without neural delays. *Journal of the Acoustical Society of America*, **86** (3), 989-1006

[249] S.A. Shamma, S. Vranic and P. Wiser (1992) Spectral gradient columns in primary auditory cortex: Physiological and psychoacoustical correlates. In *Advances in the Biosciences Volume 83*, edited by Y. Cazals, L. Demany and K. Homer, Pergamon Press, 397-406

[250] W.P. Shofner and E.D. Young (1985) Excitatory/inhibitory response types in the cochlear nucleus: Relationships to discharge patterns and responses to electrical stimulation of the auditory nerve. *Journal of Neurophysiology*, **54** (4), 917-939

[251] D.G. Sinex and C.D. Geisler (1983) Responses of auditory-nerve fibres to consonant-vowel syllables. *Journal of the Acoustical Society of America*, **73** (2), 602-615

[252] M. Slaney and R.F. Lyon (1990) A perceptual pitch detector. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 357-360

[253] M. Slaney and R.F. Lyon (1991) Visualizing sound with auditory correlograms. *Journal of the Acoustical Society of America*, submitted

[254] M. Slaney (1991) Release notes for version 2.2 of Lyon's cochlear model, *Apple Computer Technical Report*, Apple Computer, Inc.

[255] R.L. Smith (1979) Adaptation, saturation and physiological masking in single auditory nerve fibres. *Journal of the Acoustical Society of America*, **65**, 166-178

[256] R.L. Smith and M.L. Brachman (1982) Adaptation in auditory-nerve fibres: A revised model. *Biological Cybernetics*, **44**, 107-120

[257] G.A. Spirou and E.D. Young (1991) Organization of dorsal cochlear nucleus type IV unit response maps and their relationship to activation by bandlimited noise. *Journal of Neurophysiology*, **66** (5), 1750-1768

[258] H. Steiger and A.S. Bregman (1981) Capturing frequency components of glided tones: Frequency separation, orientation, and alignment. *Perception and Psychophysics*, **30** (5), 425-435

[259] K.N. Stevens (1980) Acoustic correlates of some phonetic categories. *Journal of the Acoustical Society of America*, **68** (3), 836-842

[260] R.J. Stubbs and Q. Summerfield (1990) Algorithms for separating the speech of interfering talkers: Evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, **87** (1), 359-372

[261] N. Suga (1965) Analysis of frequency-modulated sounds by auditory neurons of echo-locating bats. *Journal of Physiology (London)*, **179**, 26-53

[262] N. Suga (1990) Cortical computational maps for auditory imaging. *Neural Networks*, **3** (1), 3-21

[263] N. Suga and T. Manabe (1982) Neural basis of amplitude-spectrum representation in auditory cortex of the mustached bat. *Journal of Neurophysiology*, **47** (2), 225-255

[264] N. Suga, W. O'Neill and T. Manabe (1979) Harmonic-sensitive neurons in the auditory cortex of the mustache bat. *Science*, **203**, 270-274

[265] Q. Summerfield (1987) Speech perception in normal and impaired hearing. *British Medical Bulletin*, **43** (4), 909-925

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

[267] Q. Summerfield (1991) Roles of harmonicity and coherent frequency modulation in auditory grouping. *Paper presented at the Royal Society Meeting on the Auditory Processing of Complex Sounds*, London, 4-5 December

[268] L. Swarbrick and I.C. Whitfield (1972) Auditory cortical units selectively responsive to stimulus 'shape'. *Journal of Physiology (London)*, **224**, 68-69P

[269] B.W. Tansley and J.B. Suffield (1983) Time course of adaptation and recovery of channels selectively sensitive to frequency and amplitude modulation. *Journal of the Acoustical Society of America*, **74** (3), 765-775

[270] Y. Tougas and A.S. Bregman (1985) Crossing of auditory streams. *Journal of Experimental Psychology: Human Perception and Performance*, **11** (6), 788-798

[271] S. Ullman (1979) The interpretation of visual motion. *The MIT press*, London

[272] A.P. Varga and R.K. Moore (1990) Hidden Markov model decomposition of speech and noise. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 845-848

[273] N.F. Viemeister (1979) Temporal modulation transfer functions based on modulation thresholds. *Journal of the Acoustical Society of America*, **66** (5), 1364-1380

[274] H.F. Voigt and E.D. Young (1980) Evidence of inhibitory interactions between neurons in dorsal cochlear nucleus. *Journal of Neurophysiology*, **44** (1), 76-96

[275] G.H. Wakefield and N.F. Viemeister (1984) Selective adaptation to linear frequency-modulated sweeps: Evidence for direction-specific FM channels? *Journal of the Acoustical Society of America*, **75** (5), 1588-1592

[276] T. Watanabe and K. Ohgushi (1968) FM sensitive auditory neuron. *Proceedings of the Japan Academy Series B*, **44**, 968-973

[277] M. Weintraub (1985) A theory and computational model of monaural auditory sound separation. *Unpublished Ph.D. Thesis*, Stanford University

[278] I.C. Whitfield (1956) Electrophysiology of the central auditory pathway. *British Medical Bulletin*, **12**, 105-109

[279] I.C. Whitfield and E.F. Evans (1965) Responses of auditory cortical neurons to stimuli of changing frequency. *Journal of Neurophysiology*, **28**, 655-672

[280] S.M. Williams, R.I. Nicolson and P.D. Green (1990) Streamer: Mapping the auditory scene. *Proceedings of the Institute of Acoustics*, **12** (10), 567-575

[281] J.P. Wilson (1987) Mechanics of middle and inner ear. *British Medical Bulletin*, **43** (4), 821-837

[282] L.Z. Wise and D.R.F. Irvine (1985) Topographic organization of interaural intensity difference sensitivity in deep layers of cat superior colliculus: Implications for auditory spatial representation. *Journal of Neurophysiology*, **54** (2), 185-209

[283] M. Withgott, S.C. Bagley, R.F. Lyon and M.A. Bush (1987) Acoustic-phonetic segment classification and scale-space filtering. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 860-863

[284] T.C.T. Yin and J.C.K. Chan (1988) Neural mechanisms underlying interaural time sensitivity to tones and noise. In *Auditory Function: Neurobiological Bases of Hearing*, edited by G.M. Edelman, W.E. Gall and W.M. Cowan, Wiley, 385-430

[285] W.A. Yost and S. Sheft (1989) Across-critical-band processing of amplitude modulated tones. *Journal of the Acoustical Society of America*, **85** (2), 848-857

[286] W.A. Yost, S. Sheft and J. Opie (1989) Modulation interference in detection and discrimination of amplitude modulation. *Journal of the Acoustical Society of America*, **86** (6), 2138-2147

[287] E.D. Young (1984) Response characteristics of neurons of the cochlear nuclei. In *Hearing Science*, edited by C. Berlin, College-Hill Press, 423-460

[288] E.D. Young and M.B. Sachs (1979) Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory nerve fibres. *Journal of the Acoustical Society of America*, **66** (5), 1381-1403

[289] E.D. Young, J.M. Robert and W.P. Shofner (1988) Regularity and latency of units in ventral cochlear nucleus: Implications for unit classification and generation of response properties. *Journal of Neurophysiology*, **60** (1), 1-29