A Theory and Computational Model of Auditory Selective Attention

Stuart Nicholas Wrigley

Doctor of Philosophy in Computer Science

Department of Computer Science
University of Sheffield

August 2002

A Theory and Computational Model of Auditory Selective Attention

Stuart Nicholas Wrigley

Abstract

The auditory system must separate an acoustic mixture in order to create a perceptual description of each sound source. It has been proposed that this is achieved by a process of auditory scene analysis (ASA) in which a number of *streams* are produced, each describing a single sound source.

Few computer models of ASA attempt to incorporate attentional effects, since ASA is typically seen as a precursor to attentional mechanisms. This assumption may be flawed: recent work has suggested that attention plays a key role in the formation of streams, as opposed to the conventional view that attention merely selects a pre-constructed stream.

This study presents a conceptual framework for auditory selective attention in which the formation of groups and streams is heavily influenced by conscious and subconscious attention. This framework is implemented as a computational model comprising a network of neural oscillators which perform stream segregation on the basis of oscillatory correlation.

Within the network, attentional interest is modelled as a gaussian distribution in frequency. This determines the connection weights between oscillators and the attentional process - the attentional leaky integrator (ALI). A segment or group of segments are said to be attended to if their oscillatory activity coincides temporally with a peak in the ALI activity. The output of the model is an 'attentional stream': a description of which frequencies are being attended at each epoch.

The model successfully simulates a range of psychophysical phenomena. Furthermore, a number of predictions are made and a psychophysical experiment is conducted to investigate the time course of attentional allocation in a binaural streaming task. The results support the model prediction that attention is subject to a form of 'reset' when the attentional focus is moved in space.

Acknowledgements

The inspiration and initial psychophysical data upon which this work is based came from a presentation made by Bob Carlyon (MRC Cognition and Brain Sciences Unit, Cambridge) at the 1999 British Society of Acoustics Short Papers Meeting. I would like to thank Bob for kindly supplying me with a pre-print of his JEP:HPP paper and subsequent discussions at various stages since.

Special thanks are undoubtedly due to my supervisor, Guy Brown, for his support and inspiration throughout all stages of the project. I would also like to thank Martin Cooke and Mike Holcombe for their valuable guidance during the various progress reviews.

I would like to thank the members of the Speech and Hearing group here at Sheffield, both past and present, who have made my graduate studies, and the numerous coffee breaks, an interesting and enjoyable time.

Financial support for this work came from the University of Sheffield Hossein Farmy Scholarship and is greatly appreciated. I am grateful for the revenue generated from a feasibility study for the Defence Evaluation and Research Agency (DERA) which allowed me to attend numerous international research conferences.

My friends are also due a big thank you for supporting me, albeit occasionally with plenty of dry humour, throughout my PhD. Thanks go to, among many others, Jon, Paul, Sarah, Gillian, Philip, Kerry, Astri, Dave and Claire.

I would especially like to thank Fleur for making even the tough times enjoyable - her love and support has been invaluable in producing this thesis.

Finally, I would like to thank my parents for their love, understanding, financial support and words of encouragement just when they’re needed most.

Contents

Chapter 1. Introduction 1
Making sense of a complex environment 1
Auditory Scene Analysis 3
Computational auditory scene analysis 4
Attention in listening 5
Motivation 6
Modelling approach 7
Thesis overview 9

Chapter 2. Auditory Scene Analysis 11
Introduction 11
Perceptual grouping and streams 12
Primitive grouping 13
Common fate 13
Similarity and proximity 14
Continuity and closure 16
Simultaneous and sequential grouping 17
Schema-driven grouping 18
Computational ASA 19
Summary 21

Chapter 3. The Binding Problem 23
Introduction 23
Combinatorial solutions 25
Attentional solutions 27
Temporal correlation solutions 29
Interim summary 32

Computational models of feature binding 34
Synfire chains 34
Neural oscillators 39
Summary 44

Chapter 4. Auditory Selective Attention 47
Introduction 47
Attentional allocation in audition 48
Frequency location 48
Spatial location 50
Space-frequency allocation 52
Attentional ‘shape’ 54
Two forms of attention 54
Interim summary 55
Selective attention in stream formation 56
Perception without attention 58
Interim summary 62
Computational models of ASA incorporating attention 62
Functional approaches 62
Neural oscillator-based approaches 64
Attentional allocation in vision 66
Psychophysics 66
Computational models 67
Auditory and visual attention 68
Summary 69

Chapter 5. A Conceptual Framework for Auditory Selective Attention .71
Introduction 71
Theories of selective attention 72
Early selection 72
Late selection 75
Interim summary 76
A new conceptual framework for auditory selective attention 77

Endogenous attention 80
Summary 84

Chapter 6. A Computational Model of Auditory Selective Attention . . 87
Introduction 87
Monaural computational model 90
Auditory peripheral processing 92
Pitch and harmonicity analysis 96
Segment identification 100
Neural oscillator network 103
Segment formation and primitive grouping 105
Attentional Leaky Integrator (ALI) 107
Interim summary 109
Binaural computational model 110
Segment grouping 111
Attentional Leaky Integrator (ALI) 113
Summary 114

Chapter 7. Evaluation 117
Introduction 117
Output representation 118
Monaural stimuli 119
Two tone streaming 119
Redirection of attention 124
Harmonic segregation within a complex tone 124
Timecourse of attentional build-up 128
Binaural stimuli 129
Two tone streaming with distractor task 130
Harmonic segregation within a complex tone 132
Predictions 135
Psychophysical investigation of the spatial allocation of attention 138
Experimental method 140
Results 141

Discussion 142
Summary 142

Chapter 8. Conclusions 145
Summary 145
Original contribution 147
Limitations of model 149
Future work 151
Exclusive allocation 151
Divided attention 153
Binaural enhancements 155
Computational efficiency 157
Grouping principles 157
Conclusions 158

Chapter 9. References 159

Appendix A.Computational Model Parameters 185

Appendix B.Psychophysical Experiment Subject Responses 187
Introduction 187
Subject GB 188
Subject JE 189
Subject SC 190
Subject SM 191
Subject ST 192
All subjects 193

Chapter 1. Introduction

1.1 Making sense of a complex environment

For hundreds of years, philosophers and psychologists have argued that the act of *perception* is a process which allows us to produce a mental representation of the world from the multitude of information gathered by our senses. In turn, this implies that we have a mechanism of forming such representations and then using these to formulate plans and perform actions. In other words, we have the ability to

conceptually move from raw sensory input to some representation of the world described by that input.

Unfortunately, this is not as easy as it may first appear. In order to bridge the gap between the sensory input and the perceptual representation, we need to decide which bits of the input belong together. If we were unable to do this, everything would coalesce into an unmeaningful whole: the input contains plenty of 'meaning' but it needs to be extracted. Therefore, the problem is one of how to accurately partition the sensory input.

An overall representation of the world is made up of a number of smaller representations of individual 'things'. Visually, these could correspond to physical objects; in audition, these could be individual sounds. The important question, therefore, is how each representation is made from the sensory data. In *visual* scene analysis, one problem can be thought of as being how regions in a scene are allocated to a particular object. Visually, a simplistic answer to this could be to allocate each bounded region to an individual object. However, this is clearly overly simplified since when observing an occluded object, we are able to allocate the visible regions to a single representation.

Such fundamental problems of how to bridge the gap between the input and the representation are also present in audition. Such difficulties are clearly demonstrated in Bregman's (1990) metaphor, which makes an analogy between sound waves and the waves on a lake:

> *Your friend digs two narrow channels up from the side of a lake. Each is a few feet long and a few inches wide and they are spaced a few feet apart. Halfway up each one, your friend stretches a handkerchief and fastens it to the sides of the channel. As the waves reach the side of the lake they travel up the channels and cause the two handkerchiefs to go into motion. You are allowed to look only at the handkerchiefs and from their motions to answer a series of questions: How many boats are there on the lake and where are they? Which is the most powerful one? Which one is closer? Is the wind blowing? Has any large object been dropped suddenly into the lake?*
>
> *Bregman (1990, pp. 5-6)*

As Bregman points out, this problem seems impossible but is in fact exactly the same as that faced by the auditory system. The channels at the side of the lake correspond to the ear canals and the handkerchiefs correspond to the ear drums. Simply by analysing the motion of the ear drums, the auditory system can answer analogous questions regarding the sound environment (c.f. the lake). Are there

people talking nearby? If so, what are they saying? Is there any music playing? Which sound sources are moving?

As we have mentioned above, in order to make sense of an environment, we must be able to partition the incoming input into representations of objects (be they visual or auditory). In the lake analogy, we have seen that despite a mixture of sounds reaching the ears, the human listener can still 'pick out' a particular voice or instrument demonstrating that they can separate this complex mixture. The robustness of such an ability is demonstrated by Cherry's (1953) now famous study on the separability of concurrent conversations. He found that when two messages were recorded by the same speaker and replayed simultaneously to a listener, '*the result is a babel, but nevertheless the messages may be separated*' (p. 976) - an ability colloquially termed the *cocktail party effect*.

Our ability to perform well in this task highlights another important aspect in making sense of a complex environment: *attention*. It is through attention that we can select, or 'pick out', particular objects for more detailed study. In Cherry's work, the listener was consciously attending to one particular message. In a similar experiment, Spieth *et al.* (1954) investigated conditions under which a military communication's operator could best recognise and attend to one speech message when it was presented simultaneously with another irrelevant message. Such communication messages did not provide visual cues to aid in the identification of the sender or the perception of the message. While redundancy within a message is high, competing messages are of similar form, content, and vocabulary. Spieth *et al.* concluded that '*the operator could acquire rapidly [an attentional] "set" to listen to the "high-pitched" message only or the "low-pitched" message only*' (p. 394).

Since these initial studies, much psychoacoustic research has been conducted into the ability to make sense of a complex acoustic environment - with particular reference to trying to uncover *how* the auditory system achieves this. In other words, to investigate how the auditory system may form a representation of the acoustic world. Bregman (1990) has drawn upon this body of work to convincingly argue that the acoustic signal is subject to a similar form of scene analysis as vision.

1.2 Auditory Scene Analysis

Bregman (1990) proposed that *auditory scene analysis* (ASA) takes place in two stages. Firstly, the signal is decomposed into a number of discrete sensory *elements*. These are then recombined into *streams* on the basis of the likelihood of them

having arisen from the same physical source in a process termed *perceptual grouping*. Therefore, a stream can be considered to be a cognitive representation of a sound source.

Bregman distinguished between two different, but complementary, mechanisms involved in the grouping problem. The first is called *primitive* grouping in which decisions on how to group the sensory elements are made in a purely data-driven manner. In contrast, the other grouping mechanism, termed *schema-driven* grouping, employs prior knowledge of commonly experienced acoustic stimuli.

Primitive grouping principles are believed to be innate and can be described by the *Gestalt* principles of perceptual organisation (e.g. Koffka, 1936). The Gestalt psychologists proposed a number of rules governing the manner in which the brain forms mental patterns from elements of its sensory input. For instance:

- *Common fate*: elements whose properties change in a similar way over time tend to be grouped.
- *Similarity*: elements that are similar in physical attributes (such as timbre, pitch, or loudness) tend to be grouped
- *Proximity*: grouping strength between elements, or groups of elements, is proportional to the distance between them.
- *Continuity*: provided the changes between consecutive elements are smooth then the elements tend to be grouped.
- *Closure*: elements that form a complete, but possibly partially obscured, object tend to be grouped.

Such rules were originally formulated to explain grouping in the visual domain but, as we will see in chapter 2, they are equally applicable to auditory grouping. For example, when one observes a partially occluded object, it is usually quite easy to form a mental representation of the whole object. The same is true of auditory occlusion (masking) in which listeners are able to perceptually restore parts of sounds masked by louder sounds - a concept known as auditory induction (e.g. Miller and Licklider, 1950).

Computational auditory scene analysis

> *There are some practical reasons for trying to understand [ASA]. There are engineers that are currently trying to design computers that can understand what a person is saying. However in a noisy environment, the speaker's voice comes mixed with other sounds. To a naive computer, each different sound that the voice comes mixed with makes it sound as if different words were*

> *being spoken, or as if they were spoken by a different person. The machine cannot correct for the particular listening conditions as the human can. If the study of human audition were able to lay bare the principles that govern the human skill, there is some hope that a computer could be designed to mimic it.*
>
> *Bregman (1990, p. 3)*

As Bregman (1990) indicates, a major goal of computational auditory scene analysis (CASA) is to improve automatic speech recognition (ASR) performance in noisy environments. Early work concentrated on the separation of multiple voices using engineering solutions. For example, Parsons (1976) aimed to completely characterise the frequency spectrum of the target voiced speech which would then allow an individual speech 'stream' to be resynthesised, an approach that has seen recent refinement (Denbigh and Zhao, 1992). Work has also concentrated on producing faithful models of the human auditory periphery to both produce CASA solutions and gain insights into the mechanisms of the auditory system. Within this framework, researchers began to separate voiced speech for both stationary sounds (e.g. Scheffers 1983; Meddis and Hewitt 1992) and non-stationary sounds (e.g. Weintraub, 1985).

However, many of the early CASA solutions made rather restrictive assumptions such as there being at most two sources, both of which are voiced speech. Furthermore, the performance of these systems relied on each source having a different fundamental frequency. To avoid such limitations, a different approach to the problem was adopted by a number of workers who developed purely data-driven solutions (e.g. Cooke 1991/1993; Mellinger, 1991; Brown 1992). These systems segmented the signal into atomic units (regions of time-frequency that contain energy judged as belonging to a single source) which were then grouped together on the likelihood of them having arisen from the same source, using cues described by Bregman (1990). Recent neural network solutions (e.g. Wang 1996; Brown and Cooke, 1997; Wang and Brown, 1999), based on the neurophysiological speculations of von der Malsburg and Schneider (1986), form the foundations of a model which will be presented in later chapters of this thesis.

1.3 Attention in listening

The term '*attention*' commonly appears in every day language and can be considered to be a fundamental part of daily life. As we have seen in the work of Cherry (1953) and Spieth *et al.* (1954), it refers to both the selectivity and capacity

limitation of cognitive processing. It is widely accepted that conscious perception is selective and that perception encompasses only a small fraction of the information impinging upon the senses. The second phenomenon - that of capacity limitation - can be illustrated by the fact that two tasks when performed individually pose no problem; however, when they are attempted simultaneously, they become difficult. In other words, when one focuses attention on something, we are effectively highlighting it as requiring further processing, at the expense of everything else.

In modelling auditory scene analysis, attention has traditionally been viewed as a distinct process in which the ASA output is subjected to some form of selection to pick the 'winning' stream. Indeed, a number of computational models, which will be reviewed in chapter 4, make explicit the assumption that attentional input is distinct from the grouping process and hence leave it for future work. For example, although McCabe and Denham (1997) include an 'attentional input' in their model architecture (see chapter 4, figure 9), the authors acknowledge that the role of attention was not addressed in the model processing.

However, a recent psychoacoustic study (Carlyon *et al.*, 2001) cast doubt on this assumption when it was found that without attention, stream formation did not occur. In this case, attention can no longer be viewed as an 'optional extra': it must be incorporated into computational models of ASA if they are to behave in a similar way to human listeners. With this in mind, we have developed a model of computational auditory scene analysis in which the allocation of attention is at the heart of the stream formation mechanism.

1.4 Motivation

Since the importance of attention to auditory scene analysis has only recently been highlighted, the model presented here is motivated by two different goals. Firstly, in addition to demonstrating that attention can be successfully incorporated into the heart of an ASA system, such a model can be used to make predictions and indicate directions for future psychophysical experiments (see chapter 7).

Secondly, the model is also motivated by a practical application. Since the advent of computational models of auditory scene analysis, such systems have been considered to form the natural evolution of current hearing aid technology. Conventional hearing aids have a microphone that gathers sound, an amplifier that increases the volume of sound and a receiver that transmits this amplified sound to the ear. Clearly, such hearing aids will amplify both the source of interest (such as

speech) as well as any interfering noise. Ideally, an implementation based on the stream formation and segregation mechanisms of a computational ASA system would be preferable. Instead of amplifying the entire frequency range, such devices could segregate the acoustic environment into any number of streams, one of which (for example, a speech stream) could be selected with all the others being attenuated.

The interesting enhancement that is made possible by incorporating attention into models of ASA is to allow such an advanced hearing aid to be 'intelligent' to some degree. The processing performed by these prostheses would mimic the hearing ability of a unimpaired listener. Instead of enhancing the stream of interest at all times at the expense of the others, attentional effects such as the unconscious overruling of conscious selection (such as the startle reflex; see chapter 5) would allow the wearer to become aware of new, loud and potentially important sounds. Imagine crossing a road whilst in mid-conversation with a friend; it would certainly be advantageous for the sound of a car horn to come to the fore at the expense of the conversation stream - an effect that may not occur in a conventional streaming solution.

It ought to be noted that the incorporation of some attentional effects into models of ASA may not be desirable in all engineering paradigms. The ability of an ASA system to produce a number of streams each representing a particular sound source has been viewed as a possible robust front-end for automatic speech recognition (ASR) in noise (e.g. Barker *et al.*, 2001; Brown *et al.*, 2001). In this situation, the noise would be encapsulated in its own distinct stream allowing ASR to be performed on the relatively clean speech stream. However, in contrast to the example given above, an ASR system would not want the speech stream to be removed from the fore in favour of the car horn: the task of ASR is generally to be robust to intruding noise and be able to process as much of the speech as possible.

1.5 Modelling approach

Bregman acknowledges that the Gestalt based grouping principles he describes only provide a description of the phenomena but do not give any insight into how or why they occur. This level of explanation, together with its analogies with visual scene analysis, draws clear parallels between his framework and that of David Marr (Marr, 1982). Hence, it is all the more intriguing that Bregman makes no mention of Marr's influential theoretical framework for describing the processes of early vision (Williams *et al.*, 1990). Marr (1982) defined three levels of explanation of a system.

The *computational theory* describes the function and motivation of a system and the knowledge it has available to achieve this. In other words, the 'what and why'. The next level of explanation describes the *representation* of the data and the *algorithm* by which it is to be manipulated: 'how' the system works. Finally, the most concrete level of explanation is that of the *implementation*: the physical realisation of the data and algorithm. Furthermore, Marr argued that these three different levels are only loosely related such that explanations at each level can be largely independent from each other.

In terms of Marr's three levels, Bregman's description of auditory scene analysis is, in effect, a computational theory of auditory perception (Williams *et al.*, 1990; Cooke, 1991/1993). It is at the level of the representation and algorithm that many CASA solutions can be categorised. In other words, they describe how the general Gestalt principles of grouping can be implemented and combined in their particular data representation framework. Unfortunately, at the level of the neural implementation, researchers are only just scratching the surface in their endeavours to describe how the brain implements such complex grouping principles. Indeed, Feng and Ratnam (2000) note in their review of work in this field that '*clearly, physiological investigations of neural mechanisms of scene analysis are still in their infancy, and gaps in our understanding are profound*' (p. 717). With this in mind, the model presented in later chapters also belongs in the intermediate level of explanation. It is how attentional processes influence the scene analysis algorithm which is of interest in this work.

It ought to be noted, however, that the Marrian approach does have its difficulties. As Brown (1992) indicates, the loose coupling of Marr's three levels of explanation can lead to an unprincipled model unless care is taken to '*employ representations that are motivated by the ... organisation of the higher auditory system*' (p. 6). Churchland and Sejnowski (1992) also point out that in many studies, including those of Marr himself, the coupling between explanation levels is not as loose as Marr proposes. In a departure from the independence of the algorithmic level from its implementation, '*downward glances [to the implementation level] figured significantly in Marr's attempts to find problem analyses and algorithmic solutions*' (p. 18). Marr was significantly influenced by neurobiological considerations and the findings from such research heavily influenced his computational and algorithmic insights. Indeed, this blurring of the boundaries is what Brown (1992), and many who propose 'physiologically plausible' solutions, advocate. Churchland and Sejnowski (1992) emphasise that,

> *The issue of independence of levels marks a major conceptual difference between Marr (1982) and the current generation of researchers studying*

> *neural and connectionist models. In contrast to the doctrine of independence, current research suggests that considerations of implementation play a vital role in the kinds of algorithms that are devised and the kind of computational insights available to the scientist. Knowledge of brain architecture, far from being irrelevant to the project, can be the essential basis and invaluable catalyst for devising likely and powerful algorithms - algorithms that have a reasonable shot at explaining how in fact the neurons do the job.*
>
> *Churchland and Sejnowski (1992, p. 19)*

To summarise, Marr's framework of different explanatory levels allows the problem of human perception to be split into convenient blocks for analysis. However, the independence between these levels should not have to be strictly enforced and, in fact, often isn't. This does not necessarily weaken Marr's theory: in attempting to understand how the brain works, why should we not gain inspiration from how the brain is implemented? It is in this way that we approach the problem presented in later chapters: despite presenting a computational theory of how attention interacts with ASA, insights into brain architecture and how grouping may be signalled at a neuronal level heavily influence our algorithm and implementation.

1.6 Thesis overview

Chapter 2 explains the principles of auditory scene analysis introduced above in greater detail. As we have discussed, these principles can be considered to be a computational theory of hearing: they give no insights into how the brain may represent a number of features as belonging together. The human brain relies solely on time varying electrical impulses with no 'symbolic' input as suggested by Bregman's theory. Chapter 3 looks at a number of ways in which the brain may represent stimulus features and perform feature grouping. One possibility is *temporal synchronisation* in which neurons that code features of the same source have a synchronised firing response (phase-locked with zero phase lag) and which are desynchronised from neurons that code features of different sources. We will also look at how neurons which exhibit an oscillatory firing behaviour, *neural oscillators*, have been used in previous computational models to simulate auditory grouping and argue for their use in the model presented here. One advantage of the neural oscillator approach is that it is well suited to implementation in hardware allowing the potential of real-time performance; clearly a requirement if such systems are to be deployed in advanced hearing prostheses as suggested above.

Chapter 4 describes a number of psychophysical experiments which allow us to draw conclusions as to the way in which attentional processes influence ASA. The manner in which such influences are combined in auditory scene analysis are discussed in chapter 5.

Chapter 6 deals with the implementation of a neural oscillator model of auditory selective attention which takes into account both the 'conventional' ASA principles of Bregman (1990) and of the attentional influences described in chapters 4 and 5. The model is then evaluated in chapter 7 by simulating the outcomes of a number of psychophysical experiments described in preceding chapters. The analysis of these results are used to formulate predictions which can be made by the model and also highlight a number of areas for future psychophysical investigation.

Chapter 2. Auditory Scene Analysis

2.1 Introduction

Our ability to make sense of a world which is almost constantly filled with sounds has been investigated by numerous psychoacousticians ever since the term '*cocktail party effect*' was introduced by Cherry (1953). This effect is related to how a listener at a party, subjected to multiple sound sources, is able to separate, and participate in, a single conversation; a problem which scales to ask how a listener makes sense of any complex auditory environment.

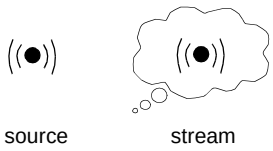


Figure 1 **The relationship between a sound source and its percept - the stream.**

The use of psychoacoustic experiments consisting of simple sounds, such as sine waves and noise bursts, has made it possible to derive very important conclusions on how the auditory system processes sound mixtures; many of these findings are incorporated into the conceptual framework of *auditory scene analysis* (ASA) (Bregman, 1990). ASA proposes that the acoustic signal is subject to a similar form of scene analysis as vision. Such auditory scene analysis takes place in two stages. Firstly, the signal is decomposed into a number of discrete sensory *elements*. These are then recombined into *streams* on the basis of the likelihood of them having arisen from the same physical source in a process termed *perceptual grouping*.

2.2 Perceptual grouping and streams

At the heart of Bregman's (1990) account of auditory scene analysis is the formation of *streams*: a perceptual unit that represents a single acoustic source (figure 1).

The word *sound* is insufficient as it is essential that the perceptual unit be able to incorporate more than one acoustic event. For example, the perception of a piano being played is a single experiential event which is made up of numerous individual sounds - notes. In this example, there is only one *source*: the piano. A source is the physical generator of a sound. It is usual for a sequence of sounds originating from the same source to be perceived as a stream. However, it is also possible for a number of sources to contribute to one stream (e.g. the perception of music) and for multiple streams to be formed from a single source (e.g. auditory streaming).

As mentioned in the introduction, the initial stage of auditory scene analysis is the decomposition of a sound into a collection of sensory elements. The second stage of processing is stream formation and segregation. The mechanism by which these sensory elements are combined is termed grouping; this is split further into *primitive* and *schema-driven* grouping.

2.2.1 Primitive grouping

Primitive grouping (bottom-up processing) encompasses the data-driven *simultaneous* and *sequential* perceptual organisations of sound. Bregman (1990) explains primitive grouping in terms of *Gestalt* principles of perceptual organisation (e.g. Koffka, 1936). Gestalt psychology was founded in Germany by Max Wertheimer, Wolfgang Köhler and Kurt Koffka at the beginning of the twentieth century as a challenge to the then-prevailing theory of *Structuralism*. A key assumption of Structuralism was *elementarism* which argued that complex perceptions could be understood by identifying the elemental components of the experience. However, Gestalt theorists argued that perception was much more than the sum of its parts. A melody transposed in key remains the same melody despite each constituent note having been shifted in frequency (and thus each note being different): the melody is an *emergent property* of the sequence of notes created by some organisation of the nervous system. In order to explain how such organisations may be formed, Wertheimer proposed a number of grouping laws by which the visual system creates perceptual wholes. He proposed that elements tend to be grouped together if they exhibit some degree of proximity or similarity; form a closed contour; or move in a similar way.

Almost a century later, Wertheimer's laws of grouping are still used to explain perceptual organisations in both visual and auditory modalities (e.g. Bregman, 1990); indeed '*not one of them has been refuted*' (Rock and Palmer, 1990). The following are some of the important grouping cues in auditory scene analysis. It should be noted, however, that a certain degree of competition occurs between cues which suggest contradictory grouping decisions and some cues dominate others (Donnelly *et al*., 1991; Elder and Zucker, 1993; Kovacs and Julesz, 1993). The method by which such competitions are resolved remains uncertain.

Common fate

Common fate describes the tendency to group components whose properties change in a similar way over time. Stimulus features which are subject to these common fluctuations are generally perceived as one object, making it difficult to focus on the individual components.

Grouping by harmonicity is an example of common fate organisation which occurs in many listening situations. Many natural sounds, especially animal communication, are caused by the vibration of some physical structure and an element of filtering and resonance (see Fant, 1960). For example, in speech the vocal cords are drawn close together and the force of the exhaled air from the lungs

causes them to vibrate. The sound wave resulting from this vibration consists of a *fundamental* and a number of related *harmonics* (components whose frequencies are integer multiples of the fundamental). Various harmonics are then 'reinforced' within the air cavities of the vocal tract. It is this reinforcement, or resonance, that produces the *formants* commonly found in speech. The presence of fundamentals and harmonics are also seen in other complex sounds, such as music.

Indeed, support for harmonicity grouping is strengthened by evidence which shows that when one listens to two steady complex sounds (e.g. two musical notes or two vowel sounds) the two are generally easily separable. This means that we are able to decide which harmonics belong to each sound percept even if the harmonics are coincident or interleaved. It is important to note that this behaviour is more easily observed if each sound has a different fundamental frequency (Broadbent and Ladefoged, 1957; Scheffers, 1983; Assmann and Summerfield, 1987, 1994; McKeown and Patterson, 1995).

Similarly, common onset and offset are also part of the wider common fate cue, and correspond to the perceptual grouping of components whose onsets and / or offsets occur simultaneously. Darwin (1984) has shown that a tone that starts or stops at the same time as a vowel sound is more likely to be heard as part of the vowel complex than if the onset and / or offset times had been different. In support of this grouping principle, Roberts and Moore (1991) demonstrated that tones added to a vowel sound had a significantly stronger effect on vowel quality when presented with identical onset and offset times.

Similarity and proximity

The degree of similarity between elements also plays a role in the grouping process. In hearing, similarity usually implies closeness of timbre, pitch, or loudness. In figure 2 (panels a and c), the formation of two groups occurs when a difference in colour is introduced. In audition, it is this ability that allows successive notes of an instrument to be segregated from notes of another instrument, even when both are played in the same register and spatial location: in this case timbre difference is the cause (Wessel, 1978). Indeed, timbre tends to be the dominant similarity cue as opposed to pitch. Van Noorden (1975) investigated this using sequences of complex tones with missing fundamentals and various missing harmonics. Tones with the same missing fundamental (and hence the same pitch) did not necessarily form a single stream whereas tones with different pitches but similar timbres did.

The Gestalt proximity principle states that the grouping strength between elements, or groups of elements, is proportional to the distance between them. For example in

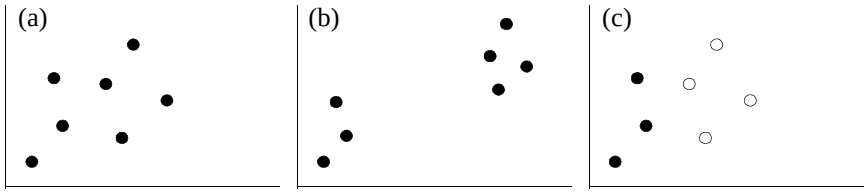


Figure 2 **Gestalt proximity describes how closer elements are more strongly grouped. Panel (a) shows one large group. Panel (b) shows how slight alteration of the element locations causes two groups to emerge. Gestalt similarity describes how similar elements are more strongly grouped. Panel (c) shows how the large group segregates when colour differences are introduced: elements of similar colour form distinct groups.**

figure 2, when the spacing between adjacent elements is similar, the grouping strength is also comparable. However, once differences in separation are introduced, two groups emerge due to members of one cluster being closer to other members of that cluster than members belonging to the other cluster. The auditory correlate of this visual cue is the separation of acoustic elements in time and frequency.

For example, the relationship between frequency proximity and temporal proximity has been studied extensively using the two tone streaming phenomenon (see Bregman, 1990 for a review). The closer in frequency two tones are, the more likely it is that they are grouped into the same stream. Similarly, the proximity of two tones in *time*, also determines likelihood of streaming. As presentation rate increases, tones of similar frequency group together into separate streams. (see figure 3).

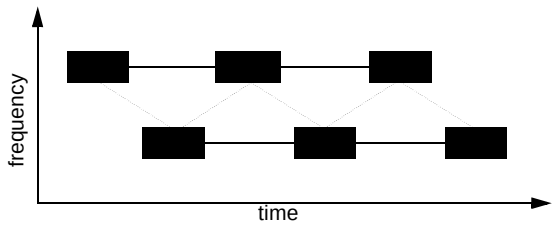


Figure 3 **Spectrogram of six alternating tones of A-B... format. Initially, the tones are temporally coherent and a single stream is present containing the entire A-B... sequence (dotted lines). If stream segregation occurs, the high tone sequence and the low tone sequence form separate streams (indicated by the solid lines) with one stream containing a A-A-A sequence and the other a B-B-B sequence.**

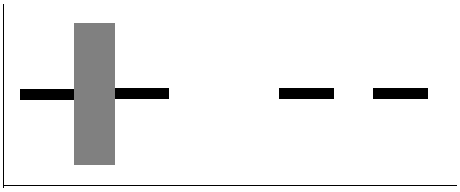


Figure 4 **Gestalt closure. The gray band present in the left diagram appears to obscure a single bar. However, when the band is removed, it becomes clear that two distinct bars are present.**

Continuity and closure

Certain interrupted and / or smoothly changing forms are perceptually grouped into a whole. For the continuity perception to occur, there must be sufficient evidence to support the hypothesis that the form is obscured rather than interrupted. Such behaviour is evident from figure 4 in which a solid bar appears to be obscured by a gray band. When the gray band is missing, it is clear that there are two distinct bars. Similarly, when the figures are considered to be pseudospectrograms (frequency on the ordinate and time on the abscissa) with the bar representing a tone and the gray band representing noise, provided the noise is sufficiently loud, the noise band will be perceived as obscuring a single tone, rather than separating two distinct tones. This effect can also been seen for speech signals which are obscured by noise. The speech sounds much more intelligible and continuous when interrupted by noise than when interrupted by silence (Miller and Licklider, 1950; see also Warren *et al.*, 1972).

Good continuation is also recognised as playing an important role in perceptual grouping. Provided the changes between consecutive elements are smooth then the elements tend to be grouped together. This can be illustrated using a second closure example. Consider the left hand diagram in figure 5. This shows the closure example as in figure 4. The second reason why the two bars form a single percept is that they also exhibit good continuation: there is no significant jump in frequency. The same is true for the middle diagram. Here, although the second bar begins at a frequency which is significantly different to that at which the first bar finished, the combination of closure gives it the appearance of a smoothly changing tone which is partially obscured. The final diagram shows that although closure could play a role in the perceptual grouping in this instance, good continuation is not present, hence the two bars are not grouped together to form a whole (see Ciocca and Bregman, 1987).

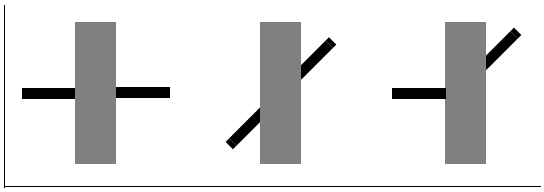


Figure 5 **Gestalt closure combined with good continuation. Only bars which exhibit a smooth change are grouped into a perceptual whole.**

Darwin and Bethell-Fox (1977) demonstrated how good continuation of F0 affects speech perception. They synthesised spectral patterns which varied smoothly and repeatedly between two vowel sounds. When the pattern had a constant F0, the sounds were perceived as belonging to a single stream. However, when the pattern exhibited a discontinuous F0 contour between the two vowel sounds, each sound was perceived as belonging to a distinct stream with illusory silences during the portions of the signal attributed to the other stream. Components were grouped only in the situation in which the signal had good continuation (i.e. smooth change) of F0.

As mentioned above, there is a degree of competition between grouping cues. Competition between good continuation and proximity can be demonstrated by the use of a frequency-alternating tone sequence. Bregman and Dannenbring (1973) showed that the tendency of such a sequence to split into two streams (see figure 3) is reduced when successive tones are connected by frequency glides. A similar effect, albeit less effective in disrupting the streaming percept, is observed when partial glides are used. Caution should be exercised when drawing conclusions from these findings; Bregman and Dannenbring originally claimed that the disruption effect was due solely to good continuation cues: the glides enhanced the predictability of the position of the next tone in frequency. However, Bregman (1990) has since suggested that the presence of glides simply reduce the frequency separation of the tones allowing the disruption of streaming to be attributable entirely to proximity cues. Indeed, unpublished experiments performed by Bregman and Dannenbring at the time of the 1973 paper support this proposal (Bregman, 1990).

Simultaneous and sequential grouping

The grouping factors described above can be considered to fall into two general categories (Bregman and Pinker, 1978). Simultaneous grouping rules (proximity and common fate) are concerned with stimulus components which occur at the

same time; sequential grouping rules (similarity and good continuation) are concerned with components which arise from the same source over time. A further example of the competition between grouping rules highlights the interaction between simultaneous and sequential grouping cues when organising a sound scene. As described above, the frequency components of a complex tone are grouped on the basis of common fate - a simultaneous grouping cue (see page 13). However, such grouping can be disrupted in the presence of sequential grouping factors. Darwin *et al*. (1989; see also Darwin *et al*., 1995) embedded a complex tone within a sequence of pure tones in order to perceptually remove a single frequency component from the complex (at the frequency of the pure tones). In the presence of the tone sequence, the sequential grouping cue of similarity relating to the tone sequence and complex component became the dominant cue, thus preventing the component from being grouped with the rest of the complex. The frequency component was grouped as an additional tone of the sequence and was hence removed from the complex.

2.2.2 Schema-driven grouping

> *Schema - A mental codification of experience that includes a particular organized way of perceiving cognitively and responding to a complex situation or set of stimuli.*
>
> *Merriam-Webster Collegiate Dictionary*

In contrast to the primitive, bottom-up, grouping mechanisms described above, the perceptual system can also employ prior knowledge about common sounds, such as speech or music, to organise the acoustic environment into streams: schema-driven grouping. It is believed that schemas directly encode the spectral characteristics of a stimulus; however, the precise information stored and how it is represented is still unknown. Such schemas can represent sounds of arbitrary duration, ranging from phonemes, to words and even entire melodies (Dowling, 1973; see also Hartmann and Johnson, 1991).

The former can be demonstrated by the perceptual restoration of a phoneme which has been replaced by a burst of noise (Warren and Warren, 1970). The stimulus used was of the form "*the *eel was on the axle*" or "*the *eel was on the orange*", where "*" indicates the noisemasked deleted phoneme. For these two examples, listeners reported hearing wheel and peel respectively: a top-down schema processes the speech before conscious perception occurs even when the disambiguating word occurred several words later than the deletion.

2.3 Computational ASA

Computational solutions to the ASA problem are generally motivated by one of two applications. The first is the goal of improving automatic speech recognition (ASR) performance in noisy environments. The accuracy of ASR systems whose input speech has been obtained in all but the quietest of backgrounds is poor compared to that of the human listener particularly if the noise is non-stationary. An ASA model that can successfully segregate speech from any number of interfering sound sources (including other speech) could be used as a first stage of pre-processing in a larger recognition system. The second motivation is to produce ‘advanced’, or ‘intelligent’, hearing prostheses. Instead of amplifying the entire frequency range, such devices could segregate the acoustic environment into any number of streams, one of which (for example, a speech stream) could be selected with all the others being attenuated. The further inclusion of attentional mechanisms to semi-automate the stream selection process would be a logical progression.

Attempts to create computer models that mimic auditory scene analysis have led to the field of study known as computational auditory scene analysis (CASA). Early work concentrated on speech separation by use of engineering solutions which didn’t draw upon any of the insights gained from analysis of the auditory system. For example, the model of Parsons (1976) aimed to completely characterise the frequency spectrum of the target voiced speech by identifying the speaker’s pitch and all associated harmonics. The use of a pitch tracker to maintain speaker continuity from frame to frame thus allowed the resynthesis of an individual speech ‘stream’. A similar approach, albeit more refined with the inclusion of spatial location information, was used in the work of Denbigh and Zhao (1992).

More recently, many researchers have been concerned with producing faithful models of the human auditory periphery to both produce CASA solutions and gain insights into the mechanisms of the auditory system. Much work has been conducted on the separation of voiced speech in which the ability to separate concurrent vowels by pitch is investigated (e.g. Scheffers 1983; Meddis and Hewitt 1992; Assmann and Summerfield 1994). However, despite success in duplicating the recognition rate for vowel-pairs as a function of fundamental-frequency difference, it should be noted that such sounds are an unrealistic test of a CASA system. Vowel sounds are ‘stationary’: their average spectral characteristics are constant over time; whereas most sounds are non-stationary.

Weintraub (1985) was one of first researchers to attempt to separate non-stationary sounds whilst drawing inspiration from the behaviour of the auditory system. His

system separated two simultaneous speakers by analysing the pitch of each voice extracted by analysing the temporal fine structure of each auditory filterbank channel.

All the solutions described above suffer from important limitations. Firstly, each model has made strong assumptions about the number and characteristics of the sources present: generally, each assumes at most two sources, both of which are periodic. The majority of environments, especially those in which ASR is to be used, contain many more than two sound sources, some of which will have non-speech characteristics. Recent work has aimed to produce more generic analysers in which no assumptions are made on the number and type of sources present.

One approach has been to model the organizational function of the auditory system to perform segregation in a purely data-driven manner (e.g. Cooke 1991/1993; Mellinger, 1991; Brown 1992). These systems pre-process the signal using an auditory 'front-end' to model the auditory periphery. This output is then segmented to form atomic units (regions of time-frequency that contain energy judged as belonging to a single source) which are then grouped together on the likelihood of them having arisen from the same source using cues described by Bregman (1990). A significant problem with these particular solutions is the lack of top-down schema-related information which could be incorporated into the grouping process. One method of avoiding this 'one-way data stream' is to use a *blackboard* architecture (Erman *et al.*, 1980; Engelmore and Morgan, 1988) in which independent knowledge sources create and modify a number of grouping hypotheses which reside on the blackboard. It is these hypotheses which constitute the entire 'state' of the analysis system at any one time. This technique has been successfully used in a number of CASA systems (e.g. Cooke *et al.*, 1993; Lesser *et al.*,1995; Ellis, 1996; Godsmark and Brown, 1999) as it allows many different types of knowledge (bottom-up primitive grouping rules, top-down schemas, etc.) to interact. Another solution related to the blackboard architecture is based on the multiagent paradigm (e.g. Nakatani *et al.* 1994; Nakatani *et al.* 1998; see also Minsky, 1986; Maes, 1991) in which an '*agent*' is dynamically allocated to a sound stream and maintains that stream on the basis of consistent attributes in the input. The entire system consists of a number of agencies, each with responsibility for extracting specific information (e.g. spatial information, stream fragment extraction). In turn, each agency is split into a number of agents, each of which extract stream fragments from the input which are consistent with some attributes (e.g. harmonicity). In addition to these, there are a number of agencies who combine the output of the other agencies to produce the final stream outputs.

Both of these types of solutions rely heavily on rule-based sound organisation. Recent neural network solutions (e.g. Wang 1996; Brown and Cooke, 1997; Wang and Brown, 1999) based on the neurophysiological speculations of von der Malsburg and Schneider (1986) move away from rule-based organisation in favour of 'emergent properties' of an array of interconnected nodes in which synchronisation of node activities signals successful grouping of components. This type of solution, and its use in the model presented in this thesis, will be discussed in more detail in later chapters.

2.4 Summary

This chapter has introduced some of the concepts involved in performing auditory scene analysis. The primitive, data-driven, grouping cues arise from the perceptual grouping rules proposed by Gestalt psychologists in the 1930s. Despite their age, they still form the core of visual and auditory grouping principles (Bregman, 1990). To complement these cues, the perceptual system also draws on schema-based information in which prior knowledge of common sounds can be used to perform grouping. The previous section briefly described a number of solutions to the auditory scene analysis problem, each of which employ different computational strategies. Despite the success of such systems on limited problem-sets, few scale well to tackle the full ASA problem and none is as robust as the human listener.

The difficulty involved in producing a computational solution is related to the mismatch between theories of perception, such as Bregman's, and the physiological processing substrate. Consider the two tone streaming stimulus (figure 3). Theories of perception are implied from experimental observations. Applying such mechanisms to figure 3, one can conclude that as the frequency separation decreases, it is more likely that the tones will be grouped together. Similarly, as temporal separation decreases, sequential tones will also be more likely to group. However, the neurophysiological mechanisms underlying auditory stream formation are poorly understood and it is not known how groups of features are coded and communicated within the auditory system. What does it mean to talk of 'frequency proximity' or 'temporal proximity'? The human brain relies solely on time-varying electrical impulses with no 'symbolic' input as suggested by Bregman's theory. The following chapter looks at a number of ways in which the brain may represent stimulus features and perform feature grouping.

Chapter 3. The Binding Problem

3.1 Introduction

In order to perceive a unified representation of an object, the brain must be able to correctly associate all the different types of features (e.g. location, colour, texture, pitch, etc.) derived from that object. Such associations, or *bindings*, are even more important when the stimulus consists of more than one object, in which case the brain must avoid incorrectly associating features from different objects: *illusory conjunctions* (e.g. Treisman and Schmidt, 1982). Evidence from work on illusory

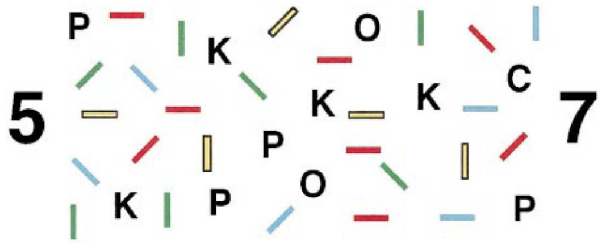


Figure 1 **Illusory conjunction demonstration. 1. There are two large numbers on the left and right of this figure. Determine if they are both odd and then read ahead for more instructions. 2. Without looking back at the figure, ask yourself if it contains the letters 'R', 'P' or 'Q'? Did you see a vertical yellow bar? Did you see a horizontal green bar? If you thought you saw an 'R', a 'Q' or a horizontal green bar, you have made an 'illusory conjunction'. From Wolfe and Cave (1999, figure 1).**

conjunctions (e.g. Wolfe and Cave, 1999) demonstrates that this task poses a problem for the brain. For example, figure 1 shows a number of coloured shapes and letters flanked by two black digits. Treisman and Schmidt (1982) instructed their subjects to report the two flanking digits. They found that the accuracy of reports were high, but reports concerning the objects between the digits (for which they were not forewarned about making) included a large number of illusory conjunctions. Such illusory conjunctions were found to involve all the features tested: size, colour, solidity and shape. Indeed, there is also significant evidence to show that even features within the same modality (such as auditory, olfactory, visual, etc.) can be encoded in widely distributed, spatially discontiguous, regions of the brain (e.g. Posner, 1980; van Essen and Maunsell, 1983; Damasio, 1985; Livingstone and Hubel, 1988).

It is this representational complexity which lies at the heart of the *binding problem*: how does the brain, confronted with many features, encoded in many different regions, draw them all together to form a perceptual whole?

Despite the importance of the binding problem in the fields of psychology, philosophy, neuroscience and computational modelling, the precise meaning of 'binding' is rarely made clear. Many different types of association fall under the general umbrella term of the binding problem. Most common is the binding of features within a modality: associating neural activities across cortical space. However, it is equally important to bind neural activities corresponding to features in other modalities such as to associate a sound with a visual object and possibly even a smell. In addition to these 'perceptual' bindings, binding is also used to refer

to the mechanism by which concepts are associated with a particular percept. For example, one binds the visual representation of an apple to all the semantic knowledge stored about it such as the fact that it is edible and how it tastes (Roskies, 1999).

The following sections introduce three broad classes of solution to the binding problem.

3.2 Combinatorial solutions

In the early 1970s a revolution in how the neuron was considered took place. The neuron had previously been thought of as a noisy indication of more basic and reliable processes involved in mental operations - the much higher reliability of the nervous system as a whole was explained by the supposed redundancy in neural circuits and averaging processes within the system. The advent of improved signal detection technology allowed physiologists to analyse the activity of single neurons and dispell this view. Neurons were no longer regarded as noisy indicators, but as the prime substrate of mental processes (e.g. Barlow, 1972).

With the evidence that the activity of a single neuron can play an important role in perception, new theories of brain function at the neuron level emerged. One popular proposal was that neural activity is organised hierarchically with progressively higher levels of processing being performed by increasingly fewer active neurons (Barlow, 1972). At the lowest level, neurons deal with the 'raw' sensory data. This information then converges on neurons with a higher level of perceptual *abstraction.* This continues until the activity of one neuron simply states the presence of a particular feature or pattern, thus allowing the binding problem to be avoided altogether. Using Barlow's example, the activity of a low-level neuron can be thought of as the occurrence of a letter, that of a high-level neuron being the occurrence of a word.

Within such a *combinatorial* framework, cells at higher processing levels are often more specific than those at lower levels: the upper levels of Barlow's hierarchy correspond to particular groups of features. An extreme example is that of the hypothetical *grandmother cell* (Barlow, 1972; see also Sherrington, 1941) which responds well to all views of grandmother's face. It has been argued that there are too few neurons to support all possible percepts (e.g. Engel *et al.*, 1992, 1997; Singer, 1993; Singer and Gray, 1995) and no site has been found which is large enough to accommodate the ultimate site of convergence (see Damasio, 1989): the

combinatorial explosion. However, this assumes that a strong interpretation of the hierarchical encoding mechanism is employed. Although our gaze will fall upon an almost infinite number of objects, the important question to be asked is not how many we will see, but how many we need to discriminate. If the perceptual system cannot distinguish between two objects, they need only have a single representation in the nervous system. Indeed, it has been estimated that people can distinguish between less than 100,000 different types of objects (Biederman, 1987) and untrained subjects cannot perceive differences between objects that are readily observable by trained subjects (Goldstone, 1998; see also Simons and Levin, 1997).

In support of a combinatorial framework, it has been found that early stages of visual cortical processing contain neurons that encode simple visual attributes whilst those at later stages encode more complex information about the visual scene (e.g. Van Essen and Gallant, 1994; Kobatake and Tanaka, 1994; Gallant *et al.*, 1996). Furthermore, there is evidence to suggest that neurons in the macaque inferotemporal cortex (IT) are tuned to views of complex objects such as faces (Bruce *et al.*, 1981). It is widely believed that the auditory system also exhibits such hierarchical processing stages (e.g. see Eggermont, 2001, for a review).

A criticism levelled at the framework is that there is no way to anticipate what different objects we will need to distinguish. In other words, there is no distinction between physically 'seeing' and cognitively 'identifying' (e.g. Kahneman *et al.*, 1992; Treisman, 1992). Novel (previously unseen or unimagined) objects are instantly seen and correctly bound even though we have no knowledge of them. For example, when viewing a previously undiscovered cell through an electron microscope, its features are immediately bound and the cell is perceived even though we have never seen it before and hence would not have a cardinal cell to represent it. A large 'reservoir' of uncommitted cells would be required for all the unseen objects which would have to maintain latent input connections from all feature-selective neurons at lower levels as well as consolidate the new perception instantaneously. Neurophysiological studies suggest that neurons can change their stimulus preferences to match the properties of a new object or feature allowing new discriminations to be made (Sakai and Miyashita, 1991, 1994; Logothetis and Pauls, 1995; Gibson and Maunsell, 1997; Kobatake *et al.*, 1998) although this would take finite time. However, novel objects are visible immediately. It has been suggested that temporal coherence (see below) could be used to disambiguate responses to stimuli before a combination-coding network for the stimulus in question has been formed (von der Malsburg, 1999) by synaptic plasticity (Konen and von der Malsburg, 1993). A further problem is how the brain stores the hierarchical structure of an object allowing a object's structure to be analysed. One possible mechanism is to encode the hierarchical path that leads to the cardinal cell

in question. However, if one begins to inspect activities at 'lower levels' in the processing hierarchy, the binding problem re-emerges: how are different low-level groups of neuronal activities made distinct?

Hierarchical representations may also conflict with how a subject draws experience from a particular event (von der Malsburg, 1999). How would the grandmother cell indicate that her face shares features with all other faces? Perceptions are not isolated; various aspects overlap giving a richness and relation to other perceptions which isolated events cannot convey. As one moves up the combinatorial hierarchy, the combinations become more complex and hence the context becomes more specialised. However, experience tends to encode information in a very general context, otherwise it cannot be exploited in other contexts without it being specifically learned in that context. For example, the *absolute* position of a pattern's constituent elements on a piece of paper is of little importance; it is the *relative* position of the pattern's elements that is of use (von der Malsburg, 1999).

Bruce *et al.* (1981) also found that the cells of the macaque responding specifically to faces also exhibited strong scale and translational invariance: changes in size and position of the facial image had no effect on the robustness of cell firing. Such invariance would allow the perceptual system to overcome the problem described above and allow it to store context-independent (to some degree) experiences. This experimental finding was used to enhance Hubel and Wiesel's (1965) original model of visual object representation and recognition to incorporate invariance (e.g. Perrett and Oram, 1993; Wallis and Rolls, 1997). This advance, however, complicated the hierarchical framework and the binding problem re-emerged.

3.3 Attentional solutions

Another class of solution to the binding problem employs attention to increase the saliency of certain features. Moran and Desimone (1985) found that when two stimuli appeared within the receptive field of a particular cell in either V4 or inferior temporal cortex, the response illicited from the cell depended on which of the two stimuli was being attended. The stimulus pair consisted of one which produced a strong response from the cell when presented alone (the *preferred* stimulus) and one which produced a weak response when presented alone (the *poor* stimulus). When attention was directed to the preferred stimulus, the cell responded strongly to the pair. However, when attention was directed toward the poor stimulus, the pair illicited a weak response, despite the preferred stimulus still being within the cell's receptive field. These findings are supported by studies of other

areas of the visual cortex (V2, V4, MT and MST) which showed that attention directed toward a preferred stimulus caused a higher cell response than when directed toward a poor stimulus (Treue and Maunsell, 1996; Luck *et al.*, 1997; Reynolds *et al.*, 1999).

From such findings, it was concluded that attention can modulate the receptive field size: when attention is directed to a particular stimulus within a cell's receptive field, this receptive field 'constricts' around the attended stimulus leaving the unattended stimulus (or stimuli) outside the receptive field. This effectively increases the spatial resolution of the visual system so that even neurons with multiple stimuli within their normally large receptive field can process information about a single stimulus at the attended location.

It ought to be noted that the ability of attention to modulate the neuronal response patterns of cells may be dependent on the processing stage in question and the task in hand: Seidemann and Newsome (1999) found similar attentional effects for an MT neuron as described above but their modulations were much weaker. Psychophysical evidence supports the theory that attention increases the spatial resolution of the visual system (Yeshurun and Carrasco, 1999; see also DeValois and DeValois, 1988; Graham, 1989) and fMRI studies have also provided support for attention modulation (Kastner *et al.*, 1998, 1999).

The studies outlined above suggest that attention has the ability to influence receptive field size. However, this says little about binding *per se*. Attention increases the saliency (represented by an increase in the neuronal response) of a particular stimulus and hence this increased neuronal response would allow the attended stimulus to win over an unattended stimulus in a competition for binding. In this 'strong' version of the framework, it is still unclear how a perceptual whole can be formed using this mechanism. When any scene is analysed (be it visual or auditory) binding produces many different grouped sets of features, only some of which are actually attended. Two difficulties arise: firstly, if an increase in neuronal response signifies a number of features which are to be bound, how are multiple, independent, groups of features to be bound without all the features forming a single, incorrectly bound, object? This implies that a form of *figure-ground* encoding is occurring. The figure-ground concept is a visual analogy in which there is one 'special' object which is the focus of one's attention and everything else falls into the background. However, the fact that the subjects in Treisman and Schmidt's (1982) study bound features into multiple objects (albeit sometimes incorrectly) demonstrates that more than two groups (equating to figure and ground) were present.

Secondly, if one sets aside the first objection, the framework implies that for any set of features to be bound, attention needs to be directed to each in turn in order to increase their neuronal responses. Hence, it is necessary for attention to be directed to all features within a scene in order to bind them into perceptual objects (no matter how loose that binding may be). This appears to conflict with the concept that attention is there to avoid such all-encompassing processing (see James, 1890).

The original hypothesis that attention is required to perform feature binding has since been relaxed in the light of challenges to the framework in which it has been shown that features can be correctly conjoined without attention and attention may not always assure correct binding (e.g. Tsal, 1989), hence resolving our second objection above. Indeed, it has been suggested that there exists a pre-attentive loose binding of features (Tsal, 1989; Cohen and Ivry, 1989; Prinzmetal and Keysar, 1989; Ashby *et al.*, 1996).

Illusory conjunctions (e.g. Treisman and Gelada, 1980) occur when bindings break down at some processing level. The role of attention is viewed not as conjoining pre-attentive features but allowing such binding failures to be rebuilt and / or maintained (Reynolds and Desimone, 1999). For example, extensions to Treisman's Feature Integration Theory (FIT; e.g. Treisman and Gelada, 1980) such as Guided Search (Wolfe *et al.*, 1989; Cave and Wolfe, 1990; Wolfe and Bennett, 1997) incorporate this loose bundling: features are not accurately bound to form objects in early 'pre-attentive' stages of processing and are merely held together on the basis of spatial proximity. It is only the deployment of attention that binds these features together into a representation that can be recognised (Nothdurft, 1993; Suzuki and Cavanagh, 1995).

Such a framework may be efficient at solving illusory conjunctions in which the spatial resolution of confusing boundaries can be increased. However, our first objection (see above) remains: how can simple increases in neuronal response allow the brain to form multiple bindings? It appears that another mechanism is still required to encode these multiple bound groups.

3.4 Temporal correlation solutions

As mentioned in section 3.2, the incorporation of invariance into the perceptual system to allow a degree of context-independent encoding, caused the binding problem to re-emerge in the combinatorial framework. Indeed,

> *'As generalisations are performed independently for each feature, information about neighbourhood relations and relative position, size and orientation is lost. This lack of information can lead to the inability to distinguish between patterns that are composed of the same set of invariant features.'*
>
> *von der Malsburg (1995, p. 523)*

An alternative mechanism of grouping to the hierarchical approach is based on the concept of an *assembly*: a large number of spatially distributed neurons (Hebb, 1949; Braitenburg, 1978; Edelman, 1978; Palm, 1981, 1990; von der Malsburg, 1986; Gerstein *et al*., 1989). The major advantage of the scheme over a hierarchical approach is the benefit of neuron 'overloading': an individual cell can participate in the representation of multiple perceptual objects. Thus assembly coding is relational because the significance of an individual neuron's response depends entirely on its context.

With a distributed representation it is necessary to be able to distinguish a neuron as belonging to one assembly or another. Therefore, the responses of related neurons must be labelled as such. It was proposed by von der Malsburg (1981; see also Milner, 1974) that the means of labelling different assemblies is by temporal synchronisation of the responses of assembly members. Segregation was hence expressed by the desynchronisation of different assembly responses. Thus, each assembly is identified as a group of synchronised neurons. The advantage of synchronisation is that the extra dimension of *phase* allows many simultaneous assemblies, each being desynchronised with the others.

However, the computational expense of evaluating synchrony between multiple spike trains is high; to alleviate this problem, von der Malsburg and Schneider (1986) proposed a mechanism in which the mean discharge response of a pool of cells is represented by an oscillator. In this manner, groups of features form streams if their oscillators are synchronised and the oscillations of additional streams desynchronise. Using this technique von der Malsburg and Schneider constructed a network of fully connected oscillators (E-cells), each receiving input from one frequency band of the auditory periphery and inhibition from an H-cell. In this framework, the global inhibitor simulates the thalamus which is known to have mutual connections with the cortex. Connections between E-cells can be modified on a fast timescale according to their degree of synchronisation. E-cells which receive simultaneous inputs synchronise through strengthened excitatory connections, and desynchronise with other cells due to inhibition. Hence, this model simulates stream segregation based upon onset synchrony.

Oscillatory activity in the brain was first observed 70 years ago from recordings made from the scalp. However, neural information was thought to be defined purely by amplitude and provenance. Hence, timing received little attention and was 'averaged out' of many studies. More recent work using multielectrode recordings to analyse internally generated covariations of firing pattern has shown that neurons can sychronise their discharges (Gray and Singer, 1989; Gray *et al.*, 1989). These studies also found that such stimulus-driven, context-dependent sychronisations were are associated with oscillatory modulation of cell firing in the γ frequency range (30-50 Hz). It was also noted that the observed sychronisation was internally generated and was not due to stimulus-locked changes in discharge rate; and the synchronisation probability changed in a systematic way in a stimulus/context-dependent manner.

Electroencephalogram (EEG) studies have also revealed prominent activity, especially in the β and γ frequency range. These so-called *40 Hz* oscillations proved to be one of the most widely recognised but least understood electrophysiological activities of the cerebral cortex. Barth and MacDonald (1996) reported that stimulation of the acoustic thalamus modulated cortex-based γ oscillations and suggest coupling of sensory processing between these cortical zones. A study by Joliot *et al.* (1994) confirmed that 40 Hz oscillatory activity was involved in human primary sensory processing and also suggested that it forms part of a solution to the binding problem. In their tests, one or two acoustic clicks were presented at varying times (3-30 ms interstimulus intervals) while a magnetoencephalograph (MEG) was used to study the auditory area of the brain. Analysis showed that at low interstimulus intervals (less than 12-15 ms) only one 40 Hz response was recorded and subjects reported only perceiving a single click. At longer intervals, each stimulus evoked its own 40 Hz response and listeners perceived two separate clicks.

The major problem associated with a temporal correlation framework is that it requires a high temporal precision of cell discharge which is exacerbated when such temporal precision is required over large cortical distances. However, cortical networks can operate with high temporal precision. Studies of single-cell response from the mammalian auditory cortex have shown pattern reproduction with millisecond accuracy from trial to trial (DeCharms *et al.*, 1998; Kilgard and Merzenich, 1998; see also Yu and Margoliash, 1996; Doupe, 1997). Furthermore, there is evidence that neurons distributed within and across cortical areas can synchronise their discharges on the basis of γ oscillations (Singer and Gray, 1995; see also Livingstone, 1996; Brecht *et al.*, 1998).

Long range temporal precision can be obtained by employing mechanisms such as *synfire chains* (Abeles, 1991; see also Shadlen and Newsome, 1998) in which

discharges are synchronised across parallel channels formed by cross-coupling parallel channels using converging and diverging axon collaterals (see below). Such a technique has been shown to perform well (Aertsen *et al.*, 1996; Diesmann *et al.*, 1997) due to the fact that it promotes highly synchronised excitatory postsynaptic potential (EPSP) barrages which are more effective than temporally distributed ones at triggering postsynaptic discharges. Thus, the temporal structure of the stimulus can be can transmitted over many synaptic junctions with little latency jitter.

Furthermore, it should be noted that perfect simultaneity would only be required if one neuron were to evaluate the object representation as a whole, which is not required in this framework. In this case, a certain degree of temporal 'drift' is acceptable as long as this happens in a continuous fashion across encoding regions (von der Malsburg, 1999).

It has been suggested that oscillatory behaviour is merely an emergent property of the experimental design: many studies use anaesthetised animals and hence oscillations may reflect a state of sleep rather than feature binding (e.g. Horikawa, 1994). However, oscillatory activity has been successfully observed in awake animals (Singer, 1993). Indeed, evidence that synchronised activity encodes salient information is supported by studies in which human subjects displayed a high correlation between perception and neuronal response synchronisation (Tononi and Edelman, 1998; Tononi *et al.*, 1998). Additionally, an engagement in cognitive tasks has also been found to increase γ oscillations in task-dependent cortical areas (e.g. Tallon-Baudry *et al.*, 1998; Keil *et al.*, 1999). Similar γ oscillation increases have also been found in states of focused attention (see Pulvermüller *et al.*, 1997).

3.5 Interim summary

In the previous sections, we have looked at three broad classes of solution to the binding problem, each of which has its advantages and disadvantages. Combinatorial coding frameworks (e.g. Barlow, 1972) allow the binding problem to be avoided altogether: every percept is represented by an individual 'cardinal' cell. Indeed, the commonly cited problem with this mechanism - insufficient numbers of neurons to encode all percepts - has been argued to be irrelevant if only *distinguishable* percepts are encoded (e.g. Biederman, 1987; Goldstone, 1998; see also Simons and Levin, 1997). However, the difficulty of rapidly encoding new percepts and analysing the hierarchical structure of objects may require a further segregation mechanism to be employed such as temporal correlation. Furthermore,

invariant object recognition (Bruce *et al.*, 1981) which has been incorporated into a number of models of visual recognition (e.g. Perrett and Oram, 1993; Wallis and Rolls, 1997) also causes problems for combination-coded representations. These solutions decompose the sensory pattern into elementary features each of which is represented by an individual cell connected to a single invariant feature cell. Therefore, the invariance cell receives the same feature from all different points on the generalisation continuum (e.g. location). However, at these higher levels, information about the relative position, size and orientation is lost (von der Malsburg, 1995) and the binding problem re-emerges. It has been suggested that temporal correlation could be used at a low processing level before generalisation has occurred. These 'spatial' patterns then resonate with and activate isomorphically arranged higher level circuits (von der Malsburg, 1981, 1988; Hummel and Biedermann, 1992).

There is also evidence that attention plays a role in solving the binding problem (e.g. Moran and Desimone, 1985; see also Pulvermüller *et al.*, 1997) by allowing groups of cells encoding particular features to be given increased saliency (by means of increased neuronal response) and hence bind them together. However, such effects have occasionally been found to be weak (Seidemann and Newsome, 1999) and indeed attention has also been found to be unnecessary in some situations (e.g. Tsal, 1989; see also Riesenhuber and Poggio, 1999). It is now believed that features are loosely bound pre-attentively and it is only the deployment of attention that binds these features together into a representation that can be recognised (Nothdurft, 1993; Suzuki and Cavanagh, 1995). As discussed above, it is unclear how multiple bound groups can be formed when the only way of signalling a to-be-bound group in increase neuronal response. One possible way of distinguishing between these groups is to use temporal coherence.

Temporal coherence was proposed by von der Malsburg (1981) to overcome the ambiguities that arise from the loss of information in combination-coded representations when feature invariance occurs. Indeed, it should be made clear that temporal coherence is *not* a solution to the binding problem *per se*, merely a way of representing the output of some solution. As mentioned above, such a framework lends itself to solving some of the fundamental problems associated with attentional- and combinatorial-based solutions.

To conclude, it is likely that a combination of all three mechanisms are used to solve the binding problem. Temporal coherence is useful for segregating many different groups (e.g. within an attentional framework, or an invariant hierarchical framework). Synchronised circuits and specialised conjunction units are compatible (Tanaka, 1996) and point toward both being used. Indeed, the existence of illusory

conjunctions (e.g. Treisman and Gelada, 1980) which can disappear given sufficient viewing time suggest that relevant combinations of features are not represented solely by combination-coding and that the brain requires time to form the correct bindings.

3.6 Computational models of feature binding

In this section, we will describe two computational models of binding based on temporal coherence. The first model - synfire chains - uses spike trains in which the correlation of the temporal fine structure of two spike trains is used to asses synchrony. The second model employs neural oscillators which demonstrate a more computationally efficient mechanism of assessing synchrony. In particular, we will look at how both models can be used to perform frequency proximity grouping of pure tones, since this relates to issues studied later in this thesis.

Synfire chains

As described above, a certain degree of long range temporal precision may be required to allow distantly separated brain regions to participate in synchronous activity. The simplest mechanism is to employ a single multi-neuron link (figure 2a) in which spikes travel from one neuron to another along a chain of neurons. However, if one neuron in the chain becomes damaged or dies, the entire chain becomes inoperative. This is significant because neurons are constantly dying and cannot be replaced - between the ages of twenty and eighty years, an average human loses one third of their cortical cells (Gerald *et al*., 1980) without significant loss in information processing ability. Therefore, some form of redundancy is required in such systems. The use of parallel serial chains (figure 2b) provides this. However, an inordinate number of neurons is required to maintain system functionality over an extended period (e.g. the life span of a human). A network of neurons connected using diverging and converging pathways (figure 2c) incorporates redundancy while limiting the number of neurons required. As mentioned above, a diverging and converging pathway topology also allows the temporal structure of a stimulus to be transmitted over many synaptic junctions with little latency jitter.

Abeles (1991) contends that information transmission in the cortex is likely to occur between sets of neurons connected by such diverging and converging pathways: *synfire* transmission (see also Griffith, 1963).

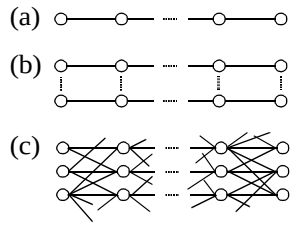


Figure 2 **Alternative connections between neurons. (a) serial; (b) parallel serial; (c) diverging/converging. After Abeles (1991), figure 6.1.1.**

For the pathways between two nodes to be a *synfire link*, the following conditions must hold,

- Whenever *n* cells of the sending node become synchronously active, at least *k* cells of the receiving node must become synchronously active.
- *k* must not be smaller than *n*.

For a network of neurons connected with diverging/converging (*feedforward*) pathways to be a *synfire chain*, the following condition must hold,

- All connections between nodes must be synfire links such that the receiving node of one link is also the sending node of the following link.

Our early work (Wrigley and Brown, 2000; see also Wrigley, 1999) considered if such a topology may also be used to explain frequency proximity grouping in the auditory system. The use of diverging projections with decreasing synaptic saliency from the central link (von der Malsburg, 1973; Hubel, 1988) appears to lend itself well to proximity segregation: more distant input (in terms of frequency separation) will experience much weaker synaptic influence from the other frequency band and hence will be less likely to produce synchronous activity. According to the temporal correlation theory, the activities of two frequency bands are said to be grouped if the output spike trains for each band display a high level of temporal correlation (figure 3).

Synfire chains can be used to demonstrate grouping by frequency proximity by ensuring that *k* is smaller than the node width. In other words, it must be possible for each group of *k* neurons to which a frequency band projects to be distinct. In this way, widely separated frequency bands each excite a group of *n* neurons whose projections to the next node do not overlap: without a significant overlap, the activities of the respective *k* neurons cannot influence each other and hence cannot synchronise.

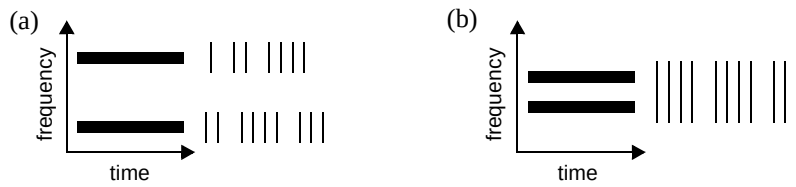


Figure 3 **Desired synfire chain network response to spectrally distant stimuli (a) and spectrally close stimuli (b).**

The network used to investigate this proposal consists of five synfire links, the first of which receives input from 50 frequency channels (figure 5). Each cell is modelled as an integrate and fire neuron (see Lapicque, 1907; Tuckwell, 1988) incorporating an absolute refractory period; input to the cell is in the form of spikes which are represented as binary values. The membrane potential E at time t is defined as

$$E_t = E_{t-1} e^{\frac{-T_s}{\mu}} + E_r \left(1 - e^{\frac{-T_s}{\mu}} \right) + H_t \qquad (1)$$

where E_r is the membrane resting potential (-60 mV); T_s is the sampling period (1 ms); μ is the decay time constant (10 ms); H_t is the increase in membrane potential due an incoming spike (8 mV) if an input spike occurred at time t, otherwise $H_t = 0$ mV. E_t becomes equal to the refractory membrane potential (-70 mV) for a period of 3 ms and a spike is produced if E_t exceeds the threshold potential (-50 mV). The behaviour of this type of neuron model is illustrated in figure 4.

Activation of each frequency band is simulated by random spike trains with an interspike interval of no less than 1 ms (absolute refractoriness). The diverging connections are subject to synaptic strengths which vary with distance in a gaussian fashion (von der Malsburg, 1973; Hubel, 1988). It is a combination of the feedforward network topology and the gaussian connection weighting that gives rise to grouping by frequency proximity. The n channels of the stimulated frequency band will project to k cells in the receiving node; therefore, for synchronous activity between the two bands to occur, the k receiving cells for each band must show a significant overlap.

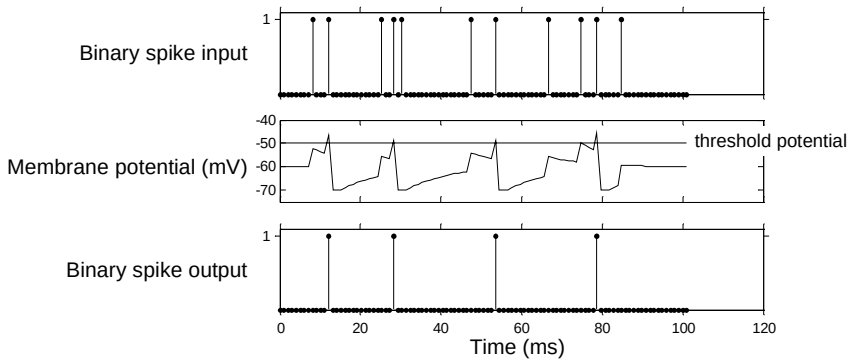


Figure 4 **Sample behaviour of the integrate and fire cell in response to a random spike input train. This cell model forms the basis of the synfire chain.**

The correlation of channel outputs X and Y is used to assess the degree of temporal synchrony between two channels,

$$C(X, Y) = \frac{\sum x(t)y(t)}{\sqrt{\sum x^2(t)y^2(t)}} \tag{2}$$

where $x(t) = X(t) - \langle X \rangle$ and $\langle X \rangle$ is the mean of $X(t)$.

Figure 6 shows that as frequency separation increases, the correlation (and hence the tendency to group) between the outputs of the centre frequencies decreases. It can also be seen that the n:k ratio (an indication of how many cells, k, a group of n cells projects to) determines the sensitivity of frequency proximity grouping: as the

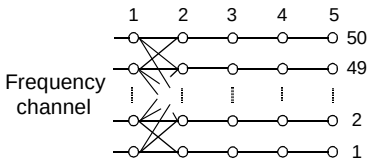


Figure 5 **The synfire chain used by Wrigley and Brown (2000) is 5 neurons long and receives input from 50 frequency channels. For reasons of clarity, only connections in the first synfire link have been shown.**

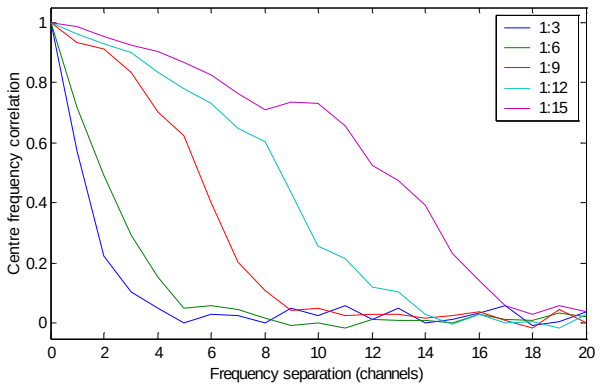


Figure 6 **Correlation between band centre frequencies decreases as frequency separation increases. The sensitivity of the network to frequency separation can be adjusted by altering the *n*:*k* ratio (legend shows *n*:*k* ratio).**

ratio tends toward 1, the network becomes more sensitive to the size of the frequency separation.

This section has described how a synfire chain (Abeles, 1991) can be used to perform grouping by frequency proximity. The 'spread' of the diverging connections within each synfire link determines the sensitivity of the network to frequency proximity: as the spread increases, widely separated bands are more likely to become synchronised and hence viewed as being grouped according to the temporal correlation theory (figure 6). However, when considering large networks of cells, the computational complexity of (2) becomes significant: to assess synchrony in the network output, (2) must be applied to every channel pair (adjacent or not): a problem with $O(n^2)$ complexity.

An alternative solution which draws upon the temporal correlation theory uses neural oscillators (e.g. von der Malsburg and Schneider, 1986) as opposed to computational models of spiking cell dynamics. This significantly reduces the complexity of synchrony assessment to $O(n)$: the application of a threshold to each channel output.

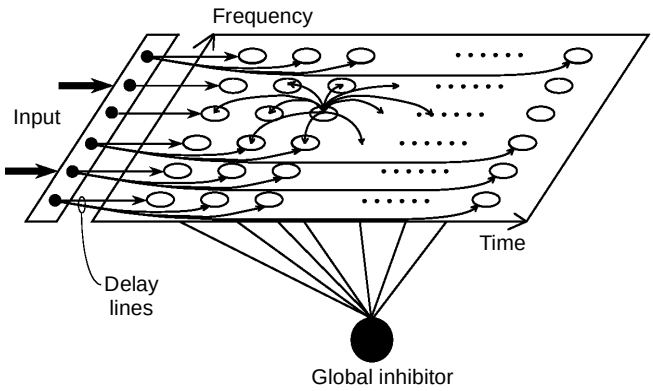


Figure 7 **Diagram of Wang's (1996) segregation network. To aid clarity, only the lateral connections from one oscillator in the network are shown. From Wang (1996).**

Neural oscillators

The previous section described the success of a synfire chain in segregating channel activities on the basis of frequency proximity. However, it also requires a mechanism for assessing synchrony within the outputs of such a model (equation 2) that is computationally expensive. As described above, von der Malsburg and Schneider (1986) proposed an alternative mechanism which extended the temporal correlation theory of von der Malsburg (1981) by using oscillators as the processing substrate. An oscillator is regarded as a model for the behaviour of a single neuron, or as a mean field approximation to a group of reciprocally connected excitatory and inhibitory neurons. Within this mechanism, the *phase* of an oscillator's activity can be used to assess synchrony, thus avoiding the need for a correlation metric: all oscillators whose activities are above a given threshold at time *t* are said to be synchronised. Hence, a set of features form a group if the corresponding oscillators oscillate in phase with zero phase lag (synchronisation); oscillators representing different groups oscillate out of phase (desynchronisation). Furthermore, by using an activity threshold, synchronisation can be determined using only the activities at time *t* whereas (2) requires a time series of previous cell activities necessitating some form of activity buffer for each channel.

Recent work by a number of researchers (Lui *et al.*, 1994; Wang, 1996; Brown and Cooke, 1997; Brown and Wang, 1999; Wang and Brown, 1999) has extended the oscillator-based stream segregation model with some success. The approach of Wang and colleagues uses a two-dimensional time-frequency network of relaxation

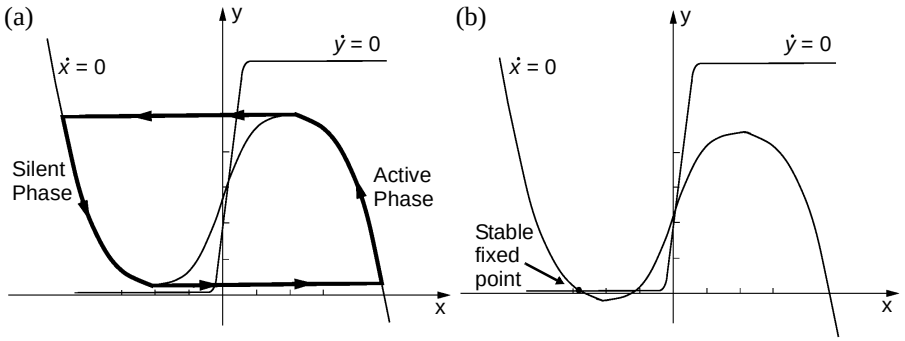


Figure 8 **Nullclines of a single oscillator. (a) The bold line shows the limit cycle of an *enabled* oscillator whose direction of motion is indicated by the arrow heads. (b) An *disabled* oscillator approaches a stable fixed point. From Wang (1996).**

oscillators in which lateral excitation connections promote synchrony and a global inhibitor aids desynchronisation (figure 7). This approach is based upon studies of locally excitatory globally inhibitory networks of relaxation oscillators (LEGION, see Terman and Wang, 1995) and has been termed the *oscillatory correlation framework* (Wang, 1996). This can be considered a special case of temporal correlation.

The building block of the network is a single oscillator, which consists of a reciprocally connected excitatory unit and inhibitory unit whose activities are represented by x and y respectively:

$$\dot{x} = 3x - x^3 + 2 - y + I + S + \rho \tag{3}$$

$$\dot{y} = \varepsilon\left[\gamma\left(1 + \tanh\frac{x}{\beta}\right) - y\right] \tag{4}$$

Here, ε, γ and β are parameters; S represents overall coupling from other oscillators in the network and ρ is a noise term used to aid desynchronisation between oscillator groups. The x-nullcline ($\dot{x} = 0$) is a cubic function and the y-nullcline ($\dot{y} = 0$) is a sigmoid function (see figure 8). When $I > 0$, the two nullclines intersect at a point along the middle branch of the cubic (see figure 8a) and give rise to a stable periodic orbit provided ε is sufficiently small. In this situation, the oscillator is said to be *enabled*. The solution of an enabled oscillator alternates between a phase of high x values (*active phase*) and a phase of low x values (*silent phase*); transitions between these two phases occur on a much faster time scale compared to time spent

in active and silent phases. When $I < 0$, the nullclines intersect on the left branch of the cubic (see figure 8b) and produce a stable fixed point at a low value of x. When the oscillator is in this state of equilibrium, it is said to be *disabled*. The parameter γ can be used to adjust the amount of time an oscillator spends in the two phases: a smaller value of γ results in a shorter active phase duration. It is clear, therefore, that oscillations are stimulus dependent: they are only observed when the external input to the oscillator is greater than zero. Because it has two timescales, the oscillator in (3) and (4) belongs to a family of relaxation oscillators. It is related to both the van der Pol oscillator and to the simplifications of the Hodgkin-Huxley equations for action potential generation in nerve membrane (van der Pol, 1926; Hodgkin and Huxley, 1952; FitzHugh 1961; Nagumo *et al.*, 1962).

Connections between oscillators are used to promote synchronous activity. The connection between oscillators i and j is in fact described by a pair of weights (Wang and Terman, 1995; Terman and Wang, 1995; see also von der Malsburg, 1981; von der Malsburg and Schneider, 1986) in which one is permanent (T_{ij}) and one is dynamic (J_{ij}). The strength of the T_{ij} connections on the grid fall off exponentially with the distance between them (figure 7) according to a two-dimensional Gaussian distribution. This endows the oscillator network with sensitivity to the frequency and temporal proximity of acoustic components. The strength of the J_{ij} connections are modified during simulation, depending on the state of synchronisation in the network, using *dynamic normalization* (Wang, 1993; 1995). This mechanism combines a Hebbian rule (Hebb, 1949) that emphasises coactivation of oscillators i and j and normalises all incoming connections to an oscillator.

The overall coupling S from other oscillators in the network to oscillator i is defined as:

$$S_i = \sum_j W_{ij} S_\infty(x_k, \theta_x) - W_z S_\infty(z, \theta_z) \qquad (5)$$

where W_{ij} is the combined T_{ij} and J_{ij} connection strength between oscillators i and j. x_k is the activity of oscillator k. The parameter θ_x is a threshold above which an oscillator can affect others in the network. Hence, the first term of (5) describes the lateral excitatory connections to oscillator i; the last term describes the inhibition from the global inhibitor. $S_\infty(x, \theta)$ is a 'squashing function' which compresses oscillator activity to be within a certain range:

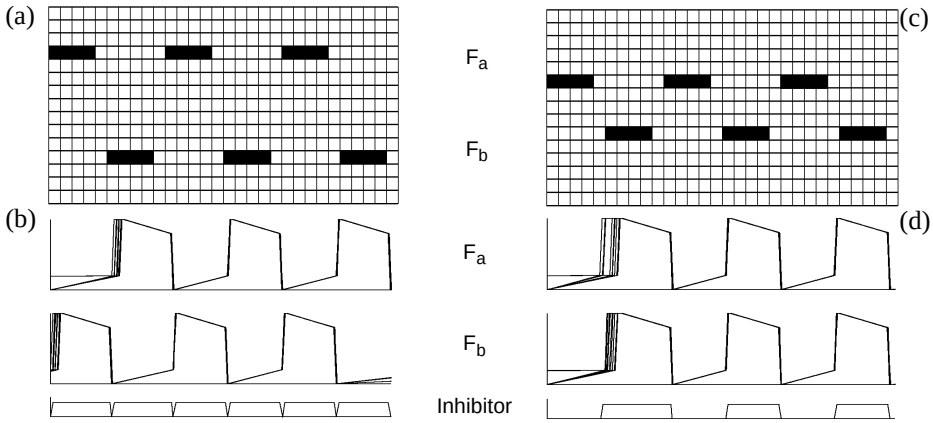


Figure 9 **Oscillator activities (b and d) in response to alternating tone sequences (a and c). From Wang (1996).**

$$S_\infty(x, \theta) = \frac{1}{1 + e^{-\kappa(x - \theta)}} \tag{6}$$

W_z in (5) is the weight of inhibition from the global inhibitor which is denoted by the unit z and is defined as:

$$\dot{z} = \phi(\sigma_\infty - z) \tag{7}$$

where $\sigma_\infty = 0$ if $x_i < \theta_z$ for each oscillator i, and $\sigma_\infty = 1$ if $x_i \geq \theta_z$ for at least one oscillator i. Hence θ_z is another threshold. If $\sigma_\infty = 1$, $z \rightarrow 1$ and every oscillator in the network receives inhibition; if $\sigma_\infty = 0$, $z \rightarrow 0$ and none of the oscillators in the network receives inhibition. ϕ determines the rate at which the global inhibitor reacts to stimulation.

The global inhibitor receives excitation from each oscillator, and in turn inhibits each oscillator of the network. Once a group of oscillators 'jump' up to the active phase, it triggers the global inhibitor, which then inhibits the entire network, thus suppressing the activity of other groups of oscillators. A group (and subsequently a stream) is formed by synchronizing oscillators in the two-dimensional time-frequency network.

As the 'frequency' of the global inhibitor activity in relation to that of the network oscillators is dictated by the total number of groups in the network, this activity also

forms a useful cue in determining how many groups exist and which oscillators belong to them.

Using this network, grouping is performed on a time-frequency pattern input: the network works on a pseudo-spectrogram with a time resolution of 40ms. It is hypothesised that the time axis is produced by a system of delay lines. When presented with binary input, the network quickly achieves a stable state in which groups of oscillators representing streams 'pop out' one after the other.

Wang (1996) investigated grouping by frequency proximity using an alternating tone sequence. Figures 9a,c show the pseudo-spectrogram diagrams of the stimuli conditions with large (figure 9a) and small (figure 9c) frequency separations as they would be mapped to the oscillator network. Each square in the figures represents the input to a single oscillator in the two-dimensional network.

Figures 9b,d show the oscillator activities for the stimulated frequency channels and the global inhibitor activity over the full duration of the stimulus. It is clear from the oscillator plots that when the frequency separation is small, the two sets of oscillators are synchronised (phase locked with zero phase lag - figure 9d). However, if the frequency separation is large, the oscillators are desynchronised (figure 9b) and the frequency bands are not considered to be grouped. In each of these plots (figures 9b,d) the activity of the global inhibitor is shown at the bottom. As noted above, the global inhibitor activity forms a useful cue in determining how many groups exist. When the two frequency bands are segregated (figure 9b) the frequency of the global inhibitor is twice that of the global inhibitor when the frequency bands are grouped (figure 9d). It is also important to note the simplicity of assessing synchronisation: oscillators which are synchronised jump up to the active phase and drop back down to the silent phase simultaneously. Furthermore, the global inhibitor ensures that only one group of synchronised oscillators are in that active phase at any one time. Hence the application of a threshold is sufficient to determine which oscillators are grouped: all oscillators in the active phase at the same time (x activities above a certain threshold) are to be grouped.

This model will be discussed further in later chapters in which the LEGION network is used as the basis for a model of auditory attention.

3.7 Summary

This chapter has presented three general classes of solution to the binding problem, each of which have their individual advantages and disadvantages. The combinatorial framework (e.g. Barlow, 1972) is based on the proposition that progressively higher levels of processing are performed by increasingly fewer active neurons. Ultimately, the highest processing level consists of a single neuron whose activity signifies the presence of a particular object in the scene. For example, the hypothetical 'grandmother cell' whose activity signals the presence of grandmother's face in the visual scene (see Lettvin, 1995). Despite refinements of the framework to incorporate a level of context-independent storage, these introduced difficulties in being able to distinguish between multiple 'invariant' feature sets. A similar problem arises with attentional solutions. This framework proposes that attention is used to increase the saliency (represented by an increase in the neuronal response of the feature-encoding cells) of features that are to be perceptually grouped. However, this mechanism suffers from the same problem as the invariant combinatorial framework: how are multiple grouped feature sets distinguished? To solve this problem, von der Malsburg (1981; see also Milner, 1974) proposed that the multiple groups of feature sets could be distinguished on the basis of the temporal fine structure of their neural responses: cells representing features which are to be grouped synchronise their activities; groups of synchronised cells representing different objects are desynchronised.

The final section of this chapter demonstrated two types of solution which draw on the temporal correlation theory of von der Malsburg (1981), and show that grouping by proximity (one of the Gestalt grouping cues described in the previous chapter) can be an emergent property of appropriately connected neural networks. Synfire chains (Abeles, 1991) demonstrate how synchronous spike trains can be transmitted over large distances with little temporal jitter. The simulation above demonstrated that they can also perform frequency proximity grouping by virtue of their diverging and converging pathways. However, the computational complexity of the synchrony assessment becomes significant: it has $O(n^2)$ complexity.

The use of oscillators (e.g. von der Malsburg and Schneider, 1986; Wang, 1996) within large networks avoids such issues. Firstly, the *phase* of an oscillator is used to assess synchrony. Within this mechanism, a set of features form a group if the corresponding oscillators oscillate in phase with zero phase lag (synchronisation); oscillators representing different groups oscillate out of phase (desynchronisation). As a result, many different groups can be represented simply by adjusting the phase of each group's oscillators (e.g. see Wang and Terman, 1995).

Secondly, the computationally expensive cross-correlation process required to evaluate network synchronisation in synfire chains can be replaced with a simple threshold and hence reduce the problem complexity to *O(n)*: all oscillators in the active phase at the same time (*x* activities above a certain threshold) are to be grouped. Of course, such considerations of computational cost may not be relevant when considering a highly parallel structure such as the brain.

The second example above showed how neural oscillators within the oscillatory correlation framework can perform grouping by frequency proximity (Wang, 1996). The two-dimensional network used by Wang (1996) consists of a grid of oscillators connected by two types of connection: permanent and dynamic. The permanent weighting between a pair of oscillators on the grid falls off exponentially with the distance between them, thus providing sensitivity to the frequency (and temporal) proximity of acoustic components. Conceptual issues related to the design of Wang's network and the implementation of an attentional mechanism presented in this paper will be discussed in more depth in chapters 5 and 6.

Chapter 4. Auditory Selective Attention

4.1 Introduction

'Everyone knows what attention is.'

William James (1890)

The term *attention* is commonly encountered in ordinary language and its meaning appears simple. Unlike other fields of research, people have strong convictions about its precise nature which haven't been arrived at by scientific investigation: it

is a fundamental part of daily life and therefore something about which one ought to know a great deal.

In common usage, attention usually refers to both selectivity and capacity limitation. It is widely accepted that conscious perception is selective and that perception encompasses only a small fraction of the information impinging upon the senses. The second phenomenon - that of capacity limitation - can be illustrated by the fact that two tasks when performed individually pose no problem; however, when they are attempted simultaneously, they become difficult. This occurs even when the two tasks are not physically incompatible such as reading a book and listening to the radio. In turn, this leads to the common conclusion that attention is a finite resource.

Awareness of stimuli details only occurs if they are attended to and the finite nature of attention leads to capacity limitation: when attending to one task there is less attention to devote to other tasks. Devotion of attention to one task is assumed to enhance performance (‘*pay attention to your driving*’) but can also be detrimental in some highly automatic tasks such as tying one's shoes.

Selectivity of auditory perception and the voluntary control of this selection are the core phenomena addressed in the remainder of this chapter. Specifically, we will look at how attention can be directed in frequency and location; evidence will be presented to suggest that this allocation occurs in the ‘shape’ of a gaussian distribution. Finally, we consider a psychophysical study that has cast doubt on the traditional role of attention and also look at how previous models of attention fail to explain findings of perceptual studies.

4.2 Attentional allocation in audition

Frequency location

It has been known for some time that listeners are better able to detect expected tones as opposed to unexpected tones. Greenberg and Larkin (1968) developed a probe-signal paradigm to assess the extent of this expectancy effect. Subjects were presented with two intervals, both filled with white noise, one of which contained a pure tone. Listeners were instructed to indicate which interval contained the tone. The subjects were led to expect the tone to be of a particular frequency (this signal was termed the *primary*). However, on less than a third of trials, the tone presented was of an unexpected frequency (this signal was termed the *probe*). Greenberg and

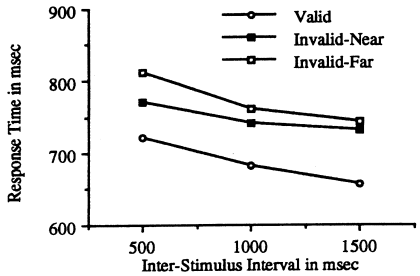


Figure 1 **Response time as a function of trial type and interstimulus interval. From Mondor and Bregman (1994), figure 1**

Larkin found that detection performance was best for primary signals, intermediate for probes within the critical band and worst for probes outside the critical band.

This form of experimental design has subsequently been used by a number of researchers (e.g. Schlauch and Hafter, 1991) to study the effect of expectancy on detection performance. In order to extend and substantiate such experiments, Mondor and Bregman (1994) investigated frequency selectivity within the context of an identification paradigm. Instead of embedding the tone in noise, the signal was presented in isolation and listeners were requested to indicate whether the target tone was longer or shorter in duration than the cue tone. On valid trials, the target and cue tones were of the same frequency. On invalid trials, the frequency separation of the two tones was manipulated to investigate the role of frequency similarity. In addition to this, Mondor and Bregman adjusted the interval between cue and target in order to determine whether the frequency selectivity effect depended on the time available to allocate attention to the cued frequency region.

Both validly and invalidly cued targets were equally likely to be one of three different frequencies. This ensured that the experiment would be able to determine whether superior performance on valid trials was due to differential familiarity with the target or allocation of attention to a cued frequency region.

Figure 1 shows that performance (indicated by the median time from target onset to listener response for each trial) declines as frequency separation increases. This implies that judgements about specific features of an auditory stimulus (in this case duration) may be facilitated (in this case made more rapidly) by orienting attention to the frequency at which the stimulus occurred. It can also be seen that increasing the duration of the cue-target interval improves performance suggesting that a finite

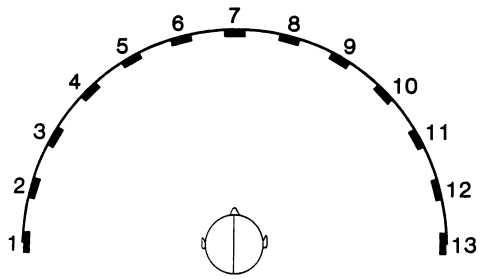


Figure 2 **Schematic description of the speaker array used in the experiments of Mondor and Zatorre. From Mondor and Zatorre (1995), figure 1.**

amount of time is required before attention is fully allocated to a particular frequency region.

Spatial location

As indicated above, cues that provide accurate frequency information lead to faster and more accurate target identification and classification. This is also true of cues to the spatial location of an acoustic target. Mondor and Zatorre (1995) examined whether the time required to perform a shift of attention was proportional to the physical distance of the shift. Each subject was placed at the centre of a semicircle of speakers (figure 2); a fixation sequence was used to control the focus of attention at the beginning of each trial: listeners were instructed to detect a drop in intensity of a steady tone presented from a particular location. Following this, a brief noise burst was delivered as a spatial cue from the spatial location from which the target tone would sound. The shift distance is defined as the spatial separation in degrees between the location of the fixation sequence and the location of the cue and target.

Despite all trials being *valid*, the cue-target interval was varied to control the amount of time available to orient attention. Figure 3a shows a similar trend to that of figure 1, in which performance increases as cue-target interval increases. However, it is important to note that the shift distance has no effect on performance. This is in direct contrast to evidence gathered by Rhodes (1987) in which she found that response times increased linearly with spatial separation up to a certain point, beyond which response times were similar. From this, Rhodes concluded that analogical shifts (in which shift time is linearly proportional to the shift distance) occurred for relatively short distances and discrete movements occurred for larger shifts. Despite accounting for her data, Rhodes' hypothesis is unique in combining the two types of shift, even when considered in the context of *visual* spatial

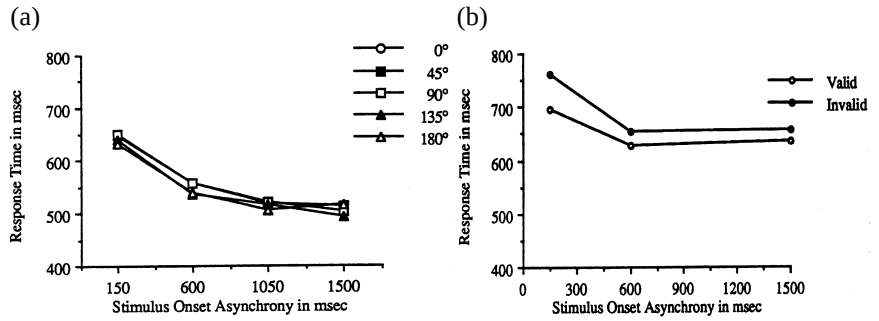


Figure 3 **(a) Response time as a function of attentional shift distance and cue-target onset interval. From Mondor and Zatorre (1995), figure 2. (b) Response time as a function of cue validity and cue-target onset interval. From Mondor and Zatorre (1995), figure 3.**

attention. Despite the mechanism for the shift of visual spatial attention having been a controversial topic of debate, Mondor and Zatorre remark that,

> *none of the investigators of visual attention has argued in favour of a model that incorporates both analogical and discrete movements*
>
> *Mondor and Zatorre (1995, p. 388).*

From this and the experimental findings of their more rigorous study, Mondor and Zatorre suggest that Rhodes misinterpreted her data, due to a correction procedure which failed to eliminate the effect of azimuthal position on localisation performance.

The most consistent explanation of Mondor and Zatorre's findings is that the cue is causing attention to be oriented to that location. However, one could also argue that the cue was simply acting as a general alert to the listener. In a minor alteration to the first experiment, Mondor and Zatorre introduced a number of inaccurate spatial cues on a portion of the trials. If an auditory cue acts solely to alert the listener, then both valid and invalid cues ought to result in identical performance. However, if auditory attention is oriented on the basis of the spatial cue, increased performance ought to be observed with valid cues. As expected, targets preceded by a valid cue were identified more quickly than those preceded by invalid cues. This provides strong evidence that listeners orient attention to the position in which the cue sounds.

In summary, Mondor and Zatorre (1995) found that performance improved as time available to shift attention to a cued spatial position increased. Furthermore,

accurate spatial cues facilitated performance more than inaccurate ones: performance declined as the distance of an unexpected target from a cue spatial location increased. Their evidence is also consistent with a discrete attention-allocation model.

Space-frequency allocation

Previous sections have demonstrated that the performance of target identification can be influenced by cues to the frequency or location of an imminent target. These results are taken to be evidence for the allocation of some form of auditory attention to single sites (Mondor and Bregman, 1994; Mondor and Zatorre, 1995). If a cue contained information from more than one modality, which would take precedence? Deutsch (1974; Deutsch and Roll, 1976) investigated this question by presenting listeners with a succession of pure tones dichotically. In one ear, an 800Hz tone was presented three times followed by two presentations of a 400Hz tone. In the other ear, a 400Hz tone was presented three times followed by two 800Hz tones. The particular frequency heard and the location from which that frequency apparently originated seemed to be governed by separate processes: the frequency heard was that presented to the listeners dominant ear and the location was that of the high tone. However, Bregman and Steiger (1980) argued that this *illusion* was likely to be a conflict of two perceptual organisation principles: grouping by frequency and grouping by location. It was suggested that this conflict was forcing the listeners' auditory system to call upon another, more reliable, form of grouping: the use of a higher harmonic to determine the location of a complex. It was unfortunate that Deutsch used an 800Hz tone which can be viewed as a harmonic of the lower, 400Hz tone. From this, Bregman and Steiger concluded that Deutsch's illusion was in fact the emergent behaviour of a preattentive process in which perceptual features are combined. Indeed, Bregman remarked,

> *The perceptual stream-forming process has the job of grouping those acoustic features that are likely to have arisen from the same physical source. Since it is profitable to attend to real qualities, locations, and so on, rather than to arbitrary sets of features, attention should be strongly biased toward listening to streams.*
>
> *Bregman (1990, p. 138)*

Mondor *et al.* (1998) investigated this interdependence of frequency and spatial information to determine if attention is directed at such streams. Using the same form of speaker array as their earlier experiments (Mondor and Zatorre, 1995), as shown in figure 2, listeners were given the task of categorising pure tones on the basis of frequency (low vs. high) or spatial location (central vs. peripheral). In

controlled conditions, no variation was made in the 'irrelevant' dimension. However, in the selective attention conditions, variations were made in the irrelevant dimension which were uncorrelated to variations in the relevant condition. If auditory attention is allocated separately to the location and frequency dimensions, then no performance degradation should be observed. Alternatively, if auditory attention cannot be allocated separately to the two dimensions, the performance will suffer interference from the variations in the irrelevant dimension.

Consistent with the notion that attention is directed toward streams, listeners found it impossible to ignore variation on an uninformative dimension while making classification judgements on the basis of a second dimension. Mondor *et al.* conclude that "*auditory attention acts to select streams*" (p. 68). When the relative salience of the frequency and location dimension was investigated, listeners were unable to guide selection independently by location or frequency. In other words, neither dimension dominates.

This suggests that auditory attention acts on groups of features rather than individual features and attending to different stimulus features which are integrated ought to result in similar cerebral activity. Zatorre *et al.* (1999) aimed to investigate this prediction by instructing listeners to perform a task which required detection of tones of a specified frequency or at a specified spatial location while undergoing positron emission tomography (PET) scanning. Their findings indicate that an auditory attentional task engages a specialised network of right-hemisphere regions, in particular the joint participation of the right parietal, frontal and temporal cortex. As expected, changes in the cerebral blood flow were very similar when subjects attended to spatial or to spectral features of the acoustic input supporting the model of Mondor *et al.* (1998) in which an initial stage of feature integration precedes selection on the basis of such streams.

Interdependencies have also been observed for pitch, timbre and loudness (Melara and Marks, 1990). This is in conflict with theories of selective attention which have arisen from visual experiments. The feature integration theory (FIT) of Treisman and colleagues (Treisman and Gelade, 1980; Treisman and Gormican, 1988) postulates that selective attention is required to perform a discrimination or detection task only when two or more features are in conjunction and not in variations of one feature. However, evidence collected by Mondor *et al.* (1998) suggests that classification cannot be based on a single feature. Furthermore, no evidence was found to support the dominant role of location present in the FIT.

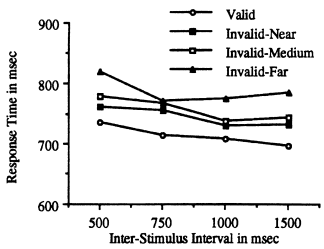


Figure 4 **Response time as a function of trial type and cue-target interval. From Mondor and Bregman (1994), figure 2.**

Attentional 'shape'

In the light of the evidence supporting the allocation of attentional resources to one or more sites simultaneously, it is interesting to consider the 'shape' of the attentional deployment. Two general classes of model have been proposed to describe the focus of attention. *Spotlight* models propose that attention is allocated to a discrete range of frequencies with an even distribution within this range. The edges of this spotlight are characterised by a sharp demarcation between attended and unattended frequencies. Alternatively, the attentional focus may be defined as a *gradient* with the density of the attentional resources being the greatest at the cued frequency and declining gradually with frequency separation from the focal point of attention.

In a similar experiment to that described above, Mondor and Bregman (1994) increased the number of possible frequency separations used on invalid trials to three. Not only is another strong cue validity effect observed (figure 4) but the effect of frequency separation is only consistent with a gradient of attention. As frequency separation increases, so too does the response time. A model incorporating a spotlight of attention with abrupt changes between attended and unattended frequencies could not account for this result. The gradient model of attentional allocation is also supported by evidence gathered by Mondor and Zatorre (1995).

Two forms of attention

It has been suggested that visual attention may be oriented by two different mechanisms which rely on differing amounts of conscious intervention by the listener. The *exogenous* system is considered to take place automatically under pure stimulus control: attention is drawn to the site of the stimulus. *Endogenous*

attention is considered to be under control of the listener, whereby attention can be consciously oriented to a particular site (Jonides and Yantis, 1988; Müller and Rabbitt, 1989). In other words, the exogenous system is engaged by peripheral cues. In contrast, the endogenous system is engaged by cues which have to be processed and interpreted before attention can be oriented, such as the selection of a single voice from a mixture. Spence and Driver (1994) have argued that these systems are also present in the allocation of auditory spatial attention. If this is true, the studies investigating frequency sensitivity (e.g. Mondor and Bregman, 1994; Schlauch and Hafter, 1991) and spatial sensitivity effects (e.g. Mondor and Zatorre, 1995) are, in fact, examining the allocation of endogenous attention.

Support for these two mechanisms can be found in the data collected by Hafter *et al*. (1993) in which the effectiveness of two types of cues for reducing frequency uncertainty was studied: *iconic* cues and *relative* cues. Iconic cues are those usually employed in the probe-signal methods described in previous sections. Relative cues were set to be two thirds the frequency of the expected signal - they acted as the symbolic cues which would stimulate the endogenous system. Hafter *et al.* found that both relative and iconic cues were successful in reducing the amount of uncertainty compared to the no-cue situation. However, it also emerged that the listening bands used with relative cues were wider than those measured for iconic cues by a factor of roughly 1.6. This suggests that the use of iconic and relative cues does indeed engage different mechanisms of attention.

4.3 Interim summary

The previous section has presented evidence that auditory attention is deployed in a number of interesting ways. Fundamentally, attention can be directed to a site of interest identified by some form of cuing (e.g. Greenberg and Larkin, 1968; Mondor and Bregman, 1994; Mondor and Zatorre, 1995). Furthermore, Mondor *et al*. (1998) have argued that there is an interdependence of frequency and spatial information consistent with the hypothesis that attention is directed at streams. Brain imaging studies (e.g. Zatorre *et al.*, 1999) have provided data consistent with this conclusion. These researchers have also shown that it is highly likely that the focus of attention can be described by a gradient model in which the density of the attentional resources is the greatest at the cued frequency and declines gradually with frequency separation from the focal point of attention. In addition to this, Jonides and Yantis (1988) and Müller and Rabbitt (1989) have argued that attention can be split into two mechanisms: unconscious and conscious allocation (exogenous and endogenous, respectively).

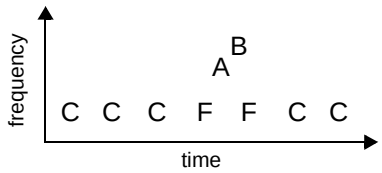


Figure 5 **Tone sequence used by Bregman and Rudnicky (1975).**

4.4 Selective attention in stream formation

We now turn to the relationship between attention and mechanisms of auditory grouping. Consider the cocktail party effect (Cherry, 1953) in which a listener has the task of following a conversation in a noisy environment. It is undoubtedly true that the process of selective attention is assisted by the speaker's voice having some acoustic properties which separate it from the other voices. Because these factors are similar to ones involved in primitive stream segregation - for example, differences in F0 - it can be argued that stream segregation is a form of selective attention. Bregman (1990) rejects this view; rather, he regards stream segregation as being largely the result of grouping by a pre-attentive mechanism. In support of this, Bregman cites an experiment by Bregman and Rudnicky (1975) in which the central part of the stimulus was a four tone pattern FABF (figure 5). Listeners were given the task of judging whether A and B formed an ascending or descending pair. In the absence of tones F, listeners found the task easy. However, in the presence of tones F, the pattern formed a single stream and the AB subpattern was found very difficult to extract. When a sequence of capturing tones C were included preceeding and following and F tones, they captured the latter into a new stream. Thus, tones A and B were separated into a different stream and their relative ordering was again found easy to judge. Bregman and Rudnicky argued that even though the stream of captor tones C was not attended to (listeners were concentrating on the occurrence of tones A and B) it was still able to capture tones F. The implication is that stream segregation can occur without attention.

Recent work by Carlyon *et al.* (2001) brings this theory into question. Their study aimed to manipulate attention more rigorously than Bregman and Rudnicky (1975) by presenting a tone sequence monaurally. When attention was to be oriented away from the tone sequence, subjects were required to perform a competing task in the contralateral ear. Specifically, a 21s sequence of A and B pure tones alternating in

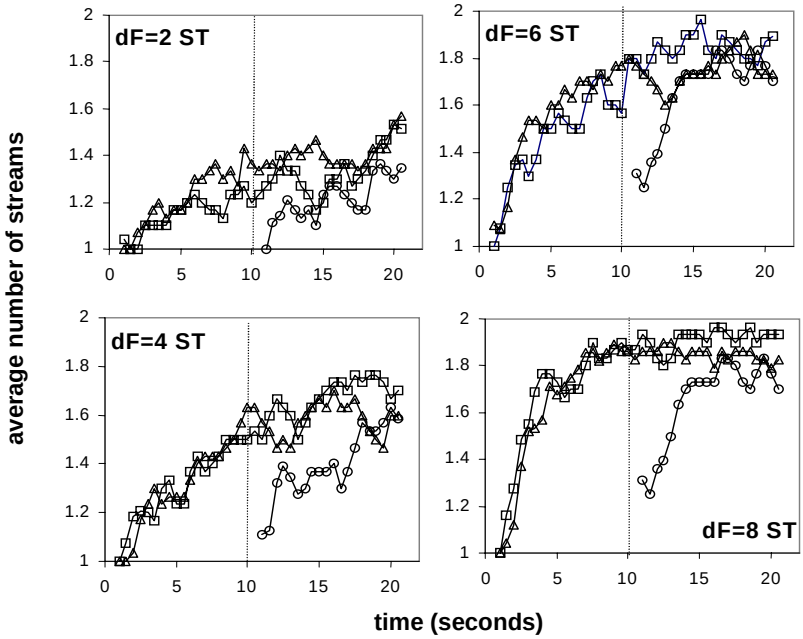


Figure 6 **Build up of streaming over time for four frequency differences. Scores are averaged across listeners and repetitions for the baseline (triangles), two-task (circles), and one-task-with-distractor (squares) conditions. From Carlyon *et al.* (2001), figure 3.**

an ABA-ABA sequence (e.g. van Noorden, 1975) was presented to the left ear. When the frequency separation of the tones is low or the repetition rate is low, the sequence is perceived as having a galloping rhythm. However, at higher repetition rates and larger separations, the sequence splits into two streams: one containing the high frequency tone and the other stream containing the low frequency tones. In this situation, the galloping rhythm is lost as only one of the streams is attended. In addition to this behaviour, there is a tendancy for the percept to build over time in such a way that at the beginning, listeners only perceive the galloping rhythm, whereas towards the end of the sequence, the rhythm is lost and only tone bursts of a single frequency stream are heard (Anstis and Saida, 1985).

In the 'baseline' condition of the Carlyon *et al*. experiments, no stimulus was presented to the right ear. Subjects were instructed to indicate whether they heard a galloping rhythm or two separate streams. In the 'two-task' condition, a series of bandpass filtered noise bursts were presented to the right ear for the first 10s of the

stimulus. The noise bursts were labelled as either *approaching* (linear increase in amplitude) or *departing* (the approaching burst reversed in time). For the initial 10s, subjects were instructed to ignore the tones in the left ear and simply concentrate on labelling the noise bursts

The labelling was achieved by pressing 'A' on a computer keyboard after each approaching noise burst, and key 'D' after each departing noise burst. After 10s the subjects switched their attention to the tone sequence. In the 'one-task-with-distractor' condition the noise bursts were presented to the right ear, as in the two-task condition, but subjects were told to ignore them and to perform the streaming task on the tones in the left ear throughout the 21s sequence. Consistent with Anstis and Saida (1985), subjects heard a single stream at the beginning of each sequence with an increased tendency to hear two streams as the sequence progressed in time. However, for the two-task condition the amount of streaming after ten seconds (during which period listeners had been concentrating on labelling the noise bursts) is similar to that at the beginning of the baseline sequence - in the absence of attention, streaming had not built up (figure 6). It can be argued that the Bregman and Rudnicky (1975) experiment was flawed as the listener did not have a competing attentional task to perform: despite the listener having been instructed to only concentrate on the A and B tones, there was no other task to distract the listeners attention from the C tones. Indeed, Carlyon *et al.* note that,

> *it seems likely that listeners were in fact attending to the C tones, as they were the only sounds present at the time, and there was no other task competing for attention.*
>
> *Carlyon et al. (2001, p. 115)*

Perception without attention

Many theories of visual perception assume that the extraction of groups of features occurs preattentively: object-based theories maintain that the visual scene is first parsed in accordance with Gestalt principles and then attention is directed to the perceptual objects that result from the parsing process (e.g. Duncan, 1984). In contrast with this view, recent work (e.g. Mack *et al.*, 1992) has suggested that little grouping, if any, occurs preattentively. Mack *et al.* developed an experimental method to investigate what can be perceived under conditions of inattention (the stimuli are within a person's visual field but no attention has been directed toward them). Subjects were presented with a difficult perceptual task, such as identifying the longest arm on a briefly presented cross, superimposed on a background of coloured dots. On the majority of trials, these dots were randomly black or white. On one trial, however, these dots would form a salient pattern if grouping occurred.

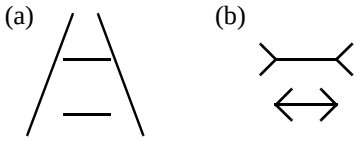


Figure 7 **Two perceptual illusions. (a) The Ponzo illusion: the line segment closer to the converging lines appears longer than the identical other. (b) The Müller-Lyer illusion: the line segment with arrowheads that point in appears longer than the identical line segment with arrowheads that point out. Redrawn from Moore and Egeth (1997) figure 2.**

At the end of the experiment, subjects were unexpectedly asked to make a forced choice decision about the background pattern. If grouping occurs without attention, then despite not attending to the background, subjects would still be able to report the pattern that had occurred.

The results from Mack *et al.*'s study show that accuracy was at chance, with a number of participants even denying that a pattern had even occurred. This is in contrast with traditional object-based theories which would have predicted a high pattern identification accuracy.

Although the subjects were unable to report the pattern, this does not imply that grouping did not occur. A possibility is that the subject may not be able to remember the pattern: the perception of the background on the critical trial may not have been encoded into memory. To control for this factor, Moore and Egeth (1997) employed a difficult line-length discrimination task on two horizontally oriented line segments superimposed on a background of dots which occasionally formed a pattern.

If grouped, these patterns in the background dots could influence the perceived lengths of the lines such as in the Ponzo and Müller-Lyer illusions (figure 7). The stimuli were arranged such that on trials in which the background did form a pattern, the lengths of the horizontal line segments were identical. If the background dots were not grouped, the discrimination task performance would be at chance. However, if the dots were grouped, this ought to influence the perceptual task and the line nearest the convergence (for the Ponzo illusion) or the inward pointing arrowheads (Müller-Lyer illusion) would be reported as longest more often than chance. Indeed, the latter was observed: the background pattern influenced the length discrimination task. Furthermore, at the end of the experiment, the subjects were asked if they saw a pattern and were given a forced choice decision over which pattern they saw. A negligible number of participants reported observing a

pattern and the identification accuracy was chance. This strongly suggests that grouping did occur during the experiment but that the patterns were either forgotten or never encoded into memory.

In summary, Mack *et al.* (1992) suggested that even salient grouping patterns are not perceived when not directly attended. Moore and Egeth (1997) extend this finding by showing that grouping of the unattended stimuli does occur: the grouped patterns influence the perceptual task. However, such preattentive grouping cannot be later reported. Moore and Egeth suggest that

> *attention [may be] required not for perceptual organisation but for encoding the results of that organisation in memory*
>
> *Moore and Egeth (1997, p. 350)*

Support for preattentive grouping and/or analysis can be found in the mismatch negativity (MMN) studies of Sussman and colleagues (Sussman *et al.*, 1998, 1999; Sussman and Winkler, 2001). The MMN is a component of event-related potentials (ERPs) which provides information about preattentive auditory processing. It is believed that the MMN is the outcome of a comparison process when the incoming stimulus differs from the memory of the stimulus in the recent past. It is considered preattentive because attention is not required to elicit a response. Sussman *et al.* (1999) presented listeners with an AB sequence (similar to that used by Carlyon *et al.*, 2001) to investigate the effect of attention on two tone streaming (Bregman and Campbell, 1971; van Noorden, 1975). In this well-studied phenomenon, listeners are more likely to hear two streams when the tones are presented at a high rate (one stream being A-A-A and the other being B-B-B). At low rates, streaming fails to occur and subjects continue to hear the AB sequence. Sussman *et al.* reasoned that if grouping occurs preattentively, it may be possible to illicit a MMN response to a deviant which could only be perceived if streaming had occurred. Furthermore, no MMN ought to be observed in the slow presentation situation. Figure 8 shows the standard AB sequence and the deviants predicted to illicit a MMN response. In all cases, the subjects performed another task and were told to ignore the stimuli. As expected, when the tones were presented at the fast pace, MMNs were detected in response to the deviant sequences occurring in both the high and low tones. When the tones were presented at the slow pace, no MMNs occurred for either the high or low tones. From this, Sussman *et al.* concluded that the streaming effect occurred automatically, (independently of attention) at, or before, the level of the MMN system.

Further work on these stimuli (Sussman *et al.*, 1998) suggested that attention could even force streaming to occur in certain occurrences of the slow pace situation.

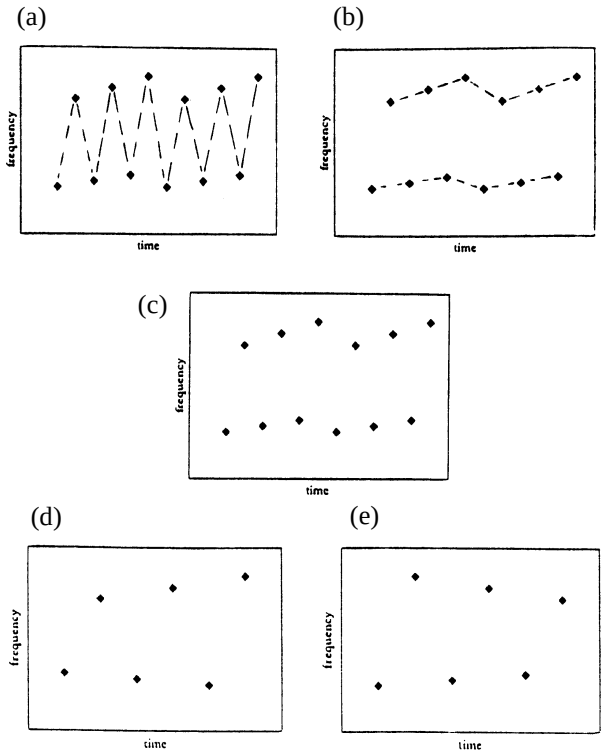


Figure 8 **(a) Perception of the AB sequence under slow presentation; (b) perception of the same sequence under rapid presentation. (c) The standard AB cycle of tones. Note both high and low tones exhibit a rising frequency trend. (d) and (e) Deviants to the standard cycle in the low tones and high tones respectively. Adapted from Sussman *et al.* (1999) figure 1.**

Listeners were instructed to attend to only the high tones and signal when they detected a deviant sequence. The results of their study show listeners could keep track of the standard three tone pattern within the high tones by employing highly focused attention. Additionally, MMNs were detected for both the attended (high) tones and the unattended (low) tones suggesting that selective attention can alter the organisation of sensory input.

The observance of a MMN in the unattended stream suggests that all groupings, whether attended or not, are subject to difference analysis. This finding is support by further MMN work by Sussman (Sussman and Winkler, 2001) in which MMN responses were detected for changes in an unattended auditory stimulus.

4.5 Interim summary

Caution should be exercised with respect to the extent to which attention was diverted away from the tone sequences in Sussman *et al.*'s (1998, 1999) experiments. In the situation when subjects were instructed to ignore the stimuli, the distracting task was to read a book - a passive, visual exercise. It was nevertheless the case that the tones were the only sounds present in the experiment. Duncan *et al.* (1997) have suggested that the visual and auditory attentional systems are largely independent and so one could speculate that some degree of attention was still being directed toward the auditory stimuli. The observance of streaming by Sussman *et al.* (1998, 1999) in a situation of inattention is in contrast to that of Carlyon *et al.* (2001) in which a more rigorous distractor mechanism was employed prevented streaming from occurring. Finally, despite the claim that mismatch negativity responses provides information about preattentive processing, it is interesting to note that Sussman *et al.* (1998) have concluded that attention can influence MMN responses (see also Alain and Woods, 1997; Trejo *et al.*, 1995).

4.6 Computational models of ASA incorporating attention

Few computational models of ASA incorporate attentional effects and those that do adhere to the traditional view that attention is used to select a pre-formed stream. In this section we will look at two functional CASA models which incorporate attentional mechanisms. Finally, we will look at two neural oscillator-based solutions, one of which demonstrates how attention is commonly viewed as a separate extension to the model and one in which attention has been incorporated into an existing framework resulting in unrealistic properties. We argue that these models have unrealistic properties and are incompatible with the findings reviewed previously in this chapter.

Functional approaches

Beauvois and Meddis (1991; 1996) contend that perceptual principles could prove to be the emergent properties of a simple low-level system. Their system is aimed specifically at the two-tone streaming problem and is intended to provide an explanation for two general principles: the perceptual accentuation of the attended stream and the apparently spontaneous shifts in attention between streams. These were investigated using a three-channel model with two centre frequencies at the tone frequencies and the other at their geometric mean. Noise is added to the output

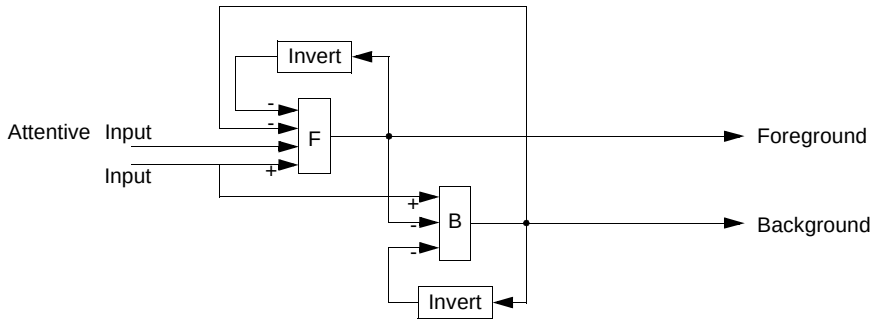


Figure 9 **Diagram of the McCabe and Denham (1997) network showing the connectivity of the streaming arrays F (foreground array) and B (background array). Excitatory and inhibitory connections are indicated by + and - symbols respectively. Redrawn from McCabe and Denham (1997).**

of the hair cell model for each channel in proportion to its activity. This is then used as the input to a leaky integrator. Finally, the dominant channel is selected and the activities of the other two channels are attenuated by 50%. The decision between streaming and temporal coherence is made on the basis of the ratio of activity in the tone channels: equal activity signifies temporal coherence, otherwise streaming.

Beauvois and Meddis showed that temporal coherence occurs when tone repetition times (TRT) are low due to the inability of the system to generate a *random walk*: long periods of silence prevent the build up of activity-related noise input. In this case the tone channels have equal activity. However, when the TRT is high, the random noise bias has little time to decay and so random walks are more likely and so, in turn, is the occurrence of streaming. Temporal coherence will also occur when the tone frequency difference is low due to the overlap of channel activation causing each tone to stimulate both its own filter and that of the other tone. In this case, the activities are equal. When the frequency difference is large, the combination of attenuation and random walk makes streaming more likely.

Despite the relative simplicity of the model, it is shown to behave consistently with a range of phenomena including grouping by frequency and temporal proximity as well as demonstrating the build up of streaming over time (Anstis and Saida, 1985). However, the model cannot simulate cross-channel grouping phenomena.

The model of Beauvois and Meddis (1991) was used as a starting point for the multichannel streaming model of McCabe and Denham (1997). Instead of using attenuation of the non-dominant channel to produce streaming, McCabe and

Denham employ inhibitory feedback signals which produce inhibition related to frequency proximity. The model also proposes that streaming occurs as a result of spectral associations and so the input to the system is represented by a multi-modal Gaussian rather than temporal fine structure as in the Beauvois and Meddis model. The model consists of two interacting arrays of neurons: a foreground array and a background array (figure 9). These terms are simply used for convenience as the system is symmetrically connected. Each array receives the excitatory tonotopic gaussian input pattern. In addition to this, the foreground array receives inhibitory input reflecting the activity of the background array and the inverse of the foreground activity. The background array receives similar inhibition. The inhibitory input to each array serves to suppress responses to those frequencies that the second array is responding to and also to suppress weak responses from itself. The streaming / temporal coherence decision is based upon the correlation between the output of the foreground array and that input. A high correlation to an input tone will mean that the tone is also present in the foreground array response. If successive tones elicit similar responses then the signal is said to be coherent; if one tone elicits a much larger response than another then streaming has occurred.

The interplay of frequency dependent inhibition and the time course of previous array activity successfully produces the two tone streaming effect and produced a good match to experimental data. Although an ‘attentional input’ was included in the model architecture (see figure 9), the authors acknowledge that the role of attention was not addressed in the model processing and remark that the influence of schema-driven grouping should not be ignored.

Neural oscillator-based approaches

In the neural oscillator model of Brown and Cooke (1997), the last stage is an attentional mechanism inspired by Crick’s (1984) proposal for an attentional ‘searchlight’. This was proposed in order to simulate the responses of thalamic cells which, when firing in synchrony with the oscillators of a particular group, caused that group to become the ‘attentional foreground’. However, Brown and Cooke concede that,

> *an attentional mechanism is not currently implemented in our computer model.*
>
> *Brown and Cooke (1997, p. 92)*

Instead, a correlation metric of the form used in the synfire example of the previous chapter is employed, thus incorporating the same limitations of complexity as described there.

Wang (1996, see also previous chapter) also proposes an attentional mechanism ('shifting synchronisation theory') as part of his network. He assumes that,

> *attention is paid to a stream when its constituent oscillators reach their active phases.*
>
> *Wang (1996, p. 444)*

Therefore, attention quickly alternates between each stream. This theory introduces unrealistic elements to the model. Wang has proposed that stream multiplexing occurs and that all streams are perceived equally at all times. This contradicts experimental findings (see Bregman, 1990) which show that listeners perceive one stream as dominant. When the two tone stimulus, described above, is presented in such a way as to promote streaming, the low tones *or* the high tones dominate. Furthermore, this theory cannot explain how attention may be redirected by a sudden stimulus. Such an event would be encoded by Wang's network as an individual stream which would be multiplexed as normal - with no attentional emphasis. The shifting synchronisation theory lies on top of the grouping layer in the network and thus operates on a two dimensional time-frequency snapshot. Therefore, attention alternates quickly between each stream in the entire snapshot in turn. Hence this forces attention to switch from a stream at the beginning of the snapshot to one at the end (later in time) and back again. This is a further consequence of the confusion in Wang's model between 'oscillator time' and 'real time'. Attention exists in real time - but in Wang's system it shifts in oscillator time.

Wang (2001) recognises that such an attentional mechanism is unrealistic when he notes that the '*treatment of auditory attention in [the] 1996 article is too preliminary*'. In later work, Wang (e.g. 1999) employs a winner take all (WTA) network topology to simulate visual selective attention. In this framework, an input pattern produces only one active neuron (or group of neurons) by means of either global inhibition or mutual inhibitory connections. However, previous WTA neural network theories of visual selective attention have not captured object-level selection: only a winning pixel (or location) is found (e.g. Koch and Ullman, 1985; Niebur, Koch and Rosin, 1993; Niebur and Koch, 1996). Wang (1999) uses a similar network to that described in the previous chapter (Wang, 1996) with the addition of a second, slower, global inhibitor. In response to a visual scene, the network inhibits all regions but the largest; in other words, the largest block sets the level of slow inhibition, which can then only be overcome by that block. Within this framework, only object selection based on the size of an object is possible, despite this only being one type of stimulus saliency (Treisman and Gormican, 1988; Desimone and Duncan, 1995). The ease with which other features, such as motion

or colour, could be parsimoniously incorporated into the selection mechanism remains to be seen.

4.7 Attentional allocation in vision

In the previous section, we saw how a computational model originally created for visual scene analysis (Wang, 1996) was adapted for an auditory scene analysis task. Before looking at a conceptual model of auditory attentional effects, it is interesting to look at the development of attentional research in the visual domain and highlight a number of contrasting concepts between visual and auditory processing.

Psychophysics

The *spotlight* concept has often been used to describe the function of visual attention: a selected region of the image is selected in a similar way to a torch illuminating an object at night. It is important to note, however, that this spotlight is not the same as visual saccades: the focus of attention can be moved independently of saccades. Posner (1980) found that signal detection efficiency increased when subjects were instructed to attend to a particular region of the visual field, whilst keeping the position of the eyes fixed. Furthermore, Posner found that a similar increase in performance was observed when the signals were presented at locations to which the subject was about to move their eyes, implying that a shift of attention precedes a saccade. As described above, auditory attention can be considered to be split into two forms: endogenous (conscious) and exogenous (subconscious). Evidence has been presented which suggests a similar split in visual attention in which one component is driven involuntarily (e.g. by the flickering of a light) and the other is under conscious control (Nakayama and Mackeben, 1989). Hence, it is an endogenous form of attention which is being oriented in Posner's study. A further analogue between auditory and visual attention can be made. In the previous section, evidence was presented which suggested that attention was required for detailed information about a auditory stimulus to be perceived. Similarly, O'Regan *et al.* (1999) found that dramatic changes to a visual scene being inspected by a subject may go unnoticed unless those changes occur at a location being attended by the subject.

Visual search and pop-out studies have been used to investigate bottom-up (stimulus-driven) processing of signals. In these studies, the observer is instructed to locate an odd target embedded within an array of distracting stimuli (Treisman and Gelade, 1980). The *conjunctive search* experiments of Treisman and Gelade

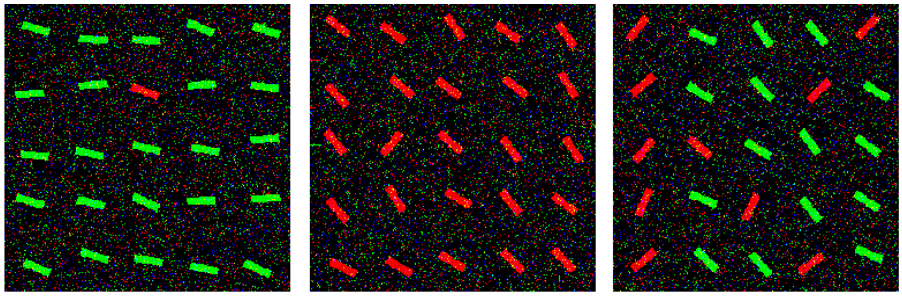


Figure 10 **Pop-out (left and middle) tasks and a conjunctive task (right) of the type pioneered by Treisman and Gelade (1980). Adapted from Itti and Koch (2000), figure 5.**

differentiated between the target and the distractors by an alteration in the conjunction of the constituent features. For example, one task may involve locating a vertical bar within an array of horizontal bars. Similarly, the distinguishing feature may be colour: identifying a red bar in the midst of identically oriented green bars (see figure 10, left panel). From these experiments, it was found that targets which differed by a single feature (such as colour) 'popped out' after a period of time independent of the number of distractors. However, when the target was only distinguishable by a combination of features (such as colour *and* orientation - see figure 10, right panel), the time required to locate the target bar increased linearly with the number of distractors. Treisman and Gelade (1980) concluded that there was a processing dichotomy in which single feature differences could be processed rapidly, in parallel, whereas conjunction searches required a slower, serial search. It was later shown that these were simply two extremes of a processing continuum based on task difficulty (see Treisman, 1988; Wolfe, 1996). However, the fundamental conclusion remained: a target which differentiates itself strongly from the surrounding distractors, can immediately attract attention to itself.

Computational models

The visual search experiments and the Feature Integrative Theory (FIT) of Treisman and Gelade (1980) formed the basis of many computational models of visual attention in which simple visual features (such as contrast, colour and orientation) are processed rapidly in parallel while attention is necessary to bind conjunctions of these features. Many models of visual attention commonly incorporate a saliency map. A saliency map is a two-dimensional topographical map which encodes stimulus saliency at every point in the incoming image. It is on the basis of this map that decisions are made as to where to orient attention. This

structure formed the basis of Koch and Ullman's (1985) computational model which was closely related to FIT. This was later developed by Itti and Koch (2000) in which all the two-dimensional topographical feature maps converge on a single saliency map. The selection of the most salient area in the image is achieved by a winner take all (WTA) neural network. To prevent a single area of the map permanently being the focus of attention, an inhibitory mechanism is incorporated which slowly suppresses the currently attended location leading to attentional scanpaths (Koch and Ullman, 1985; Itti and Koch, 2000). One danger of saliency maps is that the large number of different types of saliency projecting to a single map can make it as noisy as the original scene. An alternative to a single saliency has been proposed in which feature saliency can be encoded implicitly by modulations in each feature map (Desimone and Duncan, 1995; see also Wolfe, 1996). In this case, attentional selection occurs by top-down weighting of feature maps which are relevant to the current task.

Once a decision has been made regarding the location to which visual attention is to be oriented, a mechanism which 'routes' information flow to further processing stages is required. A model for such routing was proposed by Olshausen *et al*. (1993) in which *control neurons* (driven by a form of saliency map) modify the synaptic strengths of attended regions to allow neural responses to propagate to higher cortical areas.

Auditory and visual attention

From this brief review, it is important to note a number of factors which differentiate visual processing from auditory processing. Firstly, visual processing acts on a two-dimensional representation of the visual scene. In general, any object which appears in the visual scene is a contiguous whole; the task of binding and selecting such an object involves processing that contiguous area. It is from this property that the *spotlight* term originates: the beam of light used to illuminate an object is contiguous. However, auditory processing requires potentially widely separated discontiguous frequency regions (such as harmonics and/or fricatives) to be bound into a single object and attended. This problem of processing separated elements within a single feature dimension is exacerbated when the time-course of a stimulus is considered. In general, a visual object is observed in its entirety instantaneously. However, an auditory event requires a finite amount of time to be heard from beginning to end. The difficulty in applying visual processing and attentional principles to audition is highlighted when visual models are directly applied to auditory scene analysis and auditory attention. For example, a two-dimensional topology is commonly used in which the two spatial dimensions are replaced with 'frequency' and 'time'. Although such models have had success in

analysing the scene (e.g. Cooke, 1991/1993; Brown, 1992; Ellis, 1996; Wang, 1996; Wang and Brown, 1999) few have modelled attention. Those that have (e.g. Wang, 1996) employ visual mechanisms similar to attentional scanpaths in visual attention (Koch and Ullman, 1985; Itti and Koch, 2001) in which attention cycles through every 'object' in the two dimensional (time-frequency) scene. As described above, this mechanism does not realistically model auditory attention.

4.8 Summary

This chapter has presented a number of fundamental properties of attentional allocation. Firstly, single modality studies have produced evidence to suggest that attention can be allocated to one or more spatial locations or frequency regions; such allocation is believed to occur in the form of a gaussian distribution. When investigating more complex stimuli, it was shown that attention acts to select particular groups rather than individual features (Mondor *et al*., 1998). Secondly, evidence has been presented which suggests a certain degree of pre-attentive (exogenous) grouping of features occurs. Indeed, it is suggested that endogenous attention is required not only for stream selection and formation, but also for encoding a percept into memory: only groups which are attended are encoded into memory in detail and hence perceived (Moore and Egeth, 1997).

As described above, few models of computational auditory scene analysis incorporate the attentional findings presented in this chapter in a plausible manner, and those that use visual attentional mechanisms as their foundation produce unrealistic properties. In the next chapter, we present a conceptual framework in which attention plays a significant role in the stream formation and selection process. This will be used as the basis for a computational model which demonstrates a tight integration between attentional mechanisms and perceptual grouping.

Chapter 5. A Conceptual Framework for Auditory Selective Attention

5.1 Introduction

The studies presented in the previous chapter highlight a number of experimental factors which must be emulated by a conceptual model of auditory attention. Most important is the finding by Carlyon *et al.* (2001) that the distraction of auditory attention prevents the percept of streaming from occurring: stream formation requires attention, at least for alternating tone sequences. Furthermore, it has been suggested that attention can be engaged by two different mechanisms: the

exogenous (subconscious) and *endogenous* (conscious) systems (Jonides and Yantis, 1988; Müller and Rabbitt, 1989; see also James, 1890). These two types of attention form the basis of our model: exogenous attention accounts for subconscious (preattentive[1]) processing and the endogenous system controls which organisations are perceived and/or processed further. Before describing the new framework, an overview of previous theories of attention will be presented together with a number of studies which highlight their deficiencies.

5.2 Theories of selective attention

As noted in the previous chapter, attention is conventionally assumed to act as a selector. The point at which selection occurs has been the topic of much debate. *Early* selection (e.g. Broadbent, 1958) proposes that all stimuli reaching the sensory system are analysed for fundamental properties such as loudness and pitch. Later stages of processing are assumed to be able to process only one stimulus at a time. This approach hypothesises a filter which is used to select a particular stimulus which is to be processed on the basis of simple physical attributes. In contrast to this, theories of *late* selection (e.g. Deutsch and Deutsch, 1963) propose that all stimuli are recognised unselectively without capacity limitation, no matter how many stimuli are present. It is only later capacity-limited processing which necessitates a filtering of the processed stimuli. Since subjects are generally unaware of unattended stimuli, late selection theories propose that subsequent mechanisms are responsible for a particular stimulus to reach ‘awareness’ and be perceived (e.g. Duncan, 1980).

Early selection

Early selection is at the heart of Broadbent’s (1958) general theory of attention. It was proposed that the hypothetical filter can be tuned by the observer to any one of a number of input ‘channels’ (figure 1). Broadbent reasons that a filter is required because cognitive mechanisms (at higher levels than the filter) have finite capacity and hence a certain degree of selectivity is necessary. Only information in a single channel to which the filter has been tuned is passed to later processing stages. Broadbent suggests a wide range of possible information types that can be

1.For the purposes of this model, it is necessary to clarify the meaning of the term *preattentive*. The classical meaning of this term refers to processing which occurs before conscious intervention by the subject. In the framework of endogenous and exogenous attention, preattentive processing can be considered to be equivalent to exogenous attentive processing.

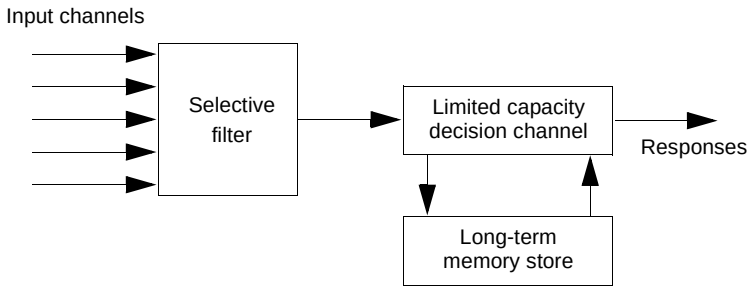


Figure 1 **Schematic of Broadbent's (1958) filter theory. After Treisman (1964b).**

conveyed within a channel such as responses from sense organs, directions in space, and voice qualities.

The findings of Cherry's (1953) perceptual studies could be explained by Broadbent's theory. Cherry presented listeners with two spoken sentences simultaneously, one to each ear. The subjects were instructed to shadow one of the sentences (immediately repeat the spoken material). It was found that despite the ease with which listeners could shadow one of the sentences, they could recall little of the ignored sentence other than the fact that sounds had been present. Indeed, subjects did not notice when the ignored speech changed language, or was played in reverse. It was only when the gender of the ignored speaker changed, or the speech was replaced by a pure tone, that listeners reliably reported an alteration. It is interesting to note that such a phenomenon is not restricted to the percept of speech. Deutsch (1986) reported that when listeners sang along with a melody presented to one ear, they were unable to give any information about the melody presented to the other ear.

Within Broadbent's framework, it can be argued that the speech from both ears is analysed for primitive features such as pitch and location, and that selection occurs at the 'selective filter' stage. Only the stimuli of the correct pitch and ear reach the limited capacity processing channel. Broadbent argues that a listener's ability to detect changes in the ignored sentence (such as pitch) is important as it aids the decision made by the selective filter, even though the actual words are ignored.

Related work by Moray (1959) suggests that Broadbent's model is too restrictive: the selective filter only allows the information within a single channel to pass through to later processing stages. In contrast to this, it has been shown that information from a 'rejected' input *can* influence later processing. Moray (1959) demonstrated that instructions presented to the unattended ear of a listener were

ignored in a similar fashion to the subjects in Cherry's study; however, when an instruction was prefaced by the listener's name, it was perceived and acted upon. Indeed, the enhanced ability to detect one's name is demonstrated by the ability to wake sleeping subjects by speaking their name (Oswald *et al.*, 1960). Note that within Broadbent's framework, information from a rejected (unattended) input channel cannot reach the later processing stage and influence the subject's response.

The sentence shadowing studies of Treisman (1960) have also identified contextual influences on the selection process. As described above, subjects were instructed to shadow the sentence in a particular ear and ignore the other. At a particular point in the presentation, the sentences were switched to the opposite ears. It was found that at the point of the switch, listeners continued to shadow the original sentence for a few words, despite it now being presented in the 'wrong' ear. Subsequently, subjects reverted to the correct ear and shadowed the 'new' sentence. This suggests that there is more than a simple channel selection process at work: contextual information is used for attentional stream[2] propagation. It is only when the listener becomes aware of the conflict between this *subconscious* decision to follow the contextually-correct sentence and the *conscious* desire to shadow a particular ear that they are able to rectify this and continue to shadow the new sentence in the correct ear.

Treisman (1960) proposed a relaxation of Broadbent's model in which the filter *attenuates* the incoming signals as opposed to blocking them completely. It is then assumed that different words require different signal intensities for recognition: a signal detection problem. An adjustable cut-off is adopted above which signals are accepted and below which they are rejected as being noise. This cut-off is adjusted on the basis of current contextual cues and word 'importance'. For example, this cut-off would be lowered for the detection of one's name. In the case of unattended streams, which have been attenuated by the filter, their information would only flow on to later processing stages if they contained words which passed the signal detection test. Therefore, Treisman proposes that the difference between an attended and an unattended stream is attenuation: the attended stream being the louder. However, at a cocktail party, the ignored (or unattended) voices do not seem less loud simply because they are not the attentional focus.

2.It is useful to define a term to describe the time course of the perceptual process; we define 'attentional stream' to refer to the time varying conscious percept. It is important to note that this is related to, but different from, the definition of a stream in ASA as defined in chapter 2. At any point in time, the attentional stream is equivalent to one attentionally selected ASA stream.

Further support for processing of unattended streams comes from later sentence-shadowing work by Treisman (1964a, 1964b). In these studies, an interesting alteration was made to the conventional shadowing task: the unattended sentence was a time-delayed version of the shadowed sentence. Despite the subjects not expecting the sentences to be the same, every subject eventually became aware of this. Indeed, when this experiment was repeated on bilingual subjects and the unattended ear was played a translation of the shadowed sentence, a number of subjects still reported the identity.

Late selection

An alternative explanation of these results is to assume that all groups are processed completely and attention then selects one to become the attentional stream: an example of *late* selection (e.g. Moray, 1959; Deutsch and Deutsch, 1963; Norman, 1968; MacKay, 1973; Duncan, 1980). However, this introduces a different conceptual problem: a significant amount of processing must be occurring in order to fully analyse every possible stream. Furthermore, if this processing must take place, why is so little of each unattended stream perceived?

Evidence in support of the same degree of semantic analysis of the rejected message as that of the attended message has often been disputed. For example, Corteen and Wood (1972) used autonomic responses to keywords (city names) which had been previously paired with mild but unpleasant electric shocks. When tested, subjects shadowed a sentence of prose whilst ignoring a contralaterally presented sentence which contained shock-associated city names, unassociated city names and unrelated nouns. A significantly higher proportion of conditioned galvanic skin responses (GSR) were detected when the unattended message contained an associated city than when the unattended message contained unrelated nouns. A similar, although smaller, effect was observed for the unassociated city names. When questioned on the content of the ignored message, subjects did not recall the presence of the shock-associated city names and no loss of shadowing performance was observed at the times of shock-associated city name presentation. This was taken as evidence for the semantic analysis of the unattended message. Dawson and Schell (1982) carefully analysed this effect and found that when lapses in selective monitoring (such as switching ears) were discounted, GSR responses to the rejected message were almost, but not completely, abolished. This suggests that Corteen and Wood (1972) may simply have observed subjects occasionally lapsing when shadowing a sentence. However, it should be noted that these findings do not necessarily contradict those of Moray (1959) in which unattended utterances were recognised when associated with keywords. Moray demonstrated this phenomenon with a highly salient keyword - the subject's name. In contrast, the keyword set

used by Corteen and Wood was arguably ‘contrived’: despite the learning period, each city name would still be of relatively low saliency to the subject when compared to their name. It is possible that the saliency of Corteen and Wood’s keyword set was not high enough to trigger recognition within the rejected (unattended) speech.

Similarly, the shadowing experiments of MacKay (1973) have been called into question, in which the meaning of an ambiguous shadowing sentence could be influenced by a word in the ignored message. Newstead and Dennis (1979) found that this effect was only observed when the disambiguating word was the only one present in the ignored message. When a full sentence was presented to the unattended ear, the effect was not observed. They concluded that the presence of a single word in the unattended ear disrupted the channel selection decision; hence, the word momentarily became the focus of attention, and was analysed.

5.3 Interim summary

None of the experimental findings presented above fully support either early or late theories of selection: a compromise between the two is necessary. A relaxation of Broadbent’s framework is required to allow a certain degree of keyword (such as one’s name) and semantic analysis (contextual influences) to occur. Indeed, the previous attentional frameworks view attention as a selector of incoming stimuli whose constituents have been analysed to a lesser or greater degree. Early selection theories assumes that stream formation occurs after selection; late selection theories assume that stream formation occurs before selection. Evidence presented in the previous chapter draws this assumption into question. Carlyon *et al.* (2001) demonstrated that attention is an integral part of the stream formation process.

In the next section a new conceptual framework for auditory selective attention will be presented in which the experimental findings of the previous chapter and the findings of early versus late selection studies will be incorporated. This framework will form the basis of the implementation presented in the next chapter.

5.4 A new conceptual framework for auditory selective attention

Before introducing the framework in detail, it will be useful to briefly review the key experimental findings which have been presented above.

- Auditory attention can be allocated, or oriented, to a specific frequency region; this allocation takes the form of a gradient with the peak centred on the frequency of choice (e.g. Mondor and Bregman, 1994).
- The allocation of attention to a particular frequency is a conscious activity; Spence and Driver (1994) have suggested that attention is comprised of two types: conscious and unconscious (endogenous and exogenous, respectively). When attention is directed toward a particular frequency, it is actually endogenous attention which is involved; exogenous attention is concerned with primitive grouping of features and the detection of gross changes in the input.
- Carlyon *et al.* (2001) have shown that endogenous attention is essential for stream formation and hence a stream's constituent features and/or groups of features can only be successfully combined at the endogenous attention stage.
- Furthermore, it is only a stream which has been formed from endogenous attentional processes that is fully perceived (e.g. Moore and Egeth, 1997).

The framework presented here assumes a number of exogenous processes can occur simultaneously and that endogenous attention is required to allow the outcome of these processes to be perceived and encoded into memory (figure 2).

Exogenous processes are responsible for performing primitive grouping (see chapter 2) of individual features within the stimulus. These groups are then passed to the endogenous processing stage, which takes into account the conscious decision of the listener, changes in the stimulus input and schema information to form an 'attentional stream': the time varying conscious percept.

Each of these endogenous factors compete to produce a single stream. For example, it is possible for a subject's attentional focus to be reoriented, contrary to their conscious preference, to an alternative grouping. Schema information can overrule a conscious decision if sufficient support for such a reorientation is provided. For example, a person's name can be considered to be an important schema: if the person's name is detected, the grouping that contained that information ought to become the attentional focus. All of the groups created by the exogenous processing stages are analysed to see if they match the schema. If one of them does, then attention is reoriented, against the subject's intentions, to the previously

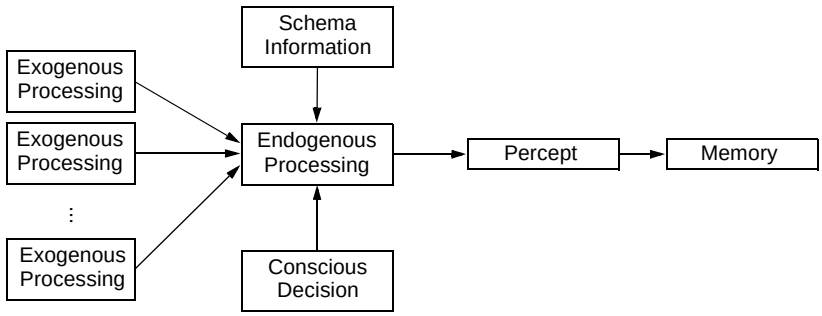


Figure 2 **Structure of the attentional model. Note that endogenous processing is required before exogenous perceptual organisations can be encoded into memory and perceived.**

unattended group. This results in the percept that the subject's attention has been drawn from the original conversation to one in which their name appeared (see Moray, 1959). Once this has occurred, it then becomes a conscious decision for the subject to decide whether to remain with the 'new' conversation or revert to their original stream. Experimental findings described previously suggest that such schema only encode small amounts of strongly salient information such as one's name, or one's partner's name: when subjects were instructed to learn arbitrary pieces of information, awareness of their presence in an unattended stream was significantly reduced (Dawson and Schell, 1982).

Another form of unconscious redirection of conscious attention can happen in virtually any environment: a loud, usually transient, sound occurring unexpectedly such as a bang or a crack which initiates a startle reflex (e.g. Winslow *et al.*, 2002). In this situation, exogenous information about the gross change in the stimulus triggers the detector bank and endogenous attention is directed to the new sound without regard for the listener's conscious preference. Indeed, physiological studies provide support for the hypothesis of exogenous monitoring of unattended stimuli (e.g. Sussman and Winkler, 2001; see also Winkler and Czigler, 1998).

Schema information can also be used to aid the grouping of the exogenous processing outputs and form the attentional stream. In particular, schemas can encapsulate semantic information about grammar and contextual meaning. For example, despite a subject's conscious intention to shadow the speech in a particular *ear*, Treisman (1960) found that when the two sentences switched ears, the subject shadowed the original *sentence*. This implies that at the stage of endogenous processing, schema information related to the sentence semantics is

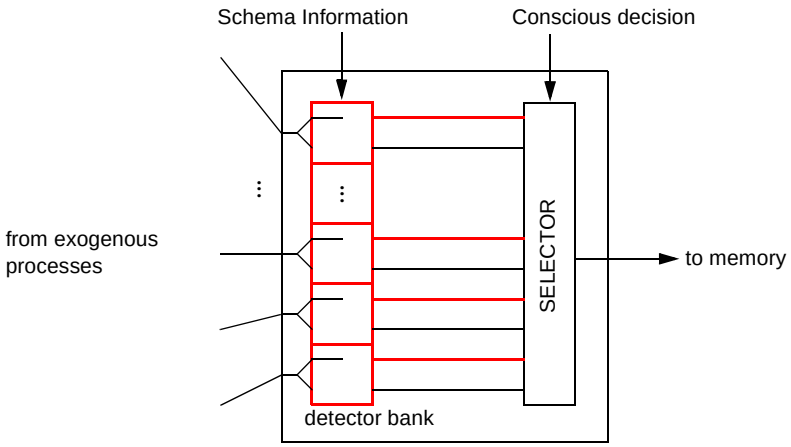


Figure 3 **Enlargement of the endogenous processing mechanism from figure 2. The *detector bank* indicated in the diagram corresponds to the proposed 'change' detectors and schema detectors of the model.**

being employed to overrule the listeners conscious decision to shadow a particular ear.

In summary, endogenous processing requires more than conscious input; it is proposed that other processes such as 'change' detectors, schema recognisers, etc. can overrule a conscious decision. This mechanism is shown in figure 3. One or more exogenous processes are responsible for grouping and streaming of stimuli reaching the ears. The perceptual organisations of these processes are directed into the endogenous selection module. It is at this stage that both conscious decisions and salient information about the incoming organisations are combined and a selection made. Salient information could be gathered from a number of sources such as difference detectors and schema recognisers (both contained within the *detector bank* abstraction of figure 3): changes in intensity or overall grouping structure may be considered in the difference detectors. Such an assumption finds support in the mismatch negativity studies conducted by Sussman and colleagues (Sussman *et al.* 1998, 1999; Sussman and Winkler, 2001) which suggest that a component of event-related potentials indicates the outcome of a comparison process when the incoming stimulus differs from the memory of the stimulus in the recent past. The fact that conscious attention is not required to illicit such a response is consistent with the assumption made in this model that such difference detectors function continually and may influence the selection process. In a similar fashion, schema driven processing can influence endogenous attention independently of the particular stream attended. As described above, a listener will

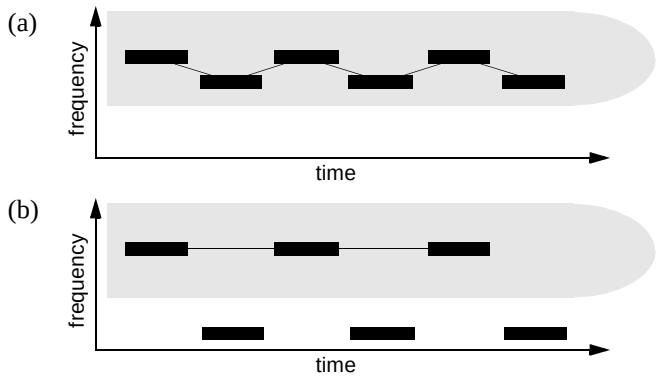


Figure 4 **Influence of attentional allocation on the formation of streams. In both situations, the allocation of attention is focused on the high frequency tones. (a) At low frequency separations, both sets of tones fall under the attentional focus (denoted by the gray area) and segregation is not observed. (b) At higher frequency separations, one set of tones does fall outside of the attentional focus and as such does not contribute to the attentional stream: segregation occurs. In both diagrams, the time-course of the attentional stream is indicated by black connecting lines.**

become aware of the information in an unattended stream only when that stream contains some highly salient information such as their name (Bregman, 1990; see also Moray, 1959). This implies that schema recognition is being carried out on all exogenous process inputs and influences the final selection process.

The last stage of the endogenous selection process is the combination of all the decision factors: schema and difference detection outcomes and conscious choice. In the absence of any evidence from the detector banks, the decision will be consistent with the conscious choice. However, should evidence appear that important information is present in a currently unattended stream, the selector overrules the conscious choice and directs attention to that stream.

Endogenous attention

In the framework proposed here, it is the manner in which endogenous attention is deployed which accounts for the finding that attention is required for stream formation. We have seen in the previous chapter that attention occurs in the ‘shape’ of a gradient and can be assigned to a particular frequency (e.g. Mondor and Bregman, 1994). We model this attentional focus ‘shape’ as a gaussian distribution. Furthermore, only feature groups which are subject to attention are perceived by the subject. Therefore, it can be theorised that group segregation on the basis of

frequency proximity occurs due to the nature of the attentional focus: features and/ or groups which fall beyond the skirts of the attentional allocation are said to be outside the influence of attention and hence do not contribute to the attentional stream. Consider the ABA-ABA two tone streaming phenomena (van Noorden, 1975) which has been described in previous chapters; the stimulus consists of a sequence of tones whose frequencies alternate over a period of time. Provided both the low and high frequency tones are sufficiently close in frequency as to fall under the attentional focus, they will both be temporally grouped (figure 4a). However, if the frequency separation is large, only one set of tones will fall under the focus of attention and be perceived as belonging to the attentional stream: the other set are outside the influence of attention and remain unattended (figure 4b). The framework can also account for the streaming of complex tones (van Noorden, 1975) in which the same form of sequence is used as above with the exception that the pure tones are replaced by complex tones. In this situation, each harmonic complex will be grouped by exogenous processes. When attention is directed to a particular frequency region containing a harmonic, that harmonic and all other grouped components contribute to the attentional stream.

A further phenomenon is observed with alternating tone sequences (van Noorden, 1975) when the frequency separation and/or repetition rate fall in what is termed the ‘ambiguous region’. In this situation, the listener can consciously decide whether to perceive temporal coherence (galloping ABA percept) or segregation (A-A-A or B-B-B). Our framework suggests that this occurs when one set of tones are presented in the tails of the gaussian attentional focus. Evidence from auditory research (see Moore, 1997) suggests that the size of the attention focus can be adapted on a task by task basis and also with training. Hence, if the size of the attentional focus is consciously adaptable, the listener would be able to either contract the width of the gaussian and promote segregation, or expand the gaussian to promote temporal coherence.

Anstis and Saida (1985) demonstrated that the segregation percept did not occur instantaneously: despite a sequence of tones being widely separated in frequency, it still took some time (up to ten seconds) for the streaming percept to occur. This agrees with Bregman’s contention that the default auditory organisation is a whole (Bregman, 1990; see also Bregman, 1978): initially, all features are contained within the attentional stream. The proposed framework accounts for this behaviour by associating a time-course to the attentional allocation. At the beginning of a stimulus, all frequencies are attended; however, over time, allocation adapts so that only frequencies ‘covered’ by the gaussian attentional focus can contribute to the attentional stream (figure 5). Further support for an attentional time course comes from evidence in which the percept of auditory streaming *decays* over a finite

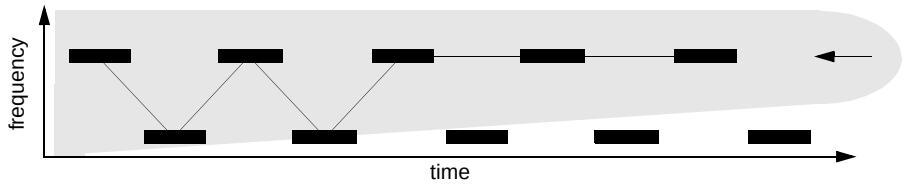


Figure 5 **Time-course of attentional allocation. Initially, all frequency regions receive equal attentional influence; over time, the allocation adapts so that the gaussian form of allocation only covers the attended frequency region. The arrow on the right hand side of the figure indicates the desired (conscious) focus of the attentional allocation. Similar to figure 4, the time-course of the attentional stream is indicated by black connecting lines.**

period of time (Cusack and Carlyon, 2001). Cusack and Carlyon presented listeners with a stimulus consisting of a repeating cycle of an ABA alternating tone sequence lasting for ten seconds followed by a gap. Four conditions were used in which the gaps were of 1, 2, 5 and 10 seconds. Throughout each presentation, listeners were instructed to signal when their percept changed from one stream to two streams and vice versa. It was found that the size of the pause between bursts of the alternating tone sequence was proportional to the likelihood of subjects hearing a single stream at the beginning of the next alternating tone burst despite having perceived two streams at the end of the preceding burst.

Jones (1976; see also Large and Jones, 1999) proposed an alternative mechanism by which such stimuli could be grouped using 'pattern' information. This theory envisages an attentional process which groups auditory events with a sequence of stimuli which are already attended to. In other words, if the 'new' events conform with the multidimensional pattern of the attended sequence, it is integrated with that sequence. This is achieved by predicting the multidimensional position of the next sound in the sequence from the preceding ones (e.g. in time, frequency, etc.). The prediction then informs the attentional process of where to shift in order to capture the next consistent sound. Such predictions are made on the basis of 'global' rules which describe the entire sequence pattern of the sounds perceived thus far. The theory also argues that as grouping rules become simpler, their corresponding sequences become easier to attend to (Jones *et al.*, 1978). However, Bregman (1990) argues that such a theory is flawed since it is too rule-based and hence not sufficiently general. For example, consider a form of the alternating tone sequence example in which a random process determines the ordering of the A and B tones. In this situation, a listener will integrate or segregate the sequence purely on the basis of the frequency separation of the tones. However, since the sequence is random, Jones' theory would fail to produce an appropriate rule by which to

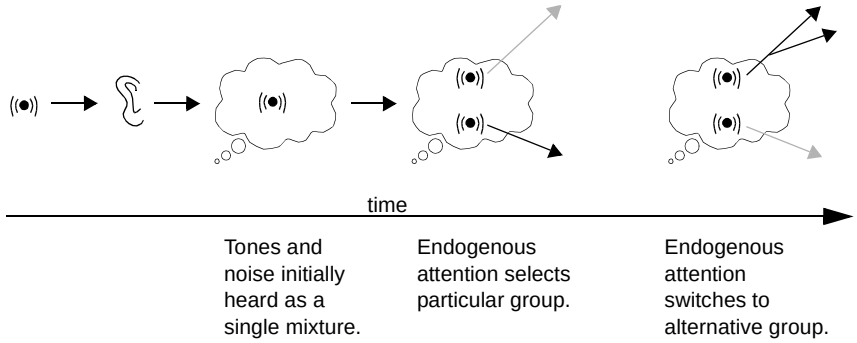


Figure 6 **Behaviour of the model in response to the Carlyon *et al.* (2001) stimulus of tone sequences and noise bursts. Gray arrows indicate exogenous processing which is not encoded into memory due to endogenous attention being directed toward other exogenous processing (black arrows).**

describe the sequence. In this case, it is unclear what her theory would predict. Despite the random process, our framework would emulate the behaviour of a listener since the integration / segregation process is based only on the frequency separation of the tones and not any temporal pattern. Jones *et al.* (1981) presented further evidence which suggested rhythm favoured segregation. They used a sequence similar to that of Bregman and Rudnicky (1975; see also previous chapter) in which a sequence of tones C were used to capture flanking tones F from a four tone pattern FABF. Jones *et al.* pointed out that the captor tone were not only related in frequency to the flanking tones but also rhythmically related. They argued that the AB tones were being segregated from the C and F tones on the basis of rhythm; a conclusion which was supported by further trials in which the rhythm of the sequence was adapted and altered the ease with which and AB tones could be segregated from the rest of the sequence. Indeed, they concluded that '*temporal predictability may be a prerequisite to the establishment of stream segregation based on frequency relationships*' (Jones *et al.*, 1981, p.1071). Bregman (1990) argues that such a conclusion may be too strong; he notes that the Bregman and Rudnicky (1975) sequence is influenced by both primitive grouping principles and exogenous attention. Bregman argues that rhythmic information 'assists the selection process *directly* rather than indirectly through its effects on primitive grouping' (Bregman, 1990, p.445, italics his). In other words, rhythm acts as a schema at the endogenous processing level of our framework rather than at the exogenous processing level. This concept can be considered to be at the core of Baird's (1996) stream segregation architecture. Baird models Jones' theory of rhythmic attention using subsets of oscillatory associative memories which form a

scale of potential rhythms and establish rhythmic expectancies. Patterns which meet these expectancies are subject to an amplitude increase in their response and become attended to. Patterns which fail to meet the expectancies cause a simulated mismatch negativity response (see previous chapter) which promotes that input to form a new stream. These oscillatory associative memories can be considered to be schemas within our framework which can aid the propagation of (rhythmic) streams and also signal high level changes (in this case, rhythmic changes) in the input to influence the endogenous decision making process.

We have looked at how the model can explain the streaming of an alternating tone sequence and also the associated time course for this percept to appear. We will now consider the more complex scenario used by Carlyon *et al.* (2001) in which noise bursts are presented simultaneously with a galloping ABA-ABA tone sequence. The initial state of perception is that of fusion: a single stream is perceived to which endogenous attention is directed by default. The listener rapidly (almost instantaneously) becomes aware of the two types of sounds: tones and noise bursts, the features of which are organised into two distinct groups by exogenous processes. At this stage, the subject has to decide as to which perceptual object to attend to (figure 6). In this example, the subject is instructed to attend to the noise bursts and perform a perceptual task on them: to classify them as approaching (linear increase in amplitude) or departing (the approaching burst reversed in time). Our framework suggests that exogenous processing occurs at all times for all streams. However, only groups subject to endogenous attention can form a stream and then be encoded into memory and perceived. Therefore, in this state of diverted attention the build up of streaming does not occur as the tones are not subject to endogenous attention. Later in the experiment, the listener is instructed to switch tasks and concentrate on the alternating tones. In this situation endogenous attention is directed toward the tones and the streaming process occurs with the associated time-course.

5.5 Summary

This chapter has presented a conceptual framework in which the allocation of attention lies at the heart of the stream formation process. It is proposed that exogenous processes perform primitive grouping of stimulus features to produce a number of groups. These are then passed to the endogenous processing mechanism which takes into account schema information extracted from the groups, exogenous factors such as gross changes in the stimulus characteristics and the conscious decisions of the subject to create an attentional stream. We suggest that this

framework can account for the two tone streaming phenomenon (van Noorden, 1975) as well as the experimental findings of Carlyon *et al.* (2001), whose results indicate that attention is a requirement of stream formation.

The next chapter will present a neural oscillator based implementation of this framework which will allow us to demonstrate the ability of the model to account for psychophysical phenomena. Indeed, this level of explanation (see Marr, 1982) will allow us to make predictions, and identify areas for further investigation, with greater detail. Furthermore, it will allow us to investigate how an attentional mechanism can be parsimoniously incorporated into the temporal correlation framework.

Chapter 6. A Computational Model of Auditory Selective Attention

6.1 Introduction

One difficulty involved in producing a computational solution to the ASA problem is the lack of a strong link between Gestalt theories of perception (see Bregman, 1990) and the underlying physiological process. Consider the two tone streaming stimulus described previously (see figure 1). Theories of perception are implied from experimental observations. Applying such theories to figure 1, one can conclude that as the frequency separation - δf - decreases, it is more likely that the

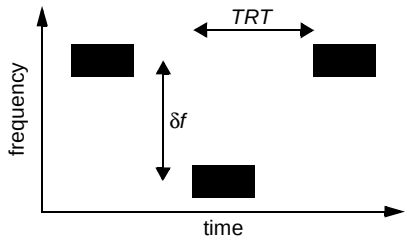


Figure 1 **Portion of an alternating tone sequence.**

tones will be grouped together. Similarly, as repetition time of the tones - *TRT* - decreases, sequential tones will also be more likely to group provided δf is not too large.

However, the neurophysiological mechanisms underlying auditory stream formation are poorly understood and it is not fully known how groups of features are coded and communicated within the auditory system. What does it mean to talk of 'frequency proximity' or 'temporal proximity'? The human brain relies solely on time varying electrical impulses with no 'symbolic', time-frequency representation as suggested by Bregman's theory. Despite this, some computational models have taken a symbolic time-frequency approach to describing an auditory scene (e.g. Cooke, 1991/1993; Brown and Cooke, 1994). For example, once a stimulus has been processed by a model of the auditory periphery, to simulate cochlear filtering and inner hair cell transduction, Brown and Cooke (1994) extract features which can be used for auditory grouping. Specifically, they use periodicities, onsets, offsets and frequency transitions to produce 'symbols' which represent individual harmonics or formants through time and frequency. One can regard this representation as a 'cartoon' of the stimulus. The final stage involves a search of all the symbols to determine which belong together based upon Gestalt grouping principles. For instance, symbols which start and stop at the same time (same onset and offsets) and belong to the same fundamental frequency are likely to have arisen from the same source and are hence grouped.

However, it is unlikely that the brain reasons about groups of auditory symbols. The temporal correlation approach draws on the suggestion that the nervous system may signal the grouping of auditory channels by the phase of firing of feature detecting cells (see chapter 3 for a discussion of temporal correlation based binding solutions). For example, the neural oscillator based solutions of Wang and collaborators (Wang and Terman, 1995; Terman and Wang, 1995; Wang, 1996; see also Brown and Wang, 1999; Wang and Brown, 1999) represent auditory activity

within a time-frequency grid. Each point in the grid is associated with a neuron that has an oscillating response pattern and the time dimension is created by a system of delay lines.

However, the use of a time-frequency grid on which to perform grouping also requires unrealistic assumptions. Firstly, no physiological evidence to support the existence of such a structure has been found. Secondly, the implementation of Wang's two dimensional grid suffers from low time resolution: the input to the model is sampled at 40 ms intervals. In Wang's architecture, input to each time slice in the time-frequency grid is a multiple of this 40 ms. However, in order to increase the time resolution of the grid, a significantly higher number of delay units would be required. The existence of such large arrays of a delay lines is yet to be determined and the ability of such delay lines to maintain temporal accuracy over large distances is uncertain.

Of more concern is the manner in which the simulated behaviour of attention is influenced by such a topology. In chapter 4, we discussed how attention has been incorporated into existing neural oscillator models. In particular, Wang's (1996) shifting synchronisation theory states that 'attention is paid to a stream when its constituent oscillators reach their active phases'. This implies that the process of sound segmentation and grouping and the process of attentional selection are linked, if not identical. Once a stream has been formed, it will be attended to when its associated set of oscillators reach the active phase. Since each synchronised block of oscillators within the time-frequency grid become active in a repeating sequence, attention quickly alternates between each different stream at different positions on the time-frequency grid. Hence, stream multiplexing occurs and all streams are perceived as equally salient at all times. This contradicts experimental findings (see Bregman, 1990) which show that listeners perceive one stream as dominant. This theory cannot explain how attention may be redirected by a sudden stimulus. Such an event would be encoded by Wang's network as an individual stream which would be multiplexed as normal - with no attentional emphasis. Furthermore, Terman and Wang (1995) demonstrated that the overall time that the system takes to represent individual patterns is no greater than m cycles of oscillations, where m is the number of patterns simultaneously presented to the network. Such rapid stream formation is problematic since Anstis and Saida (1985) have demonstrated that segregation in two tone streaming can take many seconds to occur.

These attentional problems relate directly to the use of a two dimensional time-frequency grid and as such it can be argued that time has not been properly considered in such models. The time-frequency grid is derived from the two

dimensional spatial grid used in Wang's *visual* scene analysis research (e.g. Wang, 1999; 2000). In order to apply the same architecture to *auditory* scene analysis, the acoustic input is represented as a two dimensional time-frequency pattern. Indeed, it is acknowledged that grouping of streams *over time* is not addressed in such auditory scene analysis models: '*our model does not address sequential grouping; in other words, there is no mechanism to group segments that do not overlap in time*' (Wang and Brown, 1999, p. 691).

The traditional view of attentional mechanisms within computational models of ASA is that stream formation and segregation are performed within a 'sealed unit': attention does not influence this process, it merely selects a particular stream that is generated (see chapter 4). The core of Wang's stream segregation system can be interpreted in this way: the oscillator array is a CASA preprocessor for subsequent attentional selection mechanisms. However, as such, it cannot explain attentionally modulated stream formation such as in the two tone streaming task described by Carlyon *et al.* (2001).

In summary, the success of the two dimensional network does not outweigh its conceptual disadvantages. It is useful to take a step back for a moment and consider what we are trying to achieve. Sound is a time-varying pattern of air pressure changes which is converted into a pattern of time-varying electrical signals by the ear. This is the only input available: a continual flow of electrical signals. From this, animals and humans can perform grouping and stream segregation which allows them to make sense of the most complex of sound environments. It is this idea that the following model takes as its starting point: a *one* dimensional network which receives time-varying patterns and produces a time-varying estimate of stream content at each temporal instant.

6.2 Monaural computational model

The model comprises three main stages, as shown in figure 2. The input to the model is a sound signal sampled at 8 kHz. The first stage of the model simulates peripheral auditory processing. The signal is passed through a bank of cochlear filters representing basilar membrane processing. The gain of each filter is set to reflect the outer-middle ear transfer function. An approximation to the inner hair cell transduction process is then computed from the output of each filter. This provides an estimate of auditory nerve firing response used in later stages of the model.

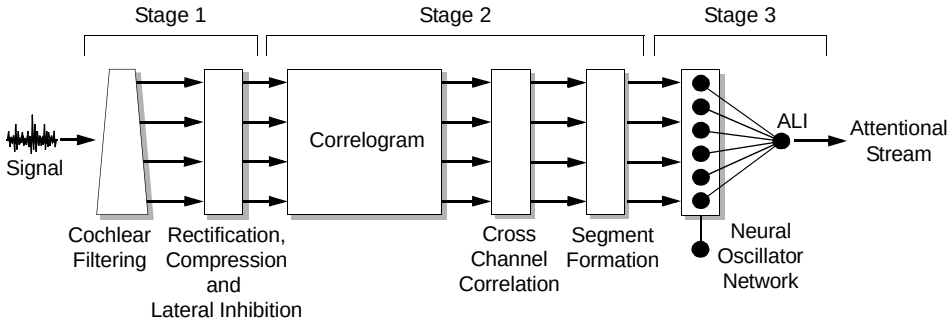


Figure 2 **Schematic diagram of the monaural model (the attentional leaky integrator is labelled ALI).**

The second stage of the model extracts periodicity information by means of a correlogram (Slaney and Lyon, 1990; Meddis and Hewitt, 1991a, 1991b; Brown and Cooke, 1994) allowing primitive grouping to be performed on the basis of harmonicity. Each channel in the correlogram is obtained by computing the autocorrelation of that channel's auditory nerve approximation. The resulting two-dimensional representation has channel centre frequency and autocorrelation lag on orthogonal axes. By summing all the frequency channels, a 'summary correlogram' is formed, the largest peak of which corresponds to the dominant fundamental frequency (F0) of the sound.

The third stage of our model is the one-dimensional neural oscillator network in which auditory grouping and segregation takes place. The network is based upon the locally excitatory globally inhibitory oscillator network (LEGION, see Terman and Wang, 1995) with two exceptions. Most importantly, the oscillator network is one-dimensional such that each frequency channel is represented by a single oscillator. The input to the network does not have any extent in time (as in Wang and colleagues' networks): the network processes the instantaneous values of the peripheral model output. Secondly, long range excitatory connections are permitted to allow harmonicity grouping. A cross channel correlation mechanism identifies contiguous regions of acoustic activity in the correlogram corresponding to Bregman's (1990) concept of *elements*: atomic parts of the acoustic scene. These are encoded in the network by locally excitatory connections and are observed as synchronised blocks of oscillators. Information from the summary correlogram is then used to group these segments on the basis of their conformity with the fundamental frequency estimate. Long range excitatory connections promote these oscillator blocks to synchronise to form an oscillatory 'group'. Blocks of oscillators which are not harmonically related desynchronise from each other.

Each oscillator in the network feeds activity to the attentional leaky integrator (ALI): the core of our attentionally motivated stream segregation mechanism. The output of the ALI is the *attentional stream* as defined in the previous chapter. The connection weights between the network and the ALI are modulated by endogenous processes including 'conscious' preference. Initially, these weights are maximal for all channels to simulate the default grouping being a whole. For example, a burst of white noise contains random cues for fusion and segregation, but is perceived as a single coherent source (Bregman, 1990). In this initial condition, all segments and groups contribute to the attentional stream. Over a period of time, these weights adapt to the gaussian shape of the endogenous attentional focus in which the peak of the gaussian is centred on the frequency channel of interest. In this situation, only oscillators under the attentional focus can influence the ALI. In terms of the computational model, the attentional stream is defined as containing all frequencies whose oscillators are synchronously active with the ALI. The reliance on synchrony to assess selection allows harmonic groups, most of whose harmonics may be outside of the attentional focus, to contribute to the attentional stream simply by attending to one harmonic.

Auditory peripheral processing

A linear system can be characterised by its response to a brief click - the *impulse response*. de Boer and de Jongh (1978) estimated the impulse response of auditory nerve fibres using a *reverse correlation* technique. The gammatone function (see Patterson *et al.*, 1988; see also Lyon, 1996, for an alternative implementation using an all-pole approach) is an analytical approximation to their (de Boer and de Jongh, 1978) physiological measurements. Here, the frequency selectivity of the basilar membrane is modelled by a bank of 128 gammatone filters with overlapping passbands (see Moore, 1997); the output of each filter represents the frequency response of the membrane at a specific position.

The gammatone filter of order n and centre frequency f Hz is given by:

$$gt(t) = t^{n-1}e^{-2\pi bt}\cos(2\pi ft + \phi)H(t) \quad (1)$$

where ϕ represents the phase, b is related to the bandwidth, n is the order of the filter and $H(t)$ is the unit step (Heaviside) function:

$$H(t) = \begin{cases} 1 & t \geq 0 \\ 0 & t < 0 \end{cases} \quad (2)$$

The name *gammatone* comes from the fact that the envelope of the filter impulse response is the statistical *gamma* function and the fine structure of the impulse response is a *tone* of frequency f and phase ϕ.

Because the gammatone filter is linear, it cannot simulate well known non-linearities such as *cochlear echoes*. For instance, Kemp (1978) suggested that cochlear mechanics are influenced by active biological processes. When a low-level auditory click was applied to the ear, sound was detected by a small microphone sealed into the ear canal following delays of between 5 to 60 ms (too long to be attributable to middle ear reflection). The reflected sound did not contain the same frequency content as the stimulating click, suggesting only particular points on the BM were involved. Furthermore, the response intensity varied nonlinearly with click intensity, in which low click levels resulted in the largest relative response levels. Another example of cochlear nonlinearity is known as a combination tone. When the ear is presented with two tones of frequency f_1 and f_2 ($f_1 < f_2$), an echo can be detected with frequency $2f_1 - f_2$ (Kim *et al.*, 1980) and subjects often report a tone of pitch $2f_1 - f_2$ (Plomp, 1965; Goldstein, 1967). Such a phenomenon suggests a nonlinearity in the auditory system which is likely to occur at the basilar membrane; a suggestion supported by observed basilar membrane responses (Robles *et al.*, 1991).

Despite its linearity, when the order of the gammatone filter (n) is in the range 3-5, its magnitude characteristic exhibits a very good fit to the *roex(p)* function commonly used to represent the magnitude characteristic of the human auditory filter shapes (see Patterson and Moore, 1986). This property justifies its use to model auditory frequency selectivity.

In our model, the gammatone filters are distributed in frequency between 50 Hz and 3.5 kHz according to their bandwidths, which increase quasilogarithmically with increasing centre frequency (see Palmer, 1987; Glasberg and Moore, 1990). The bandwidth b of a filter with centre frequency f Hz is set to the equivalent rectangular bandwidth (ERB), a psychophysical measurement of critical bandwidth in human subjects (Glasberg and Moore, 1990):

$$ERB(f) = 24.7\left(\frac{4.37f}{1000} + 1\right) \tag{3}$$

When $f >> b$, the bandwidth of the filter is proportional to b, a proportionality which depends only on the filter order n. For a gammatone filter of order 4, the bandwidth b is defined as (Patterson *et al.*, 1988):

$$b = 1.019ERB(f) \tag{4}$$

The gain of each filter is adjusted according to the ISO standard for equal loudness contours (ISO, 226) in order to simulate the pressure gains of the outer and middle ears. Since later stages of the model only require channel energy and periodicity information, the auditory nerve response is approximated by half-wave rectifying and square root compressing the output of each filter. The half-wave rectification process corresponds to the *phase-locking* action of the inner hair cells, which respond to movement of the basilar membrane in one direction for centre frequencies of up to approximately 5 kHz (see Pickles, 1988). A more detailed model of hair cell transduction (e.g. Meddis 1986, 1988; see also Hewitt and Meddis, 1991) is not used since, for our purposes, the increased accuracy of such a representation is outweighed by the computational cost. The approximation adopted here does not simulate the phenomenon of *adaptation* in which a sharp onset response in the auditory nerve is observed, which rapidly decays over the first 10-20 ms and then more slowly to a steady state over a period of 20-100 ms. It has been suggested that this form of adaptation is responsible for the reduced contribution of a mistuned component to the pitch of a complex when the mistuned component onset is slightly before that of the complex (e.g. Meddis and Hewitt, 1992). However, Ciocca and Darwin (1993; see also Darwin and Sutherland, 1984; Darwin, 1984) present evidence that suggests that adaptation only has limited influence on this phenomenon and that perceptual grouping principles play the dominant role. We argue that such findings justify the simplified approach adopted here.

To sharpen the frequency responses and reduce cross channel activity spread, the AN firing rate approximation at each time instant is convolved with a difference of gaussians ('mexican hat' shaped) kernel *d* (of size *s*):

$$d = e^{-\frac{c^2}{2\sigma^2}} - we^{-\frac{c^2}{3\sigma^2}} \quad c = -s..s \tag{5}$$

Here, σ determines the width of each gaussian in frequency (a value of *s*/4.444) and *w* is a weighting constant (a value of 0.8). The size *s* has the value 5. The resulting response is then half-wave rectified to avoid negative AN activities.

In addition to extracting the basilar membrane response from the gammatone filter, the instantaneous frequency and instantaneous envelope are used in later processing stages. The quantities are usually derived from an *analytic signal* (Gabor, 1946). The analytic signal *a(t)* of a real signal *s(t)* is defined as:

$$a(t) = s(t) - j\Pi[s(t)] \tag{6}$$

where $\Pi[\circ]$ is the Hilbert transform and j is $\sqrt{-1}$.

Cooke (1991/1993) describes a complex version of the gammatone in which the cosine term is replaced with a complex sinusoid:

$$g(t) = t^{n-1}e^{-bt}e^{j\omega t} \tag{7}$$

where ω is the radian centre frequency of the channel and b is the bandwidth. In turn, this can be rewritten as

$$g(t) = q(t)\cos\omega t + jq(t)\sin\omega t \tag{8}$$

where $q(t) = t^{n-1}e^{-bt}$. It can be shown that, under certain conditions, *g(t)* in (8) is approximately an analytic signal since the imaginary part of the gammatone filter closely approximates the Hilbert transform of the filter's real part (see Cooke 1991/ 1993 Appendix C for a derivation). Thus, it is possible to measure instantaneous frequency directly from the filter outputs and avoid having to compute a Hilbert transform. The instantaneous phase $\phi(t)$ of an analytic signal is defined as:

$$\phi(t) = \operatorname{atan}\left[-\frac{\Pi[s(t)]}{s(t)}\right] \tag{9}$$

Therefore, in terms of the gammatone filter, instantaneous phase is given by:

$$\phi(t) = \operatorname{atan}\frac{-\Im(t)}{\Re(t)} \tag{10}$$

where $\Re(t)$ and $\Im(t)$ represent the outputs of the real and imaginary parts of the gammatone filter respectively. Instantaneous frequency *v(t)* is the time derivative of instantaneous phase. Since the gammatone filter is implemented by frequency shifting by the centre frequency, lowpass filtering and then frequency shifting back (see Holdsworth *et al.*, 1988), the instantaneous frequency is found by adding the radian centre frequency ω to the derivative of (10) giving:

$$v(t) = \frac{1}{2\pi}\left(\omega + \frac{\Im(t)\dot{\Re}(t) - \Re(t)\dot{\Im}(t)}{\Im^2(t) + \Re^2(t)}\right) \tag{11}$$

The instantaneous envelope $\alpha(t)$ of the filter is given by:

$$\alpha(t) = \sqrt{\Re^2(t) + \Im^2(t)} \tag{12}$$

Pitch and harmonicity analysis

Theories of pitch perception can be considered to lie on a continuum with *pattern recognition models* and *temporal models* at the extremities. Consider the example in which a pitch of a complex tone can be perceived even when there is no energy present at the fundamental frequency due to, for example, filtering out the lower harmonics. Schouten (1940; see also Schouten, 1970) called this phenomenon the 'residue' pitch (also 'virtual' pitch). He argued that such a residue pitch did not require activity on the basilar membrane at the point which would respond maximally to a pure tone of similar pitch. The classical 'place theory' of pitch perception in which the pitch corresponds to the position of maximum excitation fails to account for this. In response to this behaviour, pattern recognition models (e.g. Goldstein, 1973) attempt to find the best fitting harmonic series to a set of resolved harmonics. For example, when analysing a complex which is missing at least its fundamental, the harmonic series which best matches the resolved harmonics present is sought. The fundamental of this series is hence said to be equivalent to the fundamental of the complex. Indeed, evidence exists that suggests that low frequency, resolved, harmonics contribute strongly to the pitch percept (see Moore, 1997). However, this type of model cannot account for residue pitches originating from complexes containing only high frequency, unresolved, harmonics (e.g. Moore and Rosen, 1979). By contrast, the temporal models of pitch perception (e.g. Schouten *et al.*, 1962) use the temporal fine structure of the auditory nerve firings to determine the pitch. They rely on the fact that the pattern of basilar membrane response due to *unresolved* harmonics overlap to a certain extent. Therefore, the basilar membrane waveform is a result of the interference (*beating*) of a number of high frequency harmonics and hence displays a periodicity equal to that of the signal's (residue) pitch. However, it ought to be noted that Schouten's model implies that a residue pitch can only be perceived if two or more (unresolved) harmonics are interacting.

Since neither theory can account for all the experimental data, there are also models of pitch perception which incorporate pattern recognition and temporal mechanisms such that temporal information from both resolved and unresolved harmonics is used to form a pitch estimate (e.g. Licklider, 1951; Slaney and Lyon, 1993; Meddis and Hewitt, 1991a, 1991b; Meddis and O'Mard, 1997). For example, Licklider's (1951) 'duplex' theory describes spectral analysis in the frequency domain occurring simultaneously with periodicity analysis in the time domain. He suggested that auditory nerve activity at each characteristic frequency is subjected

to a form of autocorrelation analysis (Licklider, 1959). Autocorrelation is the multiplication of a signal with a time-delayed version of itself. For periodic signals, a peak is observed at the time-delay relating to the fundamental frequency of the signal.

It is also known that a difference in fundamental frequency (F0) can assist the perceptual segregation of complex sounds (e.g. Broadbent and Ladefoged, 1957; Brokx and Nooteboom, 1982; Scheffers, 1983; Assman and Summerfield, 1987; McKeown and Patterson, 1995; see also Moore, 1997). For example, Scheffers (1983) found that two simultaneous vowels can be identified better when their F0s differ than when they are the same. One explanation for this result is 'harmonic sieving' (c.f. Goldstein, 1973) in which a harmonic series is detected by using a sieve which has 'holes' at integer multiples of the fundamental frequency (for an alternative based on harmonic cancellation, see de Cheveigné, 1993). Harmonics which are related to the fundamental frequency align with the holes and 'fall through' and hence contribute to the pitch percept. All other harmonics are blocked. Experiments using complexes with mistuned harmonics have been used to investigate how precisely such harmonics have to be aligned with the sieve holes. For example, data collected by Darwin *et al.* (1995) suggests that mistuning up to approximately 4%, the mistuned harmonic made a normal contribution to the percept. However, for mistuning of 4-8% this contribution declined until mistunings of more that 8% had no effect on the pitch percept of the complex. This implies that the auditory system has a certain degree of tolerance when grouping components with a common fundamental frequency.

In order to take advantage of such periodicity information in the grouping process, the second stage of the model extracts pitch information from the simulated auditory nerve responses. Similar to Licklider's model, this is achieved by computing the autocorrelation of the activity in each channel to form a correlogram (e.g. Meddis and Hewitt, 1991a; Slaney and Lyon, 1993). At time t, the autocorrelation of channel i with lag τ is given by:

$$A(i, t, \tau) = \sum_{k=0}^{P-1} r(i, t-k) r(i, t-k-\tau) w(k) \qquad (13)$$

Here, $r(i,t)$ is the auditory nerve activity in channel i at time t. The autocorrelation for channel i is computed using a 25 ms (P = 200) rectangular window w (see Assmann, 1996, p. 1151) with lag steps equal to the sampling period (0.125 ms), up to a maximum lag of 20 ms.

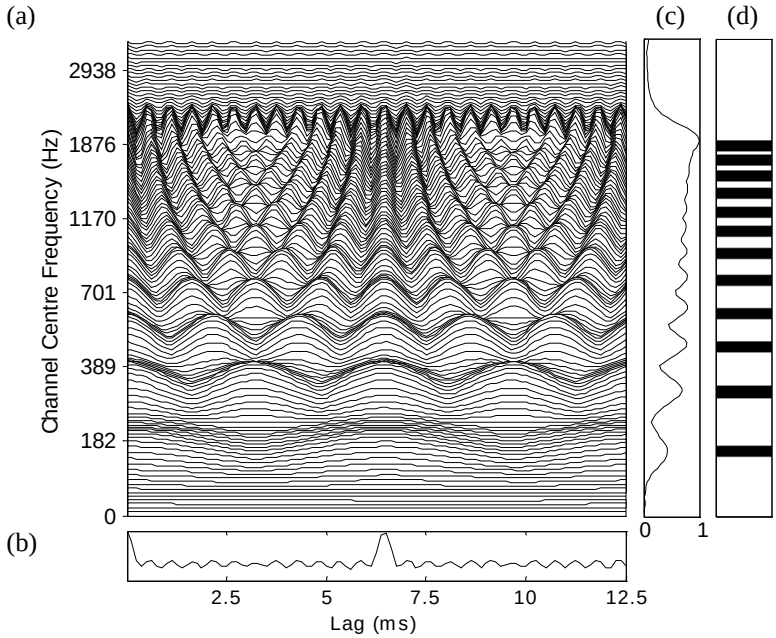


Figure 3 **A correlogram of a 12-harmonic complex tone with a fundamental frequency of 155 Hz. The main panel (a) shows the correlogram generated from the sharpened auditory nerve response; the lower panel (b) shows the summary correlogram. For clarity, only time-delay lags up to 12.5 ms have been shown. The cross correlation is shown on the right (c) with the segment 'mask' (d) in which black areas denote segments. The vertical misalignment of the correlogram compared with the cross correlation and segment diagrams is due to an offset used in generating the correlogram diagram to improve clarity.**

Figure 3 shows a correlogram for a 12-harmonic complex tone with a fundamental frequency of 155 Hz. Channel centre frequency is represented on the ordinate and autocorrelation lag is represented on the abscissa. The characteristic 'spine' (vertical alignment of peaks) can be seen in the correlogram at the lag corresponding to the fundamental period (approximately 6.5 ms). When the correlogram is summed across frequency, the resulting 'summary correlogram' exhibits a large peak at the lag corresponding to the fundamental period of the stimulus (see figure 3). The summary correlogram is calculated by

$$s(t, \tau) = \sum_{i=0}^{N-1} A(i, t, \tau) \qquad (14)$$

where N is the number of channels in the correlogram.

The peak of the summary, whose lag represents the fundamental period, is identified by normalising the summary by the energy (zero delay) and then centre clipping (Sondhi, 1968; see also Dubnowski *et al.*, 1975) by a factor of 80% to leave only the dominant peaks. Since the correlogram summary only contains positive excursions from zero, the centre clipping function is described by

$$s(t,\tau)_{clip} = \begin{cases} s(t,\tau) - \theta_{clip} & s(t,\tau) > \theta_{clip} \\ 0 & otherwise \end{cases} \tag{15}$$

where θ_{clip} is the absolute clipping value. The first local maximum is the pitch period estimate. To improve the resolution of the summary peak, a parabolic curve is fitted to the peak. The actual point at which the summary reaches a maximum may occur between two discrete values (i.e. between two lag values) and hence an accurate estimate of the maximum value and position cannot be obtained directly from the discrete summary. The parabolic curve fitted to the three samples centred on the peak has the equation

$$Y = m_0 X^2 + n_0 X + p_0 \tag{16}$$

where X represents the lag axis; m_0, n_0 and p_0 are constants. Therefore, for the three samples of concern, the equations are

$$Y_1 = m_0 X_1^{\,2} + n_0 X_1 + p_0,\ Y_2 = m_0 X_2^{\,2} + n_0 X_2 + p_0 \text{ and } Y_3 = m_0 X_3^{\,2} + n_0 X_3 + p_0$$

Algebraic manipulation of these three equations allow the values of m_0, n_0 and p_0 to be determined

$$m_0 = \frac{\dfrac{Y_1 - Y_2}{X_1 - X_2} - \dfrac{Y_1 - Y_3}{X_1 - X_3}}{X_2 - X_3} \tag{17}$$

$$n_0 = \frac{Y_1 - Y_2}{X_1 - X_2} - m_0 (X_1 + X_2) \tag{18}$$

$$p_0 = 0 \tag{19}$$

To find the position at which the true peak occurs, these values for m_0, n_0 and p_0 are substituted into the equation resulting from differentiating Y with respect to X

$$\dot{Y} = 2m_0X + n_0 \tag{20}$$

The correlogram summary reaches a maximum when (20) becomes equal to 0

$$X = \frac{-n_0}{2m_0} \tag{21}$$

By substituting the X and Y values for the three points centred on the peak into (21), an accurate estimate of the peak lag, and hence the fundamental frequency, can be found.

Segment identification

The correlogram is also used to identify formant and harmonic regions due to their similar patterns of periodicity: contiguous regions of the filterbank respond to the same spectral event. Such contiguous areas of acoustic energy are used to form 'segments' (Wang and Brown, 1999) which correspond to significant acoustic features. This is achieved by computing the cross correlations between adjacent channels of the correlogram as follows:

$$C(i) = \frac{1}{\tau_{max}} \sum_{\tau=0}^{\tau_{max}-1} \hat{A}(i, t, \tau)\hat{A}(i+1, t, \tau) \tag{22}$$

Here, $\hat{A}(i, t, \tau)$ is the autocorrelation function of (13) which has been normalised to have zero mean and unity variance to ensure that *C(i)* is only sensitive to periodicity in the correlogram and not to the mean firing rate of each channel; τ_{max} is the maximum autocorrelation lag in samples (τ_{max} = 160; equivalent to 20 ms). Figure 3c shows the cross correlation function for the 12-harmonic complex tone. It can be seen from the figure that areas of contiguous activity, especially at low frequency, are characterised as having high correlation and that these areas are separated by regions of low correlation.

Once the cross correlation *C(i)* has been computed, it is necessary to decide a 'similarity score' by which adjacent channels are deemed to be sufficiently similar to be grouped together to form a segment. This is achieved by applying a threshold to the energy-weighted cross correlation: adjacent channels whose cross

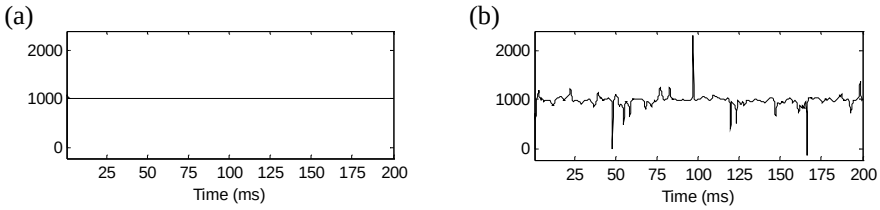


Figure 4 **Instantaneous frequency response of a gammatone filter with a centre frequency of 1 kHz to a 1 kHz pure tone (a) and broadband noise (b). It can be seen that the noise response is subject to significantly more fluctuations than that of the pure tone.**

correlations are above a certain threshold form a segment. In other words, channels i and i+1 are said to contribute to a segment when

$$C(i)A(i,t,0) > \theta_s \tag{23}$$

where θ_s is the segment membership threshold. The cross correlation is energy-weighted in order to increase the contrast between spectral peaks and spectral dips. A high threshold would result in a small number of segments as few adjacent channels would be nearly identical; as the threshold is lowered, so too is the similarity requirement and so similar adjacent channels form segments. These segments are encoded by a binary mask, which is unity when a channel contributes to a segment and zero otherwise.

In order to deal with noise stimuli, an alternative segment formation strategy is used since, by definition, periodicity information cannot be obtained from the correlogram for channels containing noise. Instead, the instantaneous frequency (see Cooke, 1991/1993) of each gammatone filter is used. In response to a pure tone, a channel's instantaneous frequency over time will be stable. However, in response to noise, it exhibits significant fluctuations (see figure 4). This property can be exploited by calculating the inverse variance of the instantaneous frequency in each channel - responses to periodic signals will produce low signal variance and hence high inverse variance. When weighted by channel energy (obtained from the correlogram), a large peak indicates periodic activity in that channel; a smaller peak indicates noise activity (figure 5). The segment estimation process occurs in two stages. First, periodic segments are identified: channels for which a peak exists in the energy weighted inverse variance function which exceeds a given 'tonal' threshold θ_t (see figure 5b). All channels under such peaks are said to constitute a tonal segment; this information is used to confirm the periodic segments identified by the cross-correlation technique described above. These peaks are then removed

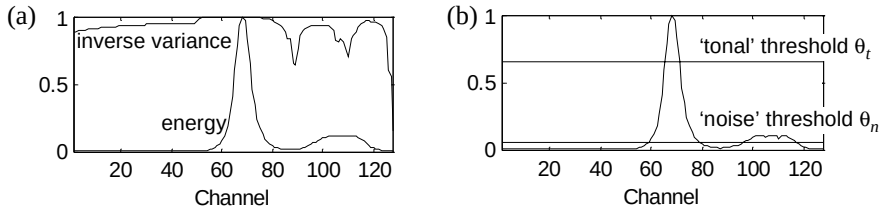


Figure 5 **Response to a mixture containing a 1 kHz tone (centred on channel 68) and a 2 - 3 kHz noise band (centred on channel 107). (a) Inverse variance and channel energy. (b) Energy weighted inverse variance function with tonal and noise thresholds.**

from the energy weighted inverse variance function. The final stage identifies noise segments: channels for which any remaining peaks in the function exceed a second, 'noise', threshold θ_n (see figure 5b). It ought to be noted that the use of instantaneous frequency is not as physiologically plausible as other processing stages in the model but provides a convenient mechanism for representing noise segments as required in later chapters.

To further improve the frequency resolution of the oscillator array, each segment is 'skeletonised': a process by which each segment is replaced by one of fixed frequency channel spread (see also Palomäki *et al.*, 2001). For example, a segment with a channel spread of five, centred on channel 54, would be replaced by one with a channel spread of three, centred on channel 54. This process is similar in principle to lateral inhibition, and leads to a sharpening of the cross-frequency responses thus reducing the tendency for segments which are close in frequency to merge together. This is especially important for elements (such as harmonics) at higher frequencies where the bandwidth of the peripheral filters are relatively large in comparison to the separation of harmonic elements. Since this model investigates how individual acoustic elements are influenced by grouping rules and attentional mechanisms, it is important that each is represented by an individual segment wherever possible. Similarly, since it is only important for the model identify the presence of a noise segment and its location in frequency, it is unnecessary to produce detailed spectral and temporal descriptions of noise regions as seen in previous CASA models (e.g. *noise clouds* of Ellis, 1996). It ought to be noted that this merely alters the *representation* of the stimulus in the model and does not alter the way in which the stimulus is conceptually processed. Figure 3d shows the segments identified by this process from the 12-harmonic complex tone used to generate the correlogram displayed in figure 3a.

Despite such models of periodicity analysis being able to account for a wide range of perceptual phenomena (see Meddis and Hewitt, 1991a, 1991b), their physiological plausibility has been questioned (e.g. Summerfield *et al.*, 1990). One such criticism is the computational complexity of computing an autocorrelation function. However, this assumes a serial processing mechanism which is in direct contrast with biological networks which are highly parallel in nature. Indeed, parallel computational autocorrelation solutions have been shown to achieve real-time performance (e.g. Slaney, 1991). A second criticism has been the lack of evidence for delay lines as suggested by Licklider (1951) for performing the time-delayed coincidence detection required for neural autocorrelations, although such structures are known to underlie binaural localisation (e.g. Konishi *et al.*, 1988; Yin and Chan, 1988). The latter issue may not be problematic since the autocorrelation function can be considered to be similar to the production of an interspike interval histogram (ISIH) of nerve firings. An ISIH represents the distribution of intervals between successive auditory nerve events and tends to be polymodal. The positions of major interval peaks correspond to the period of the pitch that is heard (i.e. the fundamental period for harmonic complexes) and its multiples (e.g. Cariani *et al.*, 1997; Cariani, 1999). These peaks are analogous to the peaks in the summary correlogram which also correspond to the pitch period and multiples thereof. Hence, it ought to be noted that the correlogram model used here should be considered a functional abstraction of the periodicity analysis occurring in the auditory system. Furthermore, this computational model does not place any restrictions on the behaviour of later stages of the implementation.

Neural oscillator network

The network consists of an array of 128 oscillators and is based upon the locally excitatory globally inhibitory oscillator network (LEGION) of Wang described in chapter 3 (e.g. Terman and Wang, 1995; Wang, 1996). Within LEGION, oscillators are synchronised by placing local excitatory links between them. Additionally, a global inhibitor receives excitation from each oscillator, and inhibits every oscillator in the network (see figure 6). This ensures that only one block of synchronised oscillators can be active at any one time. Hence, separate blocks of synchronised oscillators - which correspond to the notion of a segment in ASA - arise through the action of local excitation and global inhibition.

The model described here differs from Wang's (1996) approach in three respects. Firstly, the network is one-dimensional rather than two-dimensional; we have argued, above, that this is more plausible since it avoids the explicit representation of time on a two-dimensional, vision related grid. Furthermore, the adoption of a single dimensional array avoids the attentional multiplexing problems and

difficulties in representing the attentional time course in a two-dimensional network described above. Secondly, excitatory links are global as well as local; this allows harmonically-related segments to be grouped. Finally, we introduce an attentional leaky integrator (ALI), which selects one block of oscillators to become the attentional stream (i.e., the stream which is in the attentional foreground).

The building block of the network is a single oscillator, which consists of a reciprocally connected excitatory unit and inhibitory unit whose activities are represented by x and y respectively:

$$\dot{x} = 3x - x^3 + 2 - y + I_0 \tag{24}$$

$$\dot{y} = \varepsilon\left[\gamma\left(1 + \tanh\frac{x}{\beta}\right) - y\right] \tag{25}$$

Here, ε, γ and β are parameters. As described previously, oscillations are stimulus dependent; they are only observed when $I_o > 0$, which corresponds to a periodic solution to (24) and (25) in which the oscillator cycles between an 'active' phase and a 'silent' phase (see chapter 3, figure 8). The system may be regarded as a model for the behaviour of a single neuron in which x represents the membrane potential of the cell and y represents the inhibitory ion channel activation, or as a mean field approximation to a group of reciprocally connected excitatory and inhibitory neurons. The input I_o to oscillator i is a combination of three factors: external input I_r, network activity and global inhibition as follows:

$$I_o = I_r - W_z S(z, \theta_z) + \sum_{k \neq i} W_{ik} S(x_k, \theta_x) \tag{26}$$

Here, W_{ik} is the connection strength between oscillators i and k; x_k is the activity of oscillator k. The parameter θ_x is a threshold above which an oscillator can affect others in the network and W_z is the weight of inhibition from the global inhibitor z. Similar to θ_x, θ_z acts a threshold above which the global inhibitor can affect an oscillator. S is a squashing function which compresses oscillator activity to be within a range suitable for input to an oscillator:

$$S(m, \theta) = \frac{1}{1 + e^{-K(m - \theta)}} \tag{27}$$

Here, K determines the steepness of the sigmoidal function. The activity of the global inhibitor is defined as

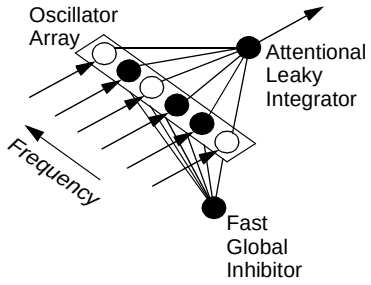


Figure 6 **The one-dimensional, single layer oscillator network performs segmentation and grouping followed by stream selection via the attentional leaky integrator (ALI) to produce an 'attentional stream'.**

$$\dot{z} = H\left(\sum_k S(x_k, \theta_x) - 0.1\right) - z \tag{28}$$

where H is the Heaviside function.

Segment formation and primitive grouping

Oscillators within a segment are synchronized by excitatory connections. The external input (I_r) of an oscillator whose channel is a member of a segment is set to I_{high} (0.2) otherwise it is set to I_{low} (-5).

A further set of connections are made between segments if a majority of channels in each segment are consistent with the pitch estimate as derived above. This involves assessing which segments conform with the pitch estimate. For each segment which covers the frequency channel closest to a harmonic of the pitch estimate, the fine structure of the autocorrelation for each channel in the segment is inspected. If the ratio of channel energy to autocorrelation value at the pitch lag is above a certain threshold θ_c (0.8) the channel is classified as being consistent with the pitch estimate (Brown and Wang, 1997). It is this tolerance in the measure of harmonicity that allows the model to account for the perceptual grouping of harmonics which have been mistuned by limited amounts (see Darwin *et al.*, 1995). In other words, channel i is consistent with a fundamental period of τ_0 when

$$\frac{A(i, t, \tau_0)}{A(i, t, 0)} > \theta_c \tag{29}$$

If the majority of segment channels are consistent, the entire segment is said to be consistent with the pitch estimate. A single connection is made between the centres of all harmonically related segments.

It is at this stage that the *old plus new* heuristic is incorporated into the model. The old-plus-new heuristic (Bregman, 1990) refers to the auditory system's preference to '*interpret any part of a current group of acoustic components as a continuation of a sound that just occurred*'. In our model, '*age trackers*' are attached to each channel of the oscillator array. The age trackers are leaky integrators:

$$\dot{B}_k = d_B(g_B[M_k - B_k]^+ - [1 - H(M_k - B_k)]c_B B_k) \tag{30}$$

Here, B_k is the age of the channel, M_k is the (binary) value of the segment mask at channel k; small values of c_B and d_B result in a slow rise (d_B) and slow decay (c_B) for the integrator. g_B is a gain factor. $[n]^+ = n$ if $n \geq 0$ and $[n]^+ = 0$ otherwise. These parameters have the values d_B=0.001, c_B=5, g_B=3. Excitatory links are placed between harmonically related segments only if the two segments are of similar age. The age of a segment is defined as:

$$AS = \frac{1}{Q} \sum_{k \in segment} B_k \tag{31}$$

where Q is the number of channels in the segment. Two segments are considered to be of similar age if:

$$|AS_1 - AS_2| < \theta_a \tag{32}$$

where AS_1 and AS_2 are the ages of the two segments and the threshold θ_a (0.1) defines the degree of similarity in age between the two segments.

Consider two segments that start at the same time; the age trackers for their constituent channels all begin receiving input at the same time and continue to receive the same input: the values of the leaky integrators will be the same. However, if the two segments had started at different times, the age trackers for the earlier segment will have already built up to a certain value when the second segment starts (whose age trackers will be initially at zero): the two ages will be different.

Attentional Leaky Integrator (ALI)

Each output channel of the oscillator array is connected to the attentional leaky integrator (ALI) by excitatory links (see figure 6); the strength of these connections is modulated by endogenous attention. Input to the ALI is a weighted version of the oscillator array output:

$$\dot{ali} = I_{ALI} - ali \tag{33}$$

Here, I_{ALI} is defined as:

$$I_{ALI} = H\left(\sum_k H(x_k)\left[\left(\frac{\alpha_k}{\theta_\alpha}\right) - T_k\right]^+ - \theta_{ALI}\right) \tag{34}$$

Here, H is the unit step, or Heaviside, function; x_k is the activity of oscillator k in the array. The parameter θ_{ALI} is a threshold above which oscillator array activity can influence the ALI. α_k is the instantaneous envelope of the gammatone filter response to the stimulus at channel k (see (12)). θ_α is a normalising factor which determines how intense a stimulus needs to be to overcome the conscious attentional interest.

T_k is the attentional threshold which is related to the endogenous interest at channel k:

$$T_k = (1 - A_k)L \tag{35}$$

Here, A_k is the endogenous attentional interest at channel k and L is the leaky integrator defined as:

$$\dot{L} = d_L(g_L[R - L]^+ - [1 - H(R - L)]c_L L) \tag{36}$$

Here, small values of c_L and d_L result in a slow rise (d_L) and slow decay (c_L) for the integrator. g_L is a gain factor. These parameters have the values d_L=0.0005, c_L=5, g_L=3. R is given by

$$R = H(x_{max}) \tag{37}$$

where x_{max} is the largest output activity of the oscillator array.

In accordance with the experimental findings of Mondor and Bregman (1994), the attentional interest is modelled as a gaussian:

$$A_k = max_{A_k} e^{-\frac{(k-p)^2}{2\sigma_{ALI}^2}} \tag{38}$$

Here A_k is the attentional interest at frequency channel k; max_{Ak} is the maximum value that A_k can attain; p is the channel at which the peak of attentional interest occurs and σ_{ALI} determines the width of the peak. In order to allow segments which are outside of the attentional interest vector, but are sufficiently loud, to overrule the attentional selection, the A_k vector must be non-zero to both sides of the peak. Hence, a minimum A_k value of min_{Ak} is enforced:

$$A_k = \begin{cases} min_{A_k} & A_k < min_{A_k} \\ A_k & otherwise \end{cases} \tag{39}$$

In the model, a segment or group of segments are considered to be attended to if their oscillatory activity coincides temporally with a peak in the ALI activity. In other words, their connection strengths to the ALI are sufficiently large to promote activity within the ALI. Initially, the connection weights between all oscillators in the array and the ALI are strong and hence all segments feed large amounts of excitation to the ALI. This means that initially all segments contribute to the attentional stream representing the default grouping percept of fusion (Bregman, 1990).

During sustained activity in the oscillator array, these weights relax toward the A_k interest vector such that strong weights exist for channels of high attentional interest and low weights exist for channels of low attentional interest. This relaxation toward the A_k interest vector is achieved by the use of the leaky integrator L. Thus, after a finite period of time, oscillators which are desynchronised to ones within the attentional interest (e.g. harmonically unrelated) will be subject to low connection weights to the ALI and will be unlikely to overcome the θ_{ALI} threshold required to influence the ALI. Such ‘relaxation’ of the connection weights towards the attentional interest vector simulates the period of build-up observed in auditory streaming (Anstis and Saida, 1985). ALI peaks will only coincide with activity within the attentional interest peak and any perceptually related (i.e. synchronised) activity outside the A_k peak. All other activity will occur

within a trough of ALI activity. This behaviour allows both individual tones and harmonic complexes to be attended using only a single A_k peak.

A table of model parameters can be found in appendix A.

6.3 Interim summary

The previous section has described the computational model which accounts for monaurally presented stimuli. It consists of three main processing stages: peripheral analysis, harmonicity and segmentation analysis and, finally, the neural oscillator array and attentional leaky integrator (ALI). The first stage - peripheral analysis - simulates the behaviour of the outer and middle ears followed by a model of the frequency analysis carried out by the cochlea. The output of this stage is an approximation to the activity of the auditory nerve at 128 centre frequencies ranging from 50 to 3500 Hz.

The second stage of the model uses this activity to produce a correlogram by computing the autocorrelation of each channel for a range of delays. This produces a two dimensional representation in which frequency is expressed on the ordinate and autocorrelation delay on the abscissa. For each channel, the first largest autocorrelation peak occurs at the period of the dominant frequency in that channel. When summed over frequency, an estimate of the dominant frequency can be obtained. In addition to estimating the dominant pitch period, the degree of correlation between adjacent channels' activities can be used to determine the presence of a region of contiguous acoustic energy: a segment (c.f. acoustic elements of Bregman, 1990). If the activities of adjacent channels exhibit a high degree of correlation it is highly likely that they are responding to the same spectral event and hence those channels are said to form a segment.

The final stage of the model consists of an array of 128 neural oscillators, each representing a particular channel centre frequency. Input to each oscillator is determined by the presence of a segment in that channel. Only oscillators which receive input are able to produce oscillations. Excitatory connections are placed between oscillators representing the same segment hence forming a synchronised block of oscillators. In addition to these, excitatory connections are also placed between segments whose channels agree with the dominant pitch estimate obtained from the correlogram. Segments which are connected by such connections then form a synchronised group of oscillator blocks. Segments which are not connected in this way become desynchronised.

Each oscillator in the array is connected to the attentional leaky integrator (ALI) by connections, the strength of which are modulated by conscious attentional interest expressed by A_k. Since the efficacy of this attentional interest requires a finite period of time to build up, the connection strengths are initially maximal to simulate the 'default' percept of fusion (see Bregman, 1990). In this case, all channels of the oscillator array feed input to the ALI and hence ALI activity is observed to be synchronous with all segments. Over time, only connections which are subject to high attentional interest can influence the ALI and so only oscillator activity corresponding to the attentionally selected frequency region can influence the ALI. From this, it can be said that any segment whose oscillator activities coincide with ALI activity are attended to. It is important to note that since harmonically related segments are temporally synchronised, if ALI activity coincides with one segment, it will also coincide with the activities of any other harmonically related segments. Thus, they are also said to be attended to. Since the lowest value of the attentional interest vector A_k is non-zero, extremely loud stimuli which are outside of the attentional interest can also influence the ALI allowing the model to account for the unconscious reorientation of attention caused by, for example, a loud bang.

6.4 Binaural computational model

In order to account for binaural grouping phenomenon (e.g. Carlyon *et al.*, 2001; Darwin *et al.*, 1995) the computational model must be able to process binaural signals. Therefore, a second processing pipeline is required for the additional signal. Figure 7 shows the schematic structure of the binaural model. Each peripheral processing pipeline incorporates basilar membrane processing and an approximation to inner hair cell transduction. Furthermore, each pipeline has its own correlogram which is used to extract periodicity information and contiguous regions of activity for segment formation. This periodicity information is then combined to produce a 'binaural' F0 estimate (e.g. Moore, 1997, p. 208) to enable binaural harmonicity grouping. The structure and functionality of these stages is identical to that of the monaural model described above. The binaural model differs in the structure of the neural oscillator network.

Figure 8 shows the network topology in more detail. Each ear has an individual oscillator array which functions in the same way as the one described for the monaural model. The mechanism by which segments in opposite ears are grouped together is described below. Importantly, both oscillator arrays share a single global inhibitor to ensure that only perceptually related groups of segments (potentially in

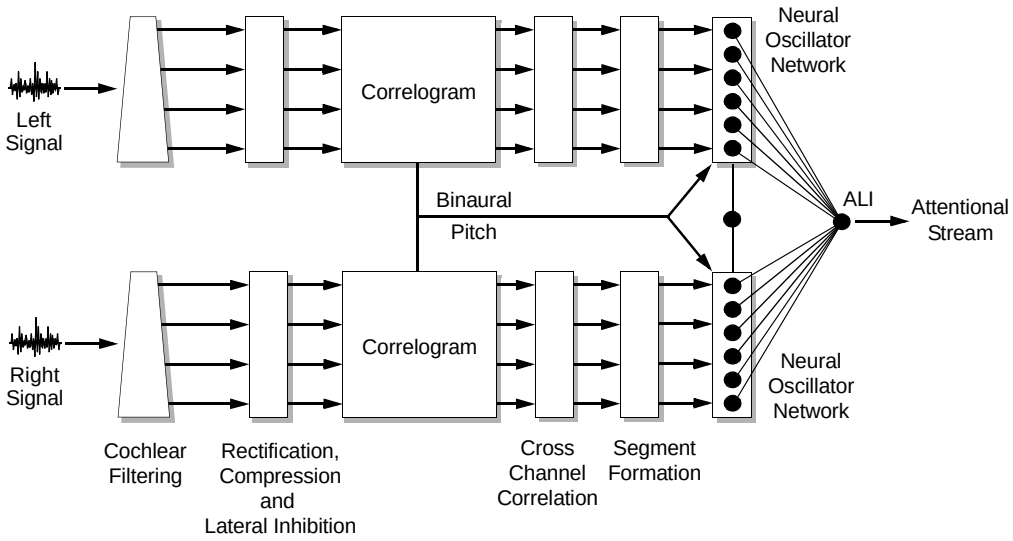


Figure 7 **Schematic diagram of the binaural model (the attentional leaky integrator is labelled ALI).**

different ears) are active simultaneously. In other words, the common global inhibitor ensures that unrelated activity, in either the same or different ears, does not occur simultaneously. Similarly, both ears share a single attentional leaky integrator (ALI) representing the central decision process.

Segment grouping

As described previously, Darwin *et al.* (1995) investigated the effect of a mistuned harmonic upon the pitch of a 12 component complex tone. As the degree of mistuning of the fourth harmonic increased towards 4%, the shift in the perceived pitch of the complex also increased. This effect was less pronounced for mistunings of more than 4%; beyond 8% mistuning, little pitch shift was observed. Apparently, the pitch of a complex tone is calculated using only those channels which belong to the corresponding stream. When the harmonic is subject to mistunings below 8%, it is grouped with the rest of the complex and so can affect the pitch percept. Mistunings of greater than 8% cause the harmonic to be segregated into a second stream, and so it is excluded from the pitch percept. This effect was found when the mistuned harmonic was presented both ipsilaterally and contralaterally.

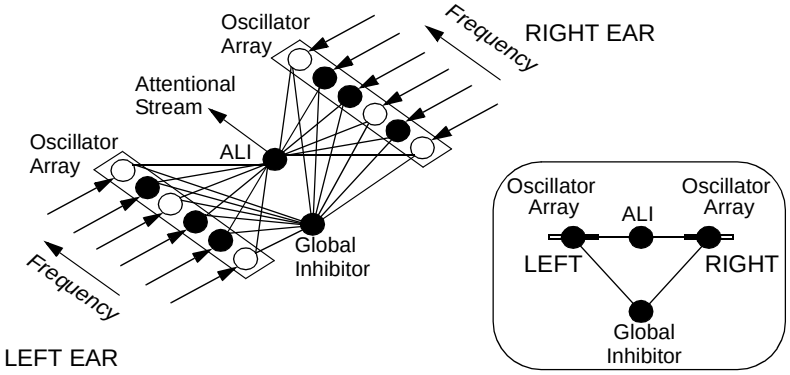


Figure 8 **An oscillator array performs segmentation and grouping for each ear. Stream formation and selection occurs by combining the activities of both layers to form a single (binaural) attentional stream via the ALI.**

This suggests that the auditory system can assess whether a tone in one ear ought to be grouped with a contralateral sound on the basis of harmonicity. In the monaural computational model, a harmonic is deemed to belong to a complex if it is consistent with the overall pitch estimate. Hence, it is assumed that binaural harmonicity grouping proceeds in the same manner: an overall pitch estimate from both ears is calculated and harmonic grouping is performed using this pitch estimate. Computationally, a 'binaural correlogram' is formed by aggregating the channels from both ears (c.f. Moore, 1997, p. 208). The summary s_b is then found by summing all the channels in the binaural correlogram:

$$s_b(t, \tau) = \sum_{e=0}^{1} \sum_{i=0}^{N-1} A_e(i, t, \tau) \tag{40}$$

where $A_e(i,t,\tau)$ is the autocorrelation of channel i at time t with lag τ for ear e. The summary is then found using the same techniques described above and used to estimate the dominant fundamental frequency. Harmonics which agree with the fundamental frequency estimate have excitatory links placed between them (as in the monaural system). Additionally, links are also placed between contralateral segments if they are harmonically related.

Attentional Leaky Integrator (ALI)

As described in chapter 4, Carlyon *et al.* (2001) demonstrated that auditory streaming (the tendency to segregate high and low tones from an alternating sequence on the basis of frequency separation and repetition rate) did not occur when listeners attended to an alternative stimulus presented simultaneously. However, when they were instructed to attend to the tone sequence, auditory streaming occurred as normal. From this, it was concluded that attention is required for stream formation and not only for stream selection.

In our model, attention can be directed toward a particular ear at the expense of the other in addition to the 'no preference' situation in which attention is allocated to both ears equally. Computationally, each ear has an attention gain, ranging from 0 to 1, representing ear preference A_{ear} such that (35) becomes

$$T_k = (1 - A_{ear} A_k) L \tag{41}$$

In principle, such a mechanism should account for both the Darwin *et al.* (1995) and Carlyon *et al.* (2001) data. The Darwin study investigated the effect of a mistuned fourth harmonic upon the pitch of a 12 component complex tone. The degree of mistuning was randomly selected (ranging from 0% to 8%) and the ear of presentation of the fourth harmonic also varied randomly (ipsilateral or contralateral to the remainder of the complex). Since the stimulus presentation was alternated randomly between the ears, listeners were unable to become accustomed to a particular presentation side: both ears had to be attended and thus A_{ear} would be high for both ears. Thus, binaural grouping of the complex and fourth harmonic in the model proceeds as in the monaural situation, with excitatory links placed between the two provided they satisfy the old-plus-new and harmonicity constraints.

In the Carlyon study, listeners are instructed to concentrate on a stimulus in a particular ear. Initially, this is a sequence of noise bursts with increasing or decreasing amplitude; the task is to classify the amplitude ramp. After a period of time (ten seconds) the subject is told to concentrate on a sequence of alternating tones in the other ear. In this situation, listeners showed a greatly reduced amount of streaming relative to the situation in which they are allowed to listen to the alternating tones from the beginning of the task. In the model, a particular ear is given a higher attentional weighting than the other by making A_{ear} high for the attended ear and low for the other; only activity in this ear can affect the ALI. Initially, the ear corresponding to the noise bursts is dominant and the ALI is stimulated by the noise segments. After the ten second period, the ear dominance is

swapped and the attentional interest peak is moved into the tone frequency region. When an ear dominance (A_{ear}) change is detected, the model's attentional mechanism 'resets' - the leaky gain factor L on the A_k vector described in (36), used to model the build up of the attentional effect, is reset to zero. Hence, after the change, the attentional interest requires time to build up before streaming of the two tones can be observed.

6.5 Summary

This chapter has described the computational implementation of the conceptual model described in the previous chapter. The model makes a number of key assumptions:

- A binaural pooling of pitch to allow across-ear grouping by harmonicity.
- Attentional build-up is modelled as a form of leaky integrator with build-up and decay timescales in the order of seconds.
- Resetting of attentional build-up when the focus of attention moves between ears.
- The old-plus-new heuristic is embodied in low-level mechanisms.

The model consists of three core stages, namely auditory peripheral processing, periodicity analysis and finally the neural oscillator network. Once the signal has been processed to simulate cochlear filtering and auditory nerve encoding mechanisms, periodicity information is extracted using the autocorrelation-based correlogram technique. Such information allows both the F0 of the signal to be extracted as well as identify areas of contiguous periodic energy which correspond to Bregman's (1990) concept of acoustic elements. Further analysis of the cochlear filtering outputs allow noise segments (i.e. non-periodic regions of energy) to be identified.

Within the network, each oscillator in the array corresponds to a particular frequency channel; segments are encoded by placing excitatory connections between the relevant oscillators which promote their individual activities to synchronise. These segments are then grouped on the basis of common harmonicity by using further excitatory connections between constituent oscillators. Each oscillator is connected to the attentional leaky integrator (ALI) by means of connections whose strengths are modulated by 'conscious' attentional interest: maximum strength occurs at the frequency of highest interest and decay over

frequency in a gaussian manner. Such a decay is only observed over a period of time in order to model the finite time course of auditory stream buildup reported by Anstis and Saida (1985). Only the activity of oscillators representing frequency channels of high attentional interest can influence the ALI and hence any activity synchronised with the ALI is said to contribute to the attentional stream. The following chapter will present a number of simulations of published experimental stimuli which demonstrate the behaviour of the model.

Chapter 7. Evaluation

7.1 Introduction

The output of the model is evaluated by inspecting the time course of the neural oscillator array activity and the attentional leaky integrator (ALI). This information allows the behaviour of the grouping and attentional processes to be compared with the findings of the relevant psychophysical studies. The simulations are presented in terms of monaural and binaural content. However, before presenting the results, we describe the format in which the oscillator array and ALI outputs will be presented.

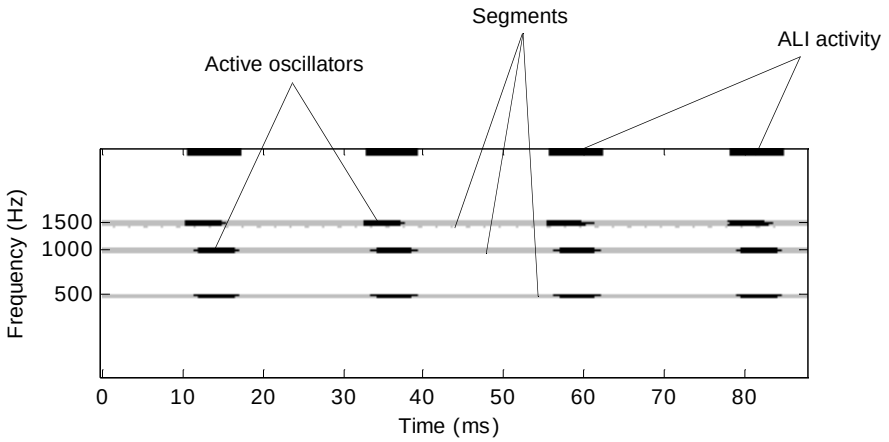


Figure 1 **Sample output of the oscillator array in response to a 3-harmonic complex tone with a fundamental of 500 Hz and a duration of 90 ms.**

Output representation

The activity of the oscillator array over the time course of a stimulus is represented in a pseudospectrogram format, as shown in figure 1. Channel centre frequency is represented on the ordinate and time is represented on the abscissa. Each diagram contains three types of information: the location in time-frequency of segments (gray pixels), the times at which channel oscillators are active (black pixels) and the activity of the ALI (black blocks along the top of the diagram). Any oscillators which are temporally synchronised with the ALI are considered to be in the attentional foreground. The presence of a segment is also equivalent to the external input to the corresponding oscillator: gray signifies I_{high} (causing the oscillator to be stimulated) and white signifies I_{low} (causing the oscillator to be unstimulated). Such diagrams are constructed on a frame by frame basis: after each stimulus sample has been processed, the state of the oscillator array is recorded and forms a vertical slice in the diagram.

As described in the previous chapter, the oscillator array consists of 128 neural oscillators, each representing one frequency channel. It is also important to recall that these channels are spaced quasi-logarithmically on the equivalent rectangular bandwidth (ERB) scale. The x activity of an oscillator is used to assess its state; figure 2 shows a typical x activity trace of a stimulated oscillator. When inspecting such traces, the important information to be extracted is the relative times at which the oscillator becomes active. Two oscillators are said to be synchronised, and

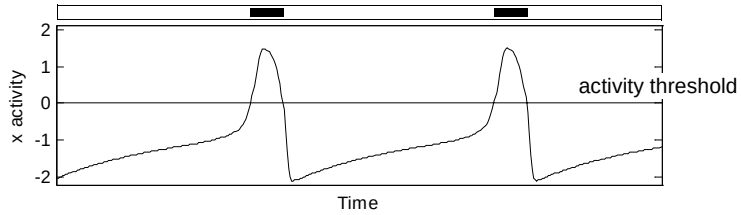


Figure 2 **Temporal activity of a neural oscillator with activity threshold. The representation of the oscillator's time course, which will be used in subsequent figures, is shown in the top set of axes. Black pixels represent times at which the oscillator is in the active phase and above a given threshold; white pixels represent times at which the oscillator is in the silent phase.**

hence perceptually grouped within the model, if they are active simultaneously. Since the output of the oscillator array involves 128 channels, possibly over a significant period of time, displaying the 'raw' *x* activity of each oscillator would produce an overly complicated diagram. The technique used in figure 1, and throughout the remainder of this chapter, is to apply an activity threshold to produce a binary output for each oscillator such that a value of 1 represents the active phase, and 0 represents the silent phase (see figure 2).

7.2 Monaural stimuli

In this section, we will demonstrate how the model performs monaural grouping and attentionally modulated streaming. The role of attention in the stream formation and segregation process is investigated by the model's response to an alternating tone sequence such as the one used by van Noorden (1975). The model requires the allocation of attentional interest for such a sequence to segregate provided the frequency separation of the tones is sufficiently large. Grouping is examined using the Darwin *et al.* (1995) mistuned harmonic stimuli in which varying degrees of mistuning have differing effects upon the perceived pitch of a complex tone. This type of stimulus is also used to demonstrate how the old-plus-new (Bregman, 1990) heuristic influences grouping by using asynchronous harmonic onsets and captor tones to segregate a harmonic from a complex tone.

Two tone streaming

The ABA-ABA two tone streaming phenomena, which has been described in previous chapters, consists of a sequence of pure tones A and B whose frequencies

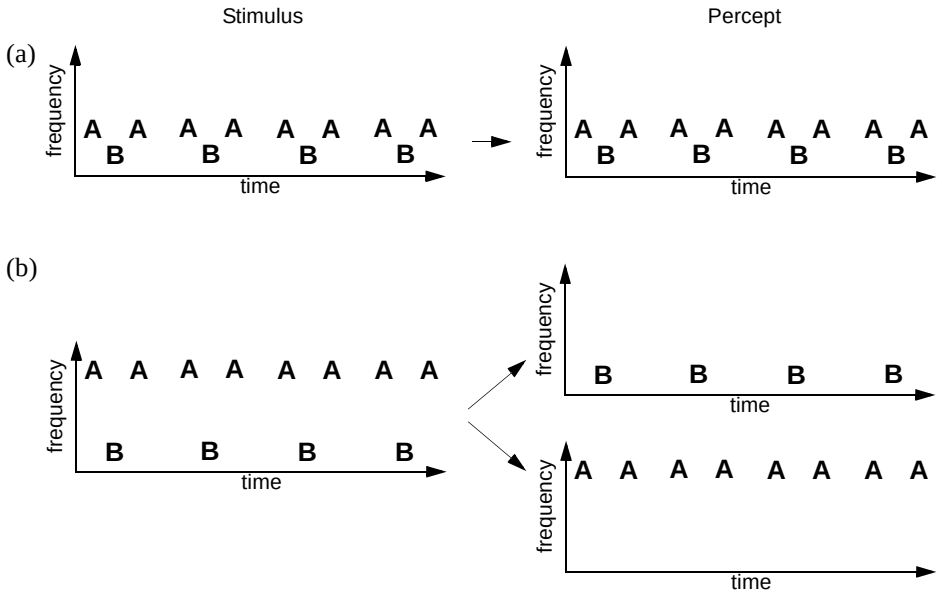


Figure 3 **Percept of an alternating sequence of tones, A and B, at two different frequency separations. When presented at low frequency separation, the tones form a single perceptual stream: temporal coherence (a). When the separation is increased, the tones are more likely to form two perceptual streams: auditory streaming (b).**

alternate over time. It has been shown that if the frequency separation of the tones is sufficiently large and the repetition rate is appropriate, the sequence will segregate into two streams, one consisting of A-A-A and the other of B---B---B where '-' indicates silence (van Noorden, 1975; see figure 3). In addition to this, it has been observed that such *auditory streaming* (Bregman, 1990) requires a finite period of time to become evident (Anstis and Saida, 1985): up to ten seconds can elapse before subjects report the percept of streaming.

Carlyon *et al.* (2001) have shown that when listeners are instructed to divert their attention away from the alternating tone sequence, streaming does not occur. From these findings, Carlyon *et al.* suggested that attention is a prerequisite for auditory streaming to occur. We have argued in previous chapters that the deployment of endogenous attention is a fundamental part of the stream formation process and hence can account for the segregation of the alternating tones in this example. Figure 4 shows the response of the oscillator network to an alternating tone sequence. In figure 4a, the frequency separation of the tones is 200 Hz

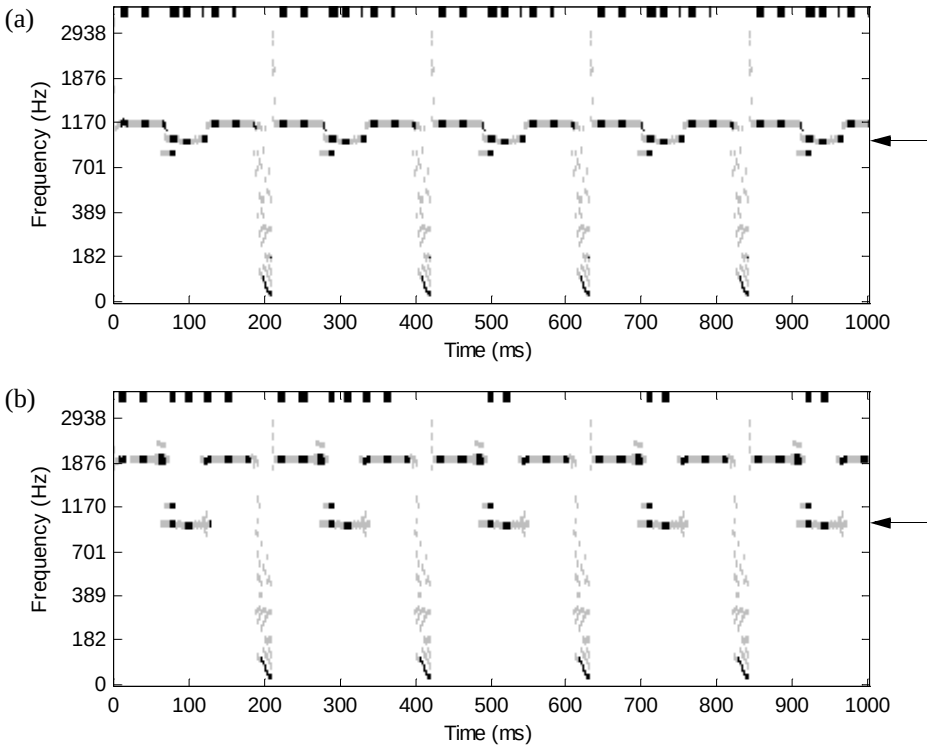


Figure 4 **Model response to ABA-ABA alternating tone sequences. In both sequences, the tones are of duration 50 ms with a intertone gap of 5 ms and an inter ABA gap of 50 ms. In both examples, tone B is 1000 Hz and endogenous attention is directed toward tone B (indicated by an arrow). In example (a), tone A is 1200 Hz. Upon inspection of the ALI activity, both the A and B tones are the subject of attentional interest throughout the signal: ALI activity is synchronous with the oscillator activity of both tones indicating both tones are in the attentional stream: temporal coherence (c.f. figure 3a). In example (b), the separation is increased such that tone A is 2000 Hz. In this situation, it can be seen that the percept of streaming builds up until half way through the signal, ALI activity only occurs during tone B: two streams have formed, one containing the B tones and the other the A tones (c.f. figure 3b). Since attention is directed toward the B tones, it is this group that becomes the attentional stream. To improve the clarity of the diagrams, the rate of endogenous adaptation has been increased so that streaming will occur over a rapid timescale.**

(approximately 3.5 semitones); in this situation, van Noorden (1975) showed that the percept of temporal coherence predominated such that both the A and B tones of the ABA sequence were contained in the same stream. For two events, separated in time, to be grouped in our model, their corresponding oscillator activities must

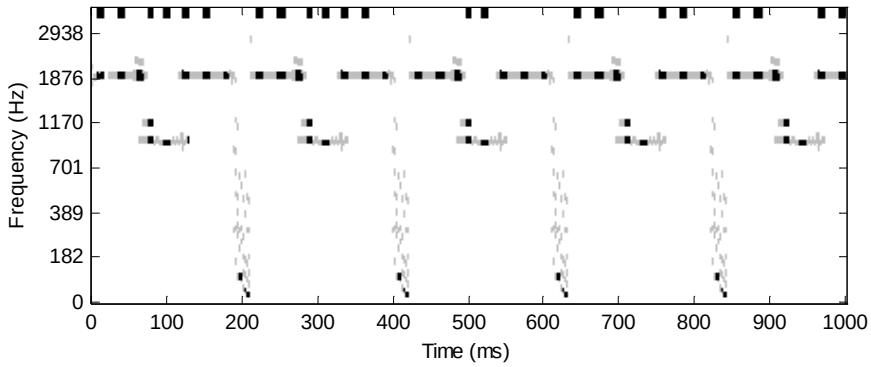


Figure 5 **Model response to the ABA-ABA alternating tone sequence presented in figure 4b. To demonstrate the ability to attend to different streams, the attentional focus is moved from the B tones to the high frequency A tones after segregation had occurred (after 600 ms).**

be synchronised to the activity of the ALI. If so, they can be said to belong to the attentional stream. ALI activity in figure 4a is synchronised to all the tones throughout the signal: both the A and B tones are present in the attentional stream and are hence perceived to be grouped: temporal coherence. By contrast, when the frequency separation of the tones is increased to 1000 Hz (13 semitones), the reported percept is one of auditory streaming in which the A and the B tones have been segregated; attention can then be directed to one of these streams. This behaviour would be indicated in the model by the activity of the ALI being synchronised to the response of only one set of tones. For example, if endogenous attention had been directed to the low frequency tones after the build-up of streaming had occurred, ALI activity would be seen only when the oscillators corresponding to the low frequency tone are active. This is seen in figure 4b in which endogenous attention is directed to the 1000 Hz tone. After a period of time, ALI activity only occurs when the oscillators corresponding to the 1000 Hz tone are active; the ALI is silent during periods of tone A oscillator activity. In this case, streaming can be considered to have occurred and the attended stream contains only the low frequency B tones.

An additional property of auditory streaming which can be observed in the model is the ability to change the stream to which attention is directed without affecting the perceptual organisation of the signal. For example, once a two tone sequence has segregated, listeners can decide which stream to attend to and switch between streams at will. In our model, this is achieved simply by moving the focus of

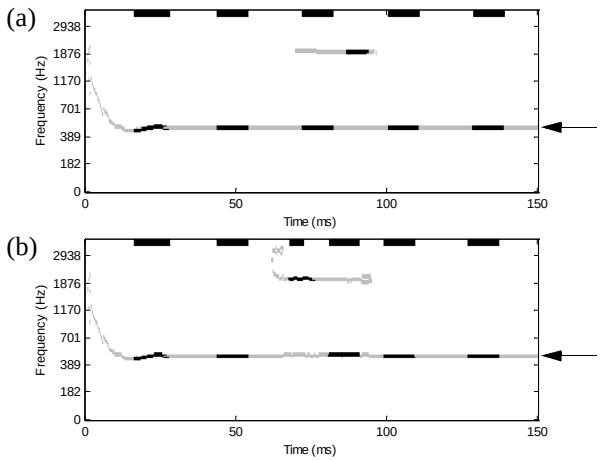


Figure 6 **Unconscious, exogenous, redirection of attention due to a loud stimulus. In (a), the later, high frequency, tone blip is not sufficiently loud to redirect endogenous attention away from the continuous low frequency tone. However, in (b), the intensity of the high frequency tone blip has been significantly increased such that the ALI is influenced by this tone and it becomes the attentional stream. In both cases, endogenous attention has fully built up and is focused on the low frequency tone (indicated by an arrow).**

attentional allocation A_k such that the peak of the gaussian falls over the alternative stream. To demonstrate this property, the two tone sequence used in figure 5 is used. Initially, the focus of attention is directed toward the low frequency B tones causing the ALI activity to be synchronised with the low frequency tone oscillators once the streaming percept has built up. However, after this has occurred, the focus of attention is moved to the high frequency A tones and they become the attentional stream with the B tones no longer being attended to. The switch in attentional focus occurs after 600 ms (see figure 5). ALI activity is synchronised to the high frequency A tones for the remainder of the signal.

It is evident from figure 5 that, occasionally, oscillator activities at the end of a segment do not influence the ALI, despite the segment being under the attentional focus. This is due to the 'dying off' of segments in which there is enough energy in the auditory nerve representation to justify a segment being formed but that this value, once scaled by θ_α, is insufficient to influence the ALI as described in (34) of chapter 6.

Redirection of attention

The unconscious, exogenous, redirection of attention due to a loud stimulus can be demonstrated by presenting the model with a pure tone and a tone blip. In the first condition, the tone blip is less intense than the attended tone. In this case, attention will not be redirected toward the new tone: attention remains with the consciously selected one. Figure 6a shows that the ALI is only influenced by the endogenously selected low frequency tone. However, when the intensity of the tone blip is increased in the second condition, endogenous selection is overruled and the ALI is influenced by the tone blip. This occurs since the intensity of the blip is sufficiently high to overcome the intensity threshold θ_α (see chapter 6) and hence the oscillators corresponding to the blip are able to provide input to the ALI. In this situation, the blip momentarily becomes the attentional stream. As can be seen in figure 6b, the low frequency, endogenously selected, tone becomes the attentional stream again after the exogenous reorientation. A future enhancement to the model could be to detect this override and make a decision on whether to stay with new stimulus or return to original by changing the A_k vector. This would be equivalent to a listener making a brief analysis of the interrupting sound and deciding whether or not it merited further investigation. Only if the new sound appeared to be more salient than the currently attended sound, would the A_k vector be consciously altered.

Harmonic segregation within a complex tone

Darwin *et al.* (1995) investigated the effect of a mistuned harmonic upon the pitch of a 12 component complex tone. As the degree of mistuning of the fourth harmonic increased towards 4%, the shift in the perceived pitch of the complex also increased. This effect was less pronounced for mistunings of more than 4%; beyond 8% mistuning, little pitch shift was observed. This suggests that the pitch of a complex tone is calculated using only those channels which belong to the corresponding stream. When the harmonic is subject to mistunings below 8%, it is grouped with the rest of the complex and so can affect the pitch percept.

For mistuning greater than 8%, the lack of influence on the pitch percept exerted by the mistuned harmonic implies that it has been perceptually segregated from the representation of the complex. In other words, for mistunings below 8%, a single group exists; mistunings beyond 8% result in two groups.

This behaviour is reproduced by our model; figure 7 shows the response of the network to stimuli subjected to mistunings of 0%, 5%, 7% and 8%.

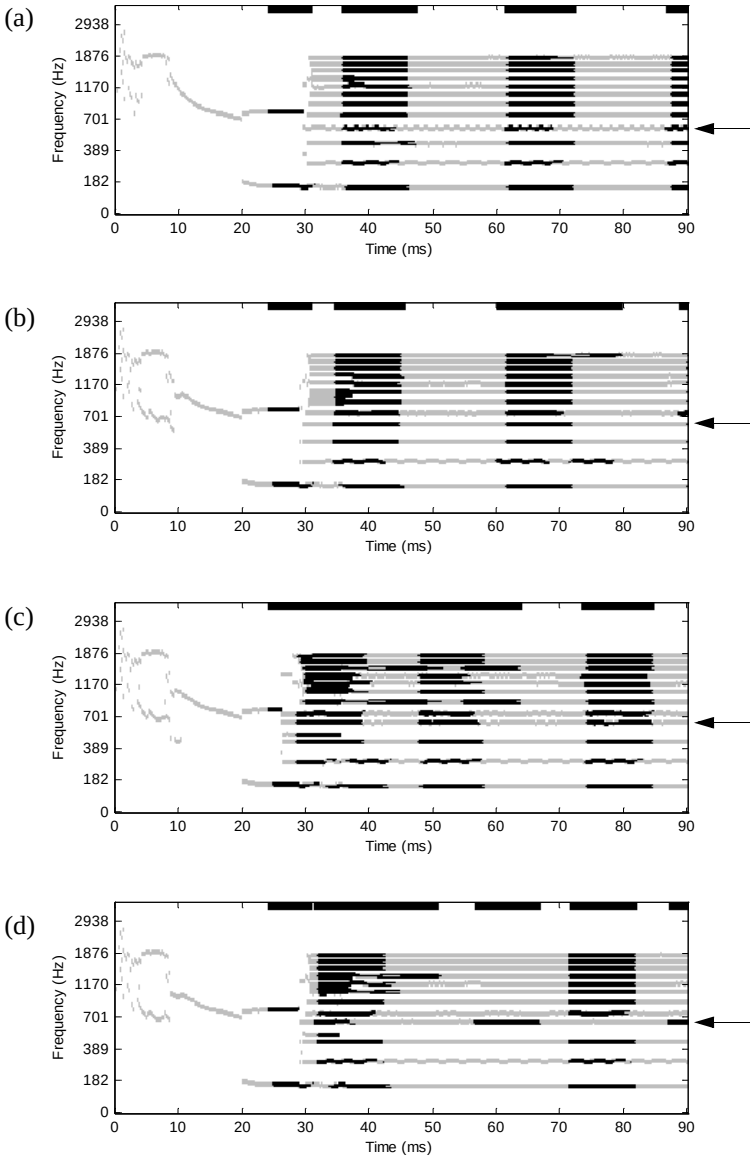


Figure 7 **Model response to a 12-harmonic complex tone whose 4th harmonic (indicated by an arrow) is mistuned by varying degrees: 0% (a), 5% (b), 7% (c) and 8% (d). Attentional interest is focused on the fourth harmonic.**

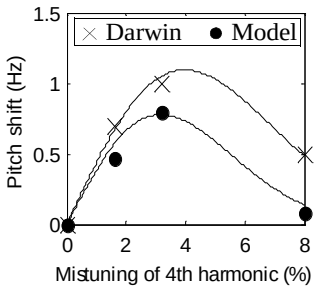


Figure 8 **Pitch shift versus degree of mistuning. A Gaussian derivative is fitted to each data set. Experimental data from Darwin *et al.* (1995).**

For mistunings of less than 8%, the network forms a single perceptual group containing all the harmonics which becomes the attentional stream (figure 7a). Recall that the diagram is a pseudospectrogram which displays information about the position in time and frequency of segments (gray areas), oscillator activities (black areas superimposed over segments) and ALI activity (black blocks along the top of the diagram). In figure 7a, the 12 harmonics of the 155 Hz complex - represented as segments - can clearly be seen. The activities of their corresponding neural oscillators exhibit temporal synchrony, indicating that all the harmonics have been grouped. Similar behaviour is observed when the harmonic is mistuned by 5% and 7%, although a greater degree of initial 'jitter' is observed before the network settles. This is especially so for 7% where clear synchronisation is only achieved after 70 ms. Nevertheless, synchronisation does occur and the harmonic is said to be perceptually grouped with the complex even at this relatively high degree of mistuning. When the degree of mistuning reaches 8%, the fourth harmonic is segregated from the rest of the complex: the oscillators corresponding to the fourth harmonic are temporally desynchronised from the remaining oscillators (figure 7d). As described in the previous chapter, it is the activity of the global inhibitor that prevents both groups from being active simultaneously. Hence, at any point in time, only oscillators corresponding to one of the groups may be active. Two distinct perceptual groups are now present: one containing the fourth harmonic and the other containing the remainder of the complex tone. ALI activity displayed at the top of the diagrams in figure 7 indicate that both the complex and the segregated harmonic are attended to since the stimuli are of insufficient duration for attentional build up to occur.

A comparison of the pitch shifts found by Darwin *et al.* (1995) and the shifts predicted by the model is shown in figure 8. The pitch of the complex was calculated by creating a summary correlogram (similar to that described in the

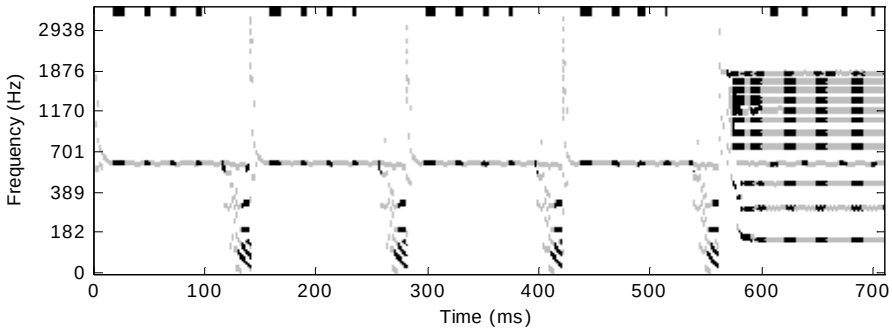


Figure 9 **Model response to a 12-harmonic complex tone whose 4th harmonic is preceded by four 'captor' tones. Note that despite the fourth harmonic being harmonically related to the complex, it is segregated (indicated by temporal desynchronisation) by virtue of the captor tones.**

previous chapter) using frequency channels contained within the complex tone group. Only segment channels below 1.1 kHz were used for this summary since low frequency (resolved) harmonics are known to dominate the pitch percept (see Moore, 1997). Indeed, Moore *et al.* (1985) showed that mistuning of low harmonics had a much greater influence over pitch percepts than the mistuning of high harmonics. They found that above the 6th harmonic, there was very little effect on the pitch percept implying that only lower harmonics are used to produce a percept of pitch

Darwin *et al.* (1995) also showed that the effect of mistuning was almost eliminated when the fourth harmonic was 'captured' from the complex by four preceding tones at the same frequency. In this situation, no matter how small the mistuning, the harmonic was segregated from the complex and did not influence the pitch percept. Figure 9 shows the capture of the fourth harmonic even when there is no mistuning. During the 550 ms before onset of the complex tone, the age tracker activities B_k for the captor tone channels build up. When the complex tone begins, there is a significant age difference between the frequency channels stimulated by the fourth harmonic and those stimulated by the remainder of the complex. Such a difference prevents excitatory harmonicity connections from being made between the fourth harmonic and the remaining harmonics. It can be seen in the diagram that the oscillator activities corresponding to the fourth harmonic temporally desynchronise from those of the remainder of the complex tone. This behaviour is consistent with the old-plus-new heuristic; a current acoustic event is interpreted as a continuation of a previous stimulus unless there is evidence to do otherwise.

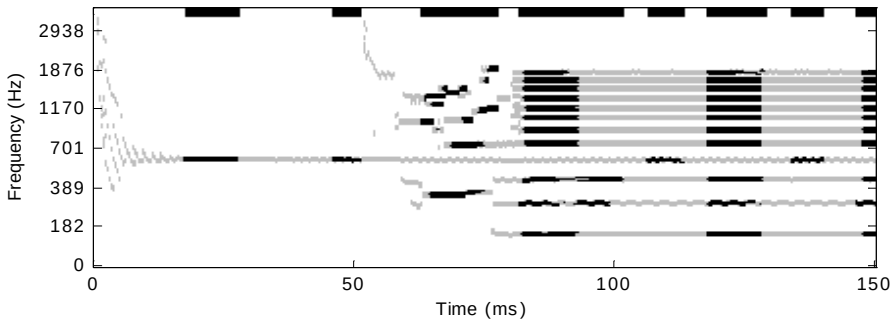


Figure 10 **Response of the model to a 12 harmonic complex tone with fundamental frequency of 155 Hz whose fourth harmonic begins 50 ms before the remainder of the complex.**

The old-plus-new heuristic can be further demonstrated by starting the fourth harmonic before the rest of the complex. Figure 10 shows the output of the model when the fourth harmonic is subject to a 50 ms onset asynchrony. During this time, the age trackers of channels excited by the fourth harmonic increase to a significantly higher value than those of the remaining harmonics. This is the same mechanism by which captor tones, in the previous example, caused the harmonic to segregate. Once again, this difference in segment activity age prevents excitatory connections from being made between the fourth harmonic and the other harmonically related segments. Thus, the early harmonic is desynchronised from the rest of the complex and two groups are formed. However, after a period of time, the importance of the onset asynchrony decreases as the channel ages approach their maximal values. Once this occurs, there is no longer any evidence to prevent excitatory links from being made between the fourth harmonic and the rest of the complex. Grouping by harmonicity then occurs for all segments: the complex and the early harmonic synchronise to form a single stream. This is consistent with experimental data (e.g. Darwin and Ciocca, 1992) in which the effect of segregation on the pitch shift of a complex due to asynchronous onset is reduced for stimuli of longer duration. In other words, the proportion of time that the harmonic is segregated relative to the overall duration of the stimuli decreases resulting in a lower overall pitch shift effect.

Timecourse of attentional build-up

As demonstrated by the model's response to an alternating tone sequence the allocation of attention to a particular frequency region occurs over a period of time

with maximal efficacy only occurring after a number of seconds. Initially, all frequency regions contribute equally to the attentional stream and hence all segment groups are attended to. Gradually, the attentional allocation becomes more focused until the desired attentional allocation, defined by A_k, is achieved. At this stage only segment groups under the focus of attentional allocation are attended to. In the previous simulations, we investigated how mistuned harmonics, asynchronous onsets and the use of captor tones influence the grouping of a harmonic complex. The durations of the stimuli used in these simulations were between 90 ms and 700 ms: much shorter than the time period required for attentional build up as demonstrated previously. As such, for the very short stimuli, both the complex and segregated harmonic are contained within the attentional stream. This means that each group is attended to equally, a percept that is observed during informal listening. The simulations in which captor tones are used to segregate a harmonic from a complex tone, the time constant controlling the attentional build up, d_L (see (36) in chapter 6), is adjusted such that the segregated harmonic becomes the attentional stream. We argue that this is plausible since cognitive processing of sound can occur at different timescales for different tasks (ranging from microseconds for sound source localisation to many seconds for stream formation and segregation).

7.3 Binaural stimuli

This section presents similar results to those described in the previous section with the exception that these stimuli exercise the model's ability to perform binaural grouping and attentional allocation. Binaural attentional allocation is simulated using the experimental stimulus of Carlyon *et al.* (2001) in which an alternating tone sequence is presented in the opposite ear to a distractor task. When instructed to perform a classification task on the distractor stimulus, the streaming percept failed to build up for the alternating tone sequence. However, when told to ignore the distractor task, segregation occurred as normal suggesting that attention plays a key role in the streaming process. Such behaviour is seen in the performance of our model. Binaural grouping is investigated using a stimulus consisting of a mistuned harmonic within a complex tone (Darwin *et al.*, 1995) is used to show how the degree of mistuning of a contralaterally presented harmonic can still influence the perceived pitch of a complex; i.e. the contralateral harmonic is being grouped with the remainder of the complex tone. Similarly, the deployment of asynchronous onsets or captor tones can remove this influence over the pitch percept by segregating the harmonic from the complex tone, even for mistunings of 0%.

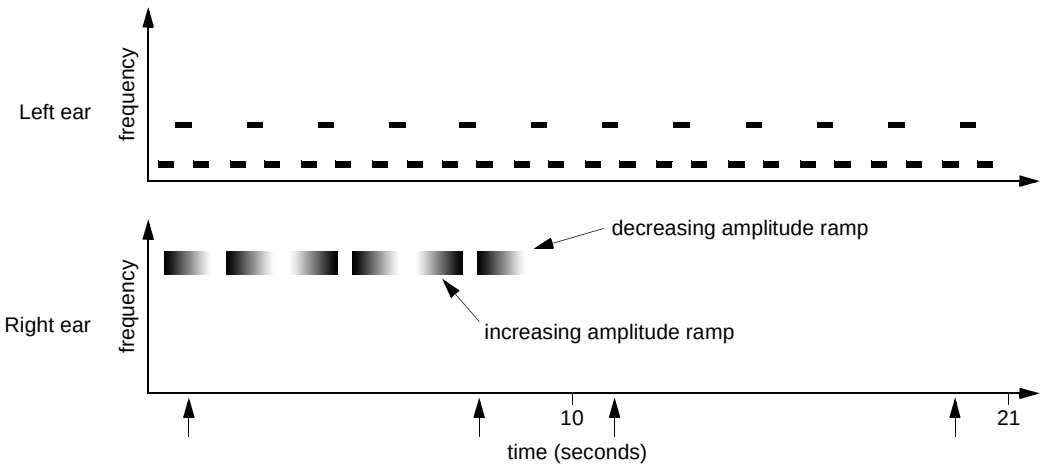


Figure 11 **Symbolic representation of the stimulus format used by Carlyon *et al.* (2001) and presented to the model. Arrows indicate the points at which excerpts of the model output are presented in figure 12.**

Two tone streaming with distractor task

Carlyon *et al.* (2001) used the alternating tone sequence stimulus described above to investigate the role of attention in the stream formation and segregation process. To achieve this, subjects were presented with a binaural stimulus in which the signal to the left ear contained the alternating tone sequence and the right ear contained noise bursts. Specifically, an alternating tone sequence, 21 seconds in duration, was presented to the left ear; bandpassed noise bursts (2 - 3 kHz band), each of duration 400 ms, were presented to the right ear for the first 10 seconds of the stimulus at a rate of one per second. For the remaining 11 seconds, silence was presented to the right ear. Each noise burst was subject to a linear amplitude ramp which was either increasing (a gain of 0 to 1 over the 400 ms) or decreasing (a gain of 1 to 0 over the 400 ms). Figure 11 shows a symbolic representation of the stimulus used. In the 'two task' condition, Carlyon *et al.* (2001) instructed listeners to attend to the noise bursts and classify the type of amplitude ramp that had been applied to each noise burst. After 10 seconds had elapsed, they were then instructed to move their attention to the alternating tone sequence in the other ear and use a computer keyboard to report the percept of the tones: one stream or two. In the 'one task with distractor', the same stimulus was used but listeners were instructed to ignore the noise bursts and concentrate on assessing the alternating tone percept for the full stimulus duration. Carlyon *et al.* (2001) found that the stream segregation

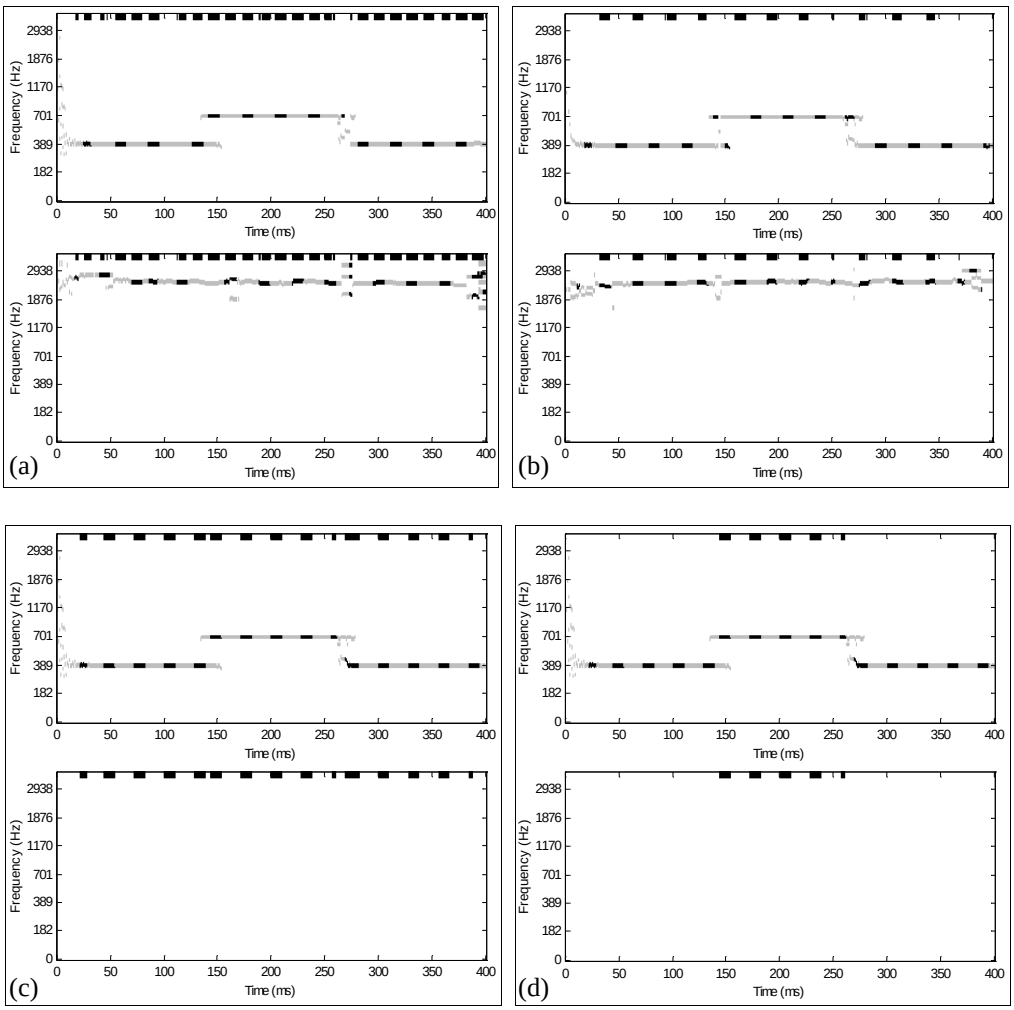


Figure 12 **Model response to the Carlyon *et al.* (2001) data after varying time periods: 1 second (a), 8 seconds (b), 11 seconds (c) and 20 seconds (d). Response of the left ear network is shown in the top panel and the right ear network in the lower panel in each case. Note that the noise bursts in the right ear cease after 10 seconds.**

percept had occurred by the 10 second point in the 'one task with distractor' condition.

However, after the 10 second point in the 'two task' condition, the percept was that of a single stream and the associated build up period was required before two streams were perceived: no build up of the streaming percept had occurred after 10 seconds when listeners were attending to the noise bursts.

The stimulus described by Carlyon *et al.* (2001) was used as input to the model and the output of the oscillator network is shown in figure 12. The movement of attention from the right ear (noise bursts) to the left ear (alternating tone sequence) was simulated by altering A_{ear} after 10 seconds. For the first half of the stimulus $A_{left} = 0$ and $A_{right} = 1$, causing attention to be directed toward the right ear; for the remainder of the stimulus $A_{left} = 0$ and $A_{right} = 1$, so that attention is directed toward the left ear.

Since the Carlyon stimulus is 21 seconds long, it is difficult to present the entire output image and maintain sufficient clarity to inspect the synchronisation of oscillators and the time course of the ALI. In view of this, figure 12 shows four representative excerpts from the network output: the state of the oscillator array after one second, eight seconds, 11 seconds and finally 20 seconds of the stimulus. Figure 12a shows the network after one second of the stimulus. At this stage, attention is directed toward the noise bursts simulating the classification task performed by the listeners in the Carlyon *et al.* experiment. Since only one second has elapsed, the level of attention build up has not reached a sufficiently high level as to exclude activity from the 'ignored' tones. However, after a period of time, this build up has occurred and the ALI is only influenced by the noise segment indicating that only the noise burst is contained in the attentional stream (figure 12b). After 10 seconds, the simulated switch of attention from the right ear to the left ear occurs and the attentional 'reset' occurs. Following this switch, the ALI is once again influenced by all the segments present (figure 12c) until the build up period has elapsed. Once the attentional build up has occurred, streaming is observed since attention is now directed toward the high frequency tone within the alternating tone sequence. This is indicated by ALI activity being synchronous only with the oscillators corresponding to the high frequency tone.

Harmonic segregation within a complex tone

In addition to the monaural experiments discussed above, Darwin *et al.* (1995) also investigated the effect of a contralaterally presented mistuned harmonic upon the pitch of a 12 component complex tone.

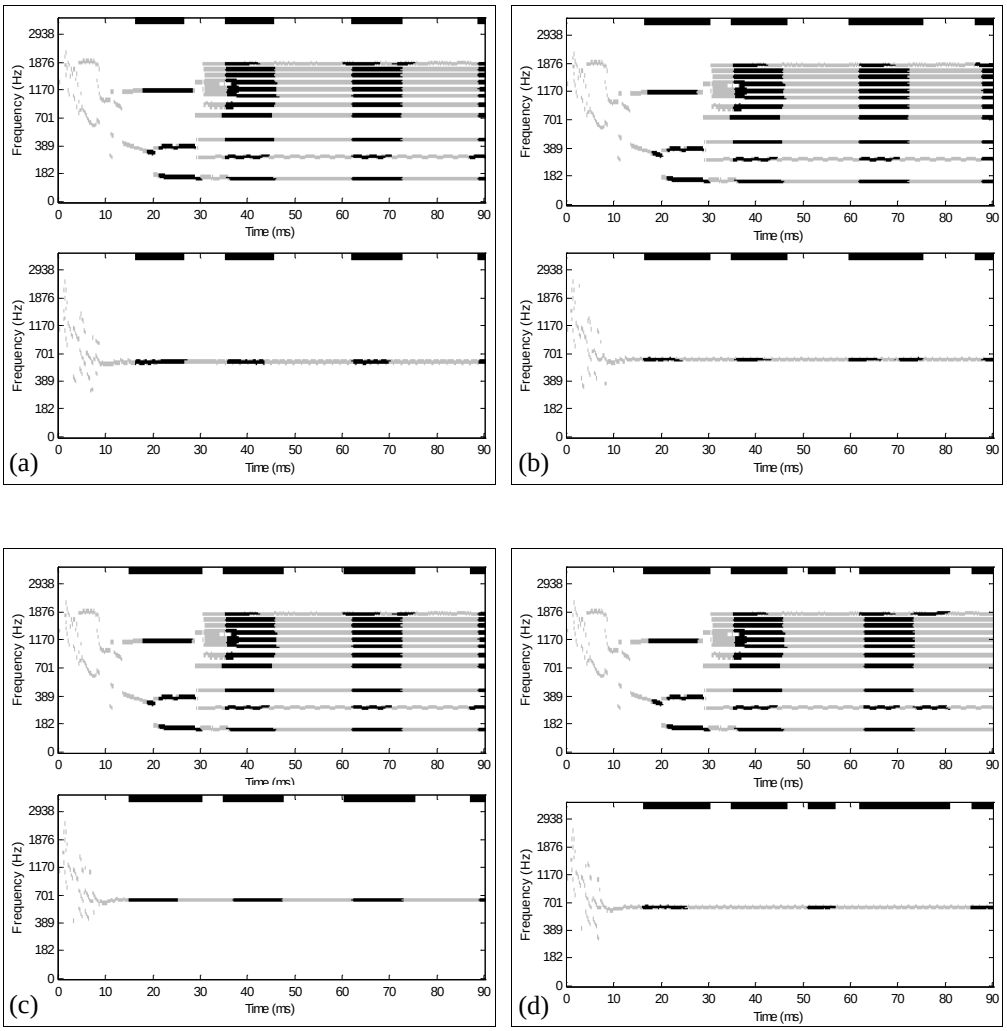


Figure 13 **Model response to a 12-harmonic complex tone (155 Hz fundamental frequency) whose contralaterally presented 4th harmonic is mistuned by varying degrees: 0% (a), 5% (b), 7% (c) and 8% (d). Response of the left ear network is shown in the top panel and the right ear network in the lower panel in each case.**

As in the monaural situation, the degree of mistuning of the fourth harmonic influenced the shift in the perceived pitch of the complex for mistunings of up to

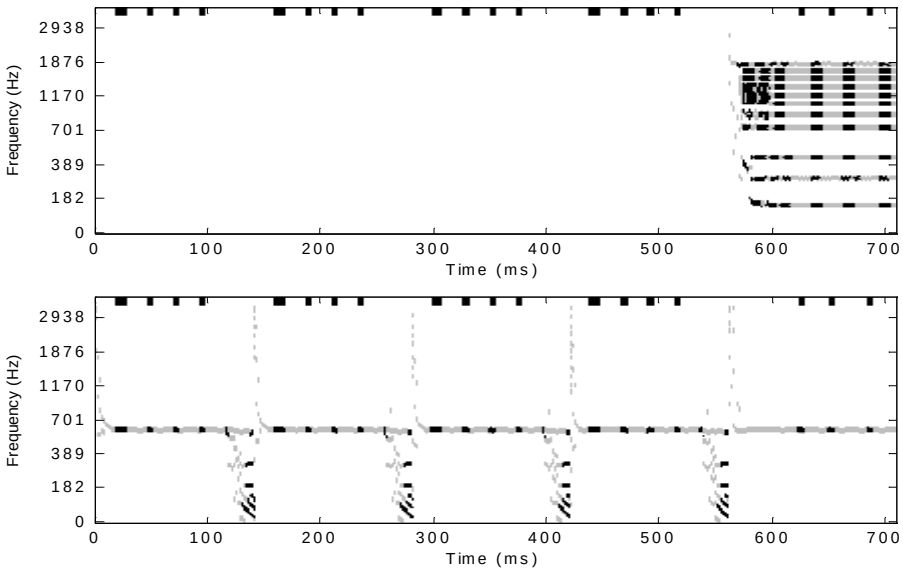


Figure 14 **Model response to a 155 Hz fundamental frequency complex tone whose fourth harmonic is presented contralaterally and is preceded by four captor tones of identical frequency.**

8%; mistunings of more than 8% had little effect of the perceived pitch. These results imply that even harmonics presented contralaterally are assessed to see if they match an overall binaural pitch estimate and are grouped accordingly. Figure 13 shows similar stimuli to those presented in figure 7 with the exception that the mistuned fourth harmonic is presented contralaterally to the remainder of the complex. Similar to figure 7a, figure 13a shows that the oscillator activities corresponding to both the complex and the contralateral fourth harmonic are temporally synchronised indicating that all the segments have been grouped on the basis of common harmonicity. Similar behaviour is observed when the harmonic is mistuned by 5% and 7% (figure 13b and 13c) and the harmonic is said to be perceptually grouped with the complex even at these relatively high degrees of mistuning. When the degree of mistuning reaches 8%, the fourth harmonic is segregated from the rest of the complex: the oscillators corresponding to the fourth harmonic are temporally desynchronised from the remaining oscillators (figure 13d). Once again, ALI activity displayed at the top of the diagrams in figure 13 indicates that both the complex and the segregated harmonic are attended to.

The effect of harmonic capturing upon the pitch percept of the complex can also be demonstrated binaurally. Darwin *et al.* (1995) found that the binaural grouping of

the complex and the contralaterally presented harmonic can be prevented if the contralateral fourth harmonic is preceded by a number of captor tones of the same frequency as the harmonic. In this situation, the mistuning of the harmonic had no effect on the pitch percept implying that it had been segregated from the complex and hence could not affect the pitch. Figure 14 shows the model response to the stimulus. It can be clearly seen that despite the fourth harmonic being subject to 0% mistuning, it is still segregated from the complex. This is indicated by the temporal desynchronisation of the oscillators corresponding to the fourth harmonic and those corresponding to the complex.

7.4 Predictions

A benefit of producing conceptual and computational models is the ability to draw predictions that can subsequently be used as the basis for further psychophysical studies. This section presents a number of predictions made by the model which could be investigated by future studies.

Nature of the attentional reset when the attentional focus is moved in space

To simulate the experimental findings of Carlyon *et al.* (2001), the model incorporates an attentional 'reset' when the attentional focus is moved from one ear to the other. This behaviour is also supported by data presented by Anstis and Saida (1985) in which the streaming build-up effect for an alternating tone sequence does not transfer between ears. They presented listeners with a tone sequence in one ear, and silence in the other ear, for a duration sufficient to induce streaming. When the sequence was then moved to the other ear, the streaming percept was no longer present. Both of these studies presented stimuli over headphones, and subjects were instructed to attend to one ear or the other. As such, stimulus presentation was dichotic, a situation which does not occur in free field listening. In free field, the stimuli could be presented via speakers positioned either side of the listener. In this case, both ears would receive both signals, albeit subject to interaural time differences and interaural intensity differences. This implies that spatial attention must be allocated more subtly than simply selecting a particular ear.

Our model predicts that when attention is moved from any spatial location to another in a discrete manner, an attentional reset occurs. This is a consequence of the attentional 'reset' described in the previous chapter in which the leaky gain factor on the A_k vector, used to model the build up of the attentional effect, is reset to zero when a change in the ear dominance (A_{ear}) is detected (see chapter 6).

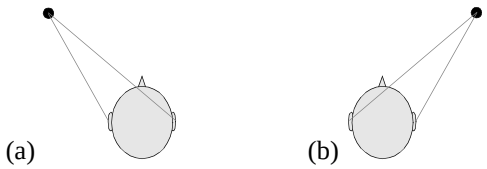


Figure 15 **Representation of the source locations to be used to investigate if a reset occurs when attention is moved. (a) Initial position of the sound source. (b) Final position of the sound source. The source is only moved to this latter position after sufficient time for a streaming percept to have been built up for position (a).**

Conceptually, one would expect such a reset to occur because an abrupt movement of the source is likely to be interpreted as the cessation of the initial source and the appearance of a new source. Furthermore, if this is the case, the principle of good continuation (see chapter 2) implies that there would be an associated threshold in which movements of the perceived source can be too small to trigger an attentional reset and would instead be interpreted as natural movement of the source.

This prediction can be investigated by using an alternating tone sequence presented over headphones in which an interaural intensity difference is applied to lateralise the signal to one side of the head. The sequence would be of sufficient duration for the streaming percept to occur. Following this, the interaural intensity difference would be altered such that the perceived location would be on the other side of the listener (see figure 15). Throughout the stimulus, listeners would be instructed to indicate, via a keyboard, times at which their percept changes from one stream to two streams. The model predicts that listeners will perceive one stream following the simulated source movement.

Streaming percept would not be reset by unconscious reorientation of attention

A reset of the attention gain factor in (36) (see chapter 6) will only occur when endogenous attention is moved to a different ear. Therefore, the model predicts that if an alternating tone sequence was momentarily interrupted by an unexpected loud noise in the other ear, any streaming percept would remain since the noise burst only became the attentional stream *exogenously.* This experiment relies on the assumption that the listener will not consciously move their attention to the ear of the interrupting noise burst even when instructed to only concentrate on the alternating tones. This prediction arises from that fact that there would be no change in the ear dominance (A_{ear}) and hence no reset of the attentional build up would occur. However, designing an experimental paradigm in which one can be

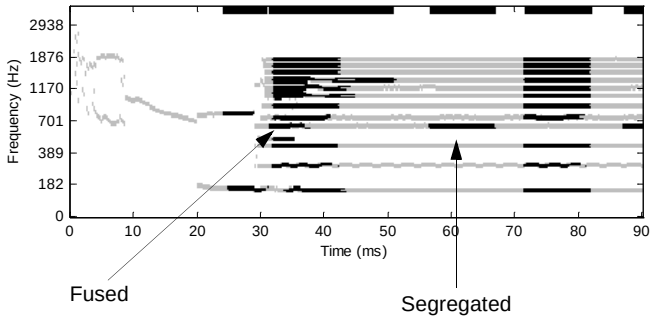


Figure 16 **In the model, a finite period of time is required for a mistuned harmonic to segregate from the complex.**

sure that the intruding sound never causes an endogenous reorientation of attention may prove difficult.

For a mistuned harmonic, segregation occurs over a finite period of time (albeit extremely rapidly) as groups of oscillators desynchronise. Does this happen with listeners?

The model predicts that the initial percept of a complex tone with a harmonic mistuning of more than 8% is fused; however, this harmonic is segregated after approximately 25 ms - the period of one oscillation (see figure 16). This is consistent with the concept that the default grouping state is ‘fusion’ (Bregman, 1990). Only when sufficient evidence has been gathered (i.e. an accurate estimate of the fundamental), does segregation take place. We suggest that this could be investigated using the pitch matching technique employed by Darwin *et al.* (1995). The effect of mistuning on the perceived pitch ought to be greater for short stimuli since the proportion of the stimulus duration in which the mistuned harmonic is grouped with the complex, and hence contributing to the pitch percept, is greater than for longer stimuli. In other words, the pitch shift is inversely proportional to the stimulus duration. Specifically, listeners would be presented with bursts of the complex of varying length, but fixed mistuning, and then instructed to determine the pitch.

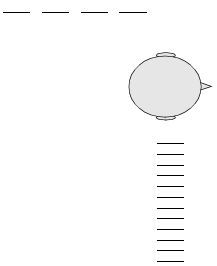


Figure 17 **Presentation of captor tones to the opposite ear to which the complex will be presented. The model predicts that the captor tones would have no segregating effect on the complex.**

Do captor tones presented to the opposite ear to complex have the same capturing effect?

The examples presented above in which captor tones are used to segregate a harmonic from a complex presented the captor tones in the same ear as the harmonic. The model predicts that captor tones presented to the ear contralateral to that to which the harmonic is to be presented (see figure 17) would have no capturing effect.

7.5 Psychophysical investigation of the spatial allocation of attention

In the previous section, we predicted that an attentional reset would occur if an attended stimulus was moved discretely in space. In this section we describe a psychophysical experiment in which an alternating tone sequence is used to test this prediction. In a similar manner to previous alternating tone sequence studies (e.g. Anstis and Saida, 1985; Carlyon *et al.*, 2001), the experiment investigates the time course of auditory stream segregation. It ought to be noted that this type of experiment differs from studies which investigate stream segregation based upon spatial location (e.g. Deutsch, 1975; van Noorden, 1975). Instead of presenting successive tones to different ears, each alternate tone in our experiment is presented at the same azimuth. The switch in position occurs once: it is the effect that this has on the time course of the streaming percept that is of interest and not the ability (or inability) to stream based purely on spatial location.

Anstis and Saida (1985) presented listeners with a tone sequence in one ear and silence in the other ear, for a duration sufficient for streaming to occur. The sequence was then moved to the other ear and the subsequent changes in the listener's percept were recorded. Throughout the experiment, listeners were instructed to use a computer keyboard to record their percept of the tone sequence: coherence or segregation. The experimental design described below is similar to that of Anstis and Saida (1985, experiment four, p. 264) but instead of presenting the stimuli dichotically, the tone sequence in our experiment was presented binaurally using interaural intensity differences to simulate movement of the stimulus from one spatial location to another. To achieve this, the tone sequence was presented to one side of the head for the first half of the stimulus and was then moved to the same spatial offset on the other side of the head by swapping the left and right ear signals.

The stimulus consisted of an alternating ABA tone sequence similar to that used by van Noorden (1975) and Carlyon *et al.* (2001) with a duration of 20 seconds. The low frequency tone was fixed at 1 kHz and the high frequency tone was set for each subject such that streaming would occur after approximately five seconds when the sequence was presented diotically (i.e. with the perceived location in the midplane).

For each presentation, an alternating tone sequence was presented to the listener in which an intensity difference between the two ears was applied. After 10 seconds, the intensity difference was reversed such that the stimulus appeared to move to the other side of the head. Each intensity difference was presented eight times in a randomised order that was different for each subject. The side of the head on which the stimulus started was also randomised. For example, on a presentation in which the intensity difference was 10 dB, the intensity difference would be +10 dB for the first ten seconds and -10 dB for the remaining ten seconds (i.e. an overall spatial movement equivalent to a 20 dB intensity difference).

A pretest was conducted, using three listeners (SC, ST and KP), in which the intensity difference between the two ears on each trial was selected from the range 0, 5, 10, 15, 20 or 25 dB allowing for a range of perceived spatial locations. The data from the pretest suggested that the reset effect becomes apparent for intensity differences between 0 dB and 10 dB: no significant change in the characteristics of the reset was observed above 10 dB. Hence, the experiment described below used intensity differences of 0, 1, 2, 4, 8 and 16 dB in order to better observe the emergence of the reset effect.

Subjects continually made forced choice judgments of stream segregation by holding down one of two computer keys over a listening period of 20 seconds. They

held down key (1) when the stimulus sounded like a single stream of tones with a galloping rhythm - *coherence* - and held down (2) when they heard two streams of tones in which one was faster than the other - *segregation*. Each subject was presented with 48 different examples (6 intensity differences x 8 presentations).

Experimental method

Subjects

5 subjects were used, namely 4 males (GB, SC, SM, ST) and 1 female (JE), all of whom had some experience with psychophysical experiments. None of the subjects reported hearing difficulties.

Stimulus

The alternating tone sequences were generated using a Tucker-Davis Technologies (TDT) System 3 RP2.1 device incorporating a TDT HB7 headphone driver and were presented to a pair of Sennheiser HD250 linear II headphones. Subjects sat in a single walled sound-attenuating booth (IAC 402-A Audiometric Booth). The amplitude of the stimuli was set to a comfortable listening level (no direct SPL measurements were taken). All stimuli were presented binaurally. The tone duration was set to 120 ms and the tone repetition time (TRT), as defined by van Noorden (1975), was also set to 120 ms. The frequency separation of the alternating tones was set for each subject during a pre-test in which the subject was presented with an alternating tone sequence of default frequency separation (75 Hz) with the low frequency tone set at 1 kHz. If stream segregation did not occur after approximately five seconds, the subject could increase the separation by 10 Hz and re-listen to the sequence. This procedure was repeated until stream segregation occurred.

Procedure

Subjects were instructed, via text presented on a computer screen, to continuously hold down one of two computer keys, changing keys to report their percepts, over a listening period of 20 seconds. They held down key (1) when the stimulus sounded like a single stream of tones with a galloping rhythm - *coherence* - and held down (2) when they heard two streams of tones in which one was faster than the other - *segregation*. For each subject, the experiment was broken up into sections consisting of five presentations. Subjects were permitted a short break between sections if desired.

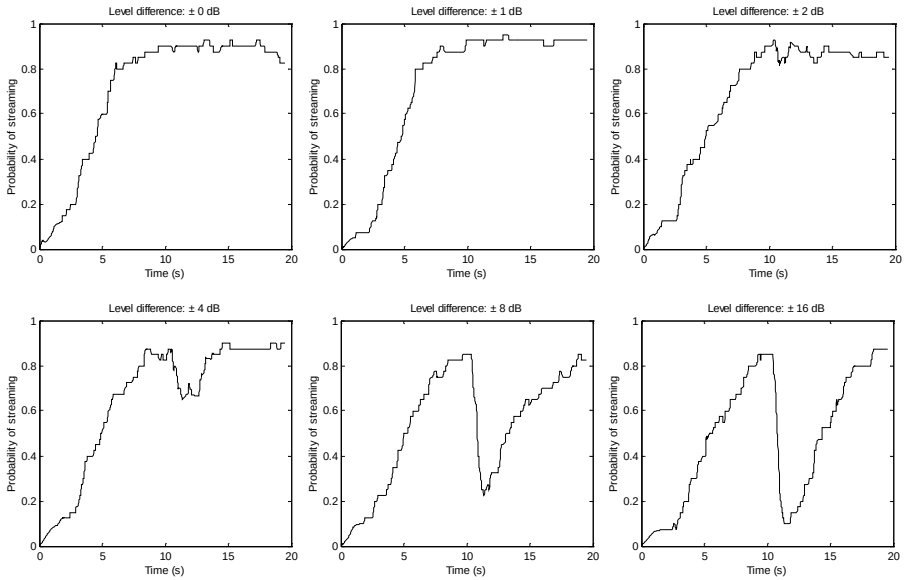


Figure 18 **Mean probability of auditory streaming for intensity differences of 0, 1, 2, 4, 8 and 16 dB. The intensity difference between the two ears was reversed after 10 seconds.**

Results

For each intensity difference, the 40 response sequences (5 subjects x 8 presentations), consisting of discrete judgements (1 or 2), were averaged and scaled to between 0 and 1. This provided the mean time course of adaptation to the alternating tone sequence in which each point is interpreted as a probability of streaming. These results are shown in figure 18. Each figure shows that the probability of hearing two streams at the beginning of the presentation was low but that this increased over time as previously demonstrated by Anstis and Saida (1985). For presentations in which the intensity difference was 4 dB or more, there was an associated decrease in the probability of hearing two streams when the intensity difference was reversed after 10 seconds; a second period of build up was then observed. This decrease in streaming probability was strongest (observed as a deeper and wider trough) for presentations in which a large intensity difference was applied. Individual subject responses can be found in appendix B.

Discussion

Our results demonstrate the time course of adaptation to an alternating tone sequence and are in agreement with the experimental data presented by Anstis and Saida (1985). Since attention appears to be required for auditory streaming (e.g. Carlyon *et al.*, 2001), our data supports the hypothesis that an attentional reset occurs when an attended stimulus is moved discretely in space. Furthermore, as we suggested in the previous section, there appears to be a tolerance associated with such an attentional reset in which small changes in source location have no effect; intermediate differences result in a small degree of reset and large differences cause a large reset effect. The data presented here suggests that this tolerance is of the order of 2 - 4 dB (i.e. an overall interaural intensity difference equivalent to 4 - 8 dB).

As described in chapter 6, our model's attentional mechanism resets when an ear dominance (A_{ear}) change is detected. This is achieved by resetting (to zero) the leaky gain factor L on the A_k vector described in (36). The results presented in this section suggest that this mechanism needs to be tuned such that only changes larger than a certain threshold (a movement equivalent to a 4 - 8 dB intensity difference) cause the attentional mechanism to be reset and that the leaky integrator L should not be reset to zero in all circumstances: the degree of reset should depend on the the magnitude of the A_{ear} change.

7.6 Summary

This chapter has described a number of simulations which have been used to demonstrate the computational model's ability to replicate the psychophysical findings. These examples have fallen into two categories: monaural and binaural stimuli. In each of these, the model's ability to perform auditory grouping and attentionally modulated stream formation and segregation has been investigated. The effect of a mistuned harmonic on the pitch percept of a complex tone has been used to investigate perceptual grouping (Darwin *et al.*, 1995). Such stimuli have been used to show how the model performs harmonicity grouping and incorporates a degree of tolerance in the harmonic relationships between components. Furthermore, the old-plus-new heuristic, which has been incorporated at a relatively low level, accounts for the perceptual segregation of components which have been subjected to asynchronous onsets or the presence of captor tones.

The requirement of attention in the stream formation process has been demonstrated both monaurally and binaurally. Figure 4 shows how differences in frequency separation can affect the ability of listeners (and the model) to form two streams from an alternating tone sequence. On the basis of psychophysical evidence (Carlyon *et al.*, 2001), we have argued that such streaming is due to the allocation of attentional resources to the alternating tones sequence. Such endogenous, conscious, attentional allocation is believed to occur in the 'shape' of a gaussian distribution which can be assigned to a particular frequency (e.g. Mondor and Bregman, 1994). Hence, the model simulates the allocation of attention in frequency by weighting the tonotopical connections between the oscillator array and the attentional leaky integrator (ALI). It is the width of this attentional allocation which accounts for the ability of the model to 'stream' widely frequency-separated tones. If a set of tones within the alternating tone sequence fall beyond the tail of the attentional gaussian, it will be unable to influence the ALI once the associated build up (Anstis and Saida, 1985) has occurred. The model also accounts for the original binaural stimulus used by Carlyon *et al*. (2001) to investigate attentional influences on auditory streaming. In this situation, the model simulates the lack of streaming build up observed in listeners when their attention is diverted away from an alternating tones sequence.

The final section of this chapter presented a selection of predictions and testable assumptions made by the computational model. A psychophysical study was presented in which one of the predictions - the hypothesis that discrete movement of the attentional focus over large spatial distances causes an associated reset of the attentional build up - was investigated. The data from this experiment supported the hypothesis and also suggested that such a reset is dependent on the degree of source movement.

In the conclusions chapter, we will discuss the outcome of this evaluation and also suggest areas of future work which will allow a clearer view of binaural attentional allocation to be formed.

Chapter 8. Conclusions

8.1 Summary

A model of auditory streaming has been presented in which the allocation of attention lies at the heart of the stream formation process. Chapter 5 presented a conceptual framework in which a number of different processes are responsible for producing the final stream estimates. Firstly, we made the distinction between *exogenous* and *endogenous* attention. Exogenous, unconscious, attention is responsible for performing primitive grouping of individual features within the

stimulus (see chapter 2). These groups are then passed to the endogenous processing stage, which takes into account the conscious decision of the listener, changes in the stimulus input and schema information to form an 'attentional stream'. It is at this stage of processing that attentional allocation influences stream formation. It is proposed that schema information is used to both aid the grouping of the exogenous processing outputs (e.g. Treisman, 1960), and to act as a form of detector in which the detection of salient information in any group reorients conscious attention (e.g. Moray, 1959). Furthermore, the conscious decision of the listener can be overruled by, and attention forcibly reoriented to, loud sounds occurring unexpectedly in the environment. This mimics the startle reflex observed in most animals (e.g. Winslow *et al.*, 2002).

The implementation of this framework as a computer model consists of three core stages. The first stage of the model simulates peripheral auditory processing using a bank of cochlear filters representing basilar membrane processing, with the gains of each filter set to reflect the outer-middle ear transfer function. The output of these filters are then passed through an approximation of inner hair cell transduction to produce an estimate of auditory nerve activity. The second stage of the model extracts periodicity information by means of a correlogram (e.g. Slaney and Lyon, 1990) in which each auditory nerve channel is subjected to an autocorrelation analysis. This produces a two-dimensional representation with channel centre frequency and autocorrelation lag on orthogonal axes. The dominant pitch for the entire signal is made explicit by forming a summary correlogram by integrating across frequency. The correlation between each adjacent channel pair is computed in order to identify channels with similar patterns of response. Such channels are responding to the same spectral event and therefore are signalled as belonging together - they are grouped in the next processing stage.

The final stage of the model is a one-dimensional neural oscillator network in which auditory grouping and segregation takes place. This neural oscillator network is based on LEGION (see Terman and Wang, 1995) but differs in its unidimensionality and the incorporation of long range inter-oscillator connections. The former avoids an explicit time axis, the problems of which were discussed in chapter 6, and the latter allows grouping by harmonicity (or other across-channel cues). Channels responding to the same spectral event (as identified from the cross-correlation mechanism in the previous stage) are encoded in the network by locally excitatory connections to form 'segments'. Further sets of excitatory connections are placed between individual segments whose periodicity conforms with the fundamental frequency estimate of the summary correlogram. The activities of oscillators with excitatory connections between them synchronise temporally to

form an oscillatory ‘group’. Blocks of oscillators which are not linked by excitatory connections desynchronise from each other.

Each oscillator in the network feeds activity to the attentional leaky integrator (ALI). The ALI is the core of our attentionally motivated stream segregation mechanism and produces an *attentional stream* as defined in chapter 5. The connection weights between the network and the ALI are modulated by endogenous processes including ‘conscious’ preference. This conscious preference is modelled by a gaussian distribution with the peak centred on the frequency of choice (see Mondor and Bregman, 1994). Initially, these weights are maximal for all channels to simulate the default grouping being a whole (see Bregman, 1990). Hence, in this initial condition, all segments and groups contribute to the attentional stream. Over a period of time, these weights adapt to the endogenous attentional focus with the effect that only oscillators under the attentional focus can influence the ALI. The reliance on synchrony allows harmonic groups, most of whose harmonics may be outside of the attentional focus, to contribute to the attentional stream simply by attending to one harmonic.

The output of the model was evaluated by inspecting the time course of the neural oscillator array activity and the ALI for a number of stimulus types. From this information, one can assess how individual spectral events have been grouped and which groups of segments contribute to the attentional stream. In other words, the activity of the ALI indicates which portions of the stimulus a listener would be attending to. To evaluate the success of the model, the network output was compared to the findings of the relevant psychophysical studies.

Accounts of previous versions of the model can be found in Wrigley and Brown (2001) and Wrigley and Brown (2002).

8.2 Original contribution

The purpose of this thesis is to present a model of auditory streaming. However, in contrast to previous CASA systems, attention plays a crucial role in the stream formation and segregation process of our model. We argue that, on the basis of psychophysical research (e.g. Carlyon *et al.*, 2001), distinct streams are not formed unless attention is directed toward particular (groups of) features. To this end, we have developed the attentional leaky integrator (ALI) and a representation of attentional allocation across frequency, A_k. A_k corresponds to the conscious preference of the listener and is modelled by a gaussian distribution in accordance

with the psychophysical findings of Mondor and Bregman (1994). Furthermore, we argue that this conscious preference cannot be deployed instantaneously: it is subject to a build-up over time. Hence, the degree to which grouped frequency channels can influence the ALI is determined by the timecourse of the A_k build-up. The output of the ALI describes the frequency content of the 'attentional stream' at each epoch. In chapter 5 (see also chapter 7) we demonstrated how this mechanism can explain both auditory streaming (van Noorden, 1975) and the associated build-up of streaming over time (Anstis and Saida, 1985).

In addition to this, it has been shown that when the ear of presentation of an alternating tone sequence is switched, a second build-up period is required immediately after the switch (e.g. Antis and Saida, 1985; Carlyon *et al.*, 2001). We contend that the buildup of attentional efficacy is subject to a 'reset' when abrupt changes in the stimulus location are consciously detected and tracked. In other words, the degree of attentional build-up is reset to its initial value following an abrupt change of spatial attentional preference. It is to be noted that such a reset is not observed when making abrupt movements in frequency.

In chapter 1, the development of an intelligent hearing aid was described as one of the motivations for the model. Specifically, such a hearing aid would incorporate attentional effects such as the unconscious overruling of conscious selection (such as the startle reflex; see chapter 5) which would allow the wearer to become aware of new, loud and potentially important sounds. This mechanism has been included in our model such that sounds presented out of the attentional focus are still able to influence the ALI and hence become attended to, provided they are sufficiently loud.

The old-plus-new heuristic has been incorporated into the model and influences the ability of the network to form harmonically related groups. If a segment is deemed to be 'older' than other harmonically related segments, it is prevented from becoming a member of that group. Previous ASA work has treated the old-plus-new heuristic as being a 'high level' grouping principle similar to common fate or harmonicity (see Bregman, 1990). However, we adopt an approach in which the old-plus-new heuristic is embodied as a low level mechanism. Each channel in our model has a leaky integrator associated with it which acts as an 'age tracker'. The leaky integrator only receives input when its associated channel is contributing to a segment. Thus, the relative ages of concurrent segments can be assessed allowing segments which are either too old or too young to be excluded from a particular group.

In contrast to previous neural oscillator models of ASA (e.g. Wang, 1996; Wang and Brown, 1999), the network presented here is one dimensional which removes the necessity for an explicit time axis. In chapter 6, the problems associated with two dimensional networks were discussed. Specifically, there is no physiological evidence for large arrays of delay lines necessary to produce the time axis, and the ability of such delay lines to maintain temporal accuracy over large distances is uncertain. Furthermore, attempts to incorporate attention into such a network (e.g. Wang, 1996) introduce unrealistic properties (see chapter 6). Our single dimensional network processes the stimulus on a sample by sample basis and, in turn, produces an estimate of the segments present and which of these are contributing to the attentional stream at any epoch. Hence, the timecourse of auditory organisation and attentional influence are implicit in the network output.

Many previous ASA systems exploit the redundancy found in the auditory nerve response. For example, Brown (1992) identifies channels which are responding to the same spectral dominance by performing a cross-correlation of their autocorrelations. However, in order to identify noise segments, we also use channel instantaneous frequencies (see Cooke, 1991/1993) of the gammatone filter. By assessing the variance of the instantaneous frequency response over time, it is possible to determine whether the channel is responding to a periodic or nonperiodic stimulus.

Section 7.5 presented a psychophysical study in which the time course of auditory streaming was analysed. Although similar to that of Anstis and Saida (1985), the experiment presented binaural stimuli as opposed to dichotic stimuli allowing particular spatial locations to be simulated using interaural intensity differences. This has extended the findings of Anstis and Saida (1985) by suggesting that an attentional reset does occur, and that the degree of this reset is related to the magnitude of the source movement in space.

8.3 Limitations of model

The model presented in this thesis is subject to a number of possible limitations which are described in this section. A number of these are also discussed in the future work section (section 8.4).

The peripheral processing stage of the model uses a bank of gammatone filters which are linear. As discussed in chapter 6, auditory filters are known to exhibit non-linearities such as *cochlear echoes*. However, the linear gammatone filter is a

good first-order approximation since its magnitude characteristic exhibits a very good fit to the *roex(p)* function commonly used to represent the magnitude characteristic of the human auditory filter shapes (see Patterson and Moore, 1986).

Only the primitive grouping principles of harmonicity and the old-plus-new heuristic are incorporated into the model. The model does not take into account schema driven grouping. It would be possible to incorporate additional primitive grouping principles by adding more analysis stages, the output of which would influence the excitatory connections which are placed between segments to form groups. It is also possible to incorporate schema influences on grouping using the same temporal correlation framework implementation as employed in our model. For example, the neural oscillator based vowel recognition model of Liu *et al.* (1994) can be interpreted as a form of schema detector in which the identification of a vowel is signalled by a particular pattern of synchronised oscillations. Such synchronised oscillations could be used to reinforce primitive grouping decisions and ensure a pattern of grouping consistent with the recognised sound. Similarly, Wang and colleagues have used neural oscillators to store particular patterns in associative memory (e.g. Wang *et al.*, 1990; Wang and Liu, 2002). This memory network is then able to influence the grouping and segmentation (including pattern completion) of novel patterns represented in a LEGION network.

The neural oscillator network enforces the principle of exclusive allocation (Bregman, 1990). This property was observed in chapter 7 when simulating the effect of a mistuned harmonic on the pitch percept of a complex tone. Experimental evidence suggests that the mistuned harmonic can contribute to more than one group. Moore *et al.* (1986) found that for mistunings greater than 3%, the harmonic tends to be heard as a pure tone standing out from the complex tone. However, that harmonic still makes a contribution to the pitch of the complex (Moore *et al.*, 1985).

The model processes the stimulus on a sample by sample basis with little regard for previous activity. As a result, the segment estimates at each epoch do not explicitly take into account previous segment estimates. The effect of this can be observed in the model output in which some segments can be seen to step up and down in frequency over a period of time.

In chapter 7, we demonstrated how the model can simulate the streaming of an alternating tone sequence and its associated build-up timecourse. However, the model does not incorporate a mechanism for the involuntary movement of attention between the two streams as observed in some listeners. For example, Beauvois and Meddis (1996) suggest that this phenomenon is due to the probabilistic nature of

spike generation within the auditory system and as such they introduce a random variation into the system. Beauvois and Meddis (1996) draw on this to produce a 'random walk' which determines which stream is in the foreground. It is difficult to incorporate a similar mechanism into our model since we contend that streaming is due to the allocation of endogenous attention which is, by definition, not stimulus driven. We contend that top-down, albeit possibly unconscious, influences are responsible for such spontaneous switches - an explanation acknowledged by Beauvois and Meddis (1996) but not incorporated into their model. It is conceivable that analysis is conducted on the attentionally selected stream and an alternative stream becomes the foreground if novel characteristics are not encountered over a certain time period.

Chapter 6 presented both monaural and binaural versions of the model. The binaural model is made of two monaural systems which share the same global inhibitor and attentional leaky integrator (ALI). This architecture is adequate for simulating the perception of dichotic stimuli commonly used in psychophysical experiments, especially those conducted over headphones (e.g. Darwin *et al.*, 1995; Carlyon *et al.*, 2001). However, since the model does not incorporate a full model of binaural processing, it cannot simulate the responses to 'true' binaural stimuli in which the left ear and right ear signals are similar except for timing and level differences. The model as it is presented here cannot represent a single acoustic event which is present in both ears simultaneously as a single perceptual object.

8.4 Future work

Exclusive allocation

As we have seen previously, our model enforces the principle of 'exclusive allocation' which prevents energy at a particular frequency from being allocated to more than one percept. However, there are examples in which this principle does not apply, ranging from the perception of mistuned harmonics (e.g. Moore *et al.*, 1985, 1986) to the perception of speech sounds (Rand, 1974; Liberman, 1982; Mann and Liberman, 1983). Therefore, the model ought to be able to allow the joint allocation of channels when necessary.

Within the neural oscillator framework, Brown and Wang (1996) have proposed a mechanism by which oscillators within a network can be associated with more than one group. Their approach is to require that an oscillator be capable of jumping back up into the active phase very quickly after having jumped down. To achieve

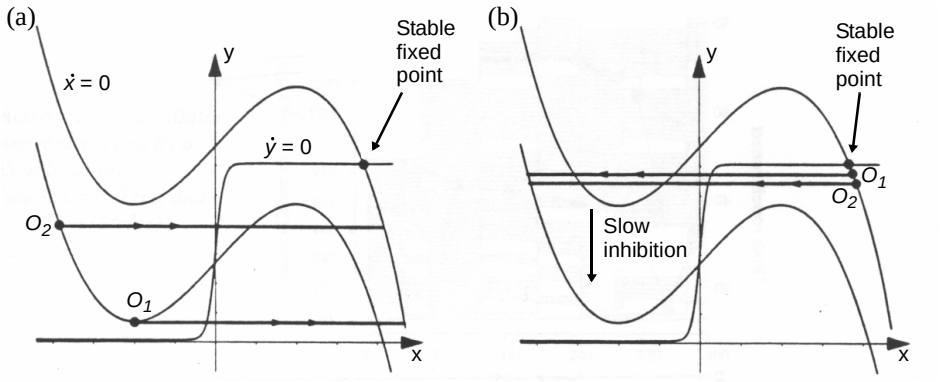


Figure 1 **Oscillator synchronisation with large phase variations. (a) Sycnchronisation is jumping is achieved by mutual excitation between O1 and O2. The raised cubic intersects with the sigmoid, creating a stable fixed point. (b) Synchronisation in jumping down is achieved by slow inhibition. As O1 and O2 near the stable fixed point, they jump down to the silent phase when the slow inhibitor is activated. From Brown and Wang (1996) figure 7.**

this, Brown and Wang raise the oscillator's cubic by a large amount such that it intersects with the *y*-nullcline creating a stable fixed point (see figure 1). To explain how this works, consider two oscillators O_1 and O_2 with mutual excitation. Assume that O_2 has just jumped down from the active phase and that O_1 is just about to jump up to the active phase. Since O_2 is to also be associated with O_1, O_2 must be able to jump back up more quickly than normal. O_1 enters the active phase at a relatively low position; due to the mutual excitation, O_2 receives a large input which raises its cubic to such a high position that it is able to jump up to the active phase. Note that O_2 has entered the active phase at a relatively high position. The mutual excitation ensures that O_1 now follows the trajectory of the higher cubic. As O_1 and O_2 move toward the stable fixed point, their phase difference decreases and they become synchronised.

To prevent O_1 and O_2 from remaining at the stable fixed point, a second global inhibitor, with a slower timescale, is introduced which lowers the cubics of O_1 and O_2 to their original levels thus making them jump down to the silent phase simultaneously. The slow inhibitor timescale is set such that the slow inhibitor is activated when O_1 and O_2 approach the fixed point. Brown and Wang (1996) successfully demonstrated how this technique can be used to model joint allocation of frequency channels when modelling the segregation of double vowels (e.g.

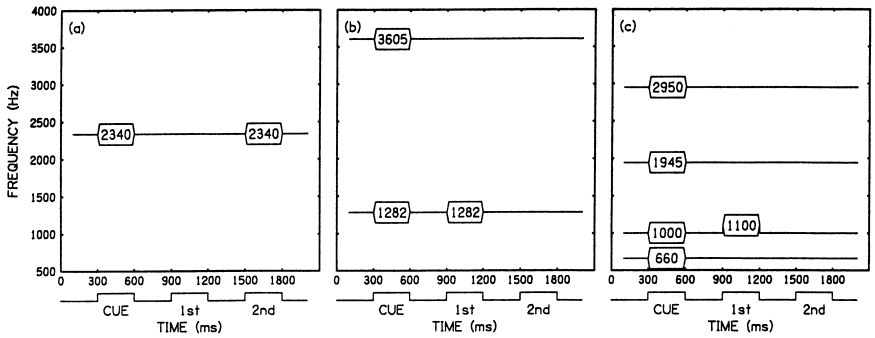


Figure 2 **Schematic of three types of trials. (a) *M*=1 and an expected target. (b) *M*=2 and an expected target. (c) *M*=4 and a probe signal. Probe signals are always 1.1 times one of cue frequencies. From Schlauch and Hafter (1991), figure 1**

Scheffers, 1983; Assmann and Summerfield, 1990). Hence, such an approach could be employed in the model presented here to allow it to account for joint allocation.

Divided attention

The model we have presented demonstrates how selective attention influences the stream formation and segregation process. However, it makes no attempt to investigate the influence of *divided* attention on auditory perception.

In addition to studies concentrating on the role of single site attentional allocation (see chapter 4), work has also been conducted in which there is uncertainty about the frequency of the signal to be detected. Results from these types of experiments are often compared to *ideal* listeners defined by assumptions regarding the nature of the detecting mechanism (for example, parameters of a bank of bandpass filters). A popular model to arise from this is based on the listener who monitors *M* orthogonal bands (MOB), only one of which contains the target signal (Green and Swets, 1966). Solutions from these models agree qualitatively with the observation from human listeners that the signal level of the target must be increased for detection as the number of monitored bands increases. However, Green (1960, 1961) had found that the predicted loss in sensitivity as *M* increased was in fact larger than that seen in human listeners. He accounted for this by proposing that there was an intrinsically high amount of uncertainty present in all conditions. According to Schlauch and Hafter (1991) Green's experiments were flawed: they did not control for the possible cognitive influences on listening bands. A listening band is defined as the band on whose output a listener decides whether or not a signal has occurred. Green employed a traditional probe-signal method in which the

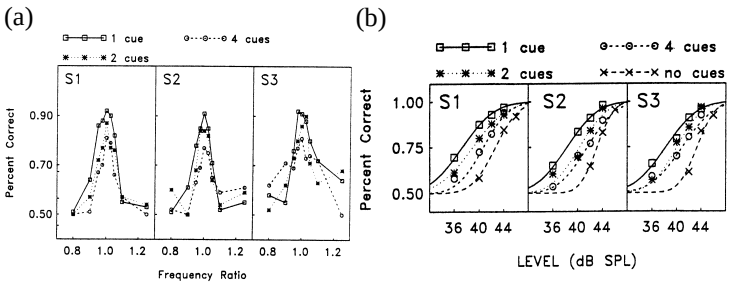


Figure 3 **(a) Data from three subjects using one-, two- and four-tone complexes as a cue. Abscissae represent ratio of signal to be detected to target frequency. (b) Psychometric functions for the three subjects. Abscissae represent dB SPL level of a 500Hz tone. From Schlauch and Hafter (1991).**

cue signal was not presented on a trial-by-trial basis, forcing subjects to rely on memory to monitor the appropriate frequency bands. The use of only two frequencies to monitor also limited the fall in performance due to uncertainty.

Schlauch and Hafter (1991) controlled for these factors by employing trial-by-trial cuing of subjects (Greenberg and Larkin, 1968) and selecting the frequencies to be monitored at random from a wide range of possibilities to avoid memory effects. Finally, to increase the possible amount of performance loss due to uncertainty, the number of bands to be monitored was doubled to four. Figure 2 shows a schematic of their signals. As observed in single site expectancy experiments, detection performance dropped for frequencies above and below the expected frequencies (figure 3a). In addition to this data, Schlauch and Hafter also calculated the *psychometric function* for each cued condition. It has been classically considered that a threshold is that intensity above which a stimulus can be heard and below which it can not. This is an oversimplification: if the intensity of a stimulus is slowly increased from a low value, there is no well-defined point at which subjects suddenly report the stimulus to be detectable. Instead, there is a range of intensities over which the subject will sometimes declare the stimulus detectable and at other times declare it undetectable. When the responses to a number of trials are plotted with percent 'detectable' responses on the ordinate and signal magnitude on the abscissa, a distinctive sigmoidal shape is produced. This plot is termed a psychometric function. This allowed the performance in dB of probes relative to expected targets to be inferred. Additionally, this allowed the hypothesis that the slope of the psychometric function increases with *M* to be tested (Green and Swets, 1966).

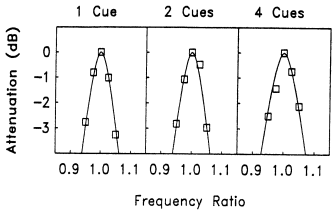


Figure 4 **Average listening bands for one, two and four cues. Best fit lines are provided by a ROEX(p) function. From Schlauch and Hafter (1991) figure 4.**

Figure 3b shows that the slope of the psychometric functions does indeed increase as *M* increases which is consistent with evidence from Johnson and Hafter (1980) reporting increased slopes as uncertainty increased.

The performance in dB of probes relative to expected targets also produced interesting results. The functions shown in figure 4 were provided by adjusting the value of *p* in the ROEX(*p*) filter (Patterson and Moore, 1986). Patterson and Moore, (1986) showed that the equivalent rectangular bandwidth (ERB) of a filter is (4/*p*)*F* where *F* is the centre frequency. From this, it was found that the width of the filters were 12%, 12.4% and 13.7% of the centre frequency for cases *M*=1, *M*=2 and *M*=4 respectively which are essentially the same as those obtained using notch-noise masking (Moore and Glasberg, 1983).

In summary, Schlauch and Hafter (1991) show that there is a significant fall in detection performance due to increasing *M* but that it was indeed possible to monitor a number of harmonically unrelated frequency bands simultaneously. The performance decrease as probes move out of these bands is consistent with the hypothesis that attention is allocated to a number of discrete frequency regions but that its effectiveness, or the ‘amount’ of attention available for each frequency region, decreases as *M* increases. It would be interesting to investigate how the division of attentional allocation influences the streaming process. Only then can a model which encompasses both selective and divided attention be produced.

Binaural enhancements

As discussed in the previous section, the model cannot process binaural stimuli: it can only simulate the perception of dichotic signals. The inclusion of a binaural peripheral front end which would take into account interaural time and level differences would provide useful information on the spatial location of particular segments. Although the cue of spatial location is not believed to play a role in the

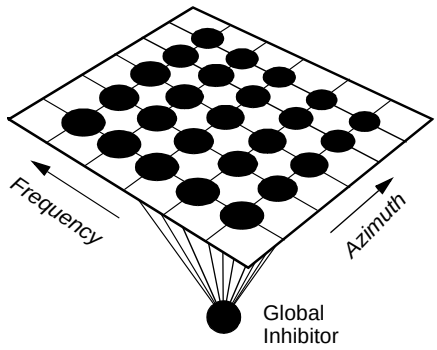


Figure 5 **Two dimensional binaural neural oscillator network. Each oscillator in the grid represents a particular frequency-azimuth combination.**

grouping process (Darwin and Hukin, 1999), it is vital in forming a representation of the stimulus in which a single acoustic event, detected by both ears, is encoded as a single event. The current model is unable to identify segments which appear in both ears but represent the same acoustic event.

With the inclusion of a binaural front end, the ability to allocate attention to a particular spatial region would also be possible. A possible way in which the neural oscillator network could be developed to incorporate this information is shown in figure 5. The proposed network would be two-dimensional with frequency on one axis and azimuth on the other. Note that this does not incur the unrealistic properties associated with two-dimensional networks discussed in previous chapters since the second dimension here is azimuth and not time: the network will still process the stimulus on a sample-by-sample basis. Despite the requirement of time delays to generate azimuth estimates, these would be much shorter than those required to model an explicit time axis. Furthermore, there is evidence to support the existence of such short delay lines for this task (e.g. Carr and Konishi, 1988).

The time course of attentional allocation in space remains unclear and is a topic which ought to receive investigation. The Carlyon *et al.* (2001) study demonstrated that attention is required for auditory streaming to occur. In their study, attention was allocated spatially (albeit to a dichotic stimulus). Our approach to explain how a buildup of attention is necessary, but does not transfer when attention is moved from the distracting task to the alternating tones, is to employ a reset mechanism. When the model detects a movement of spatial ‘interest’, the build up of attention is reset. This assumption is also supported by the experimental findings of Anstis and Saida (1985) in which the build up of the streaming percept did not transfer

between ears even when attention was always directed to the alternating tone sequence. Once again, Anstis and Saida (1985) presented their stimuli over headphones dichotically. Psychophysical data presented in the previous chapter suggests that such a reset also occurs under binaural conditions. Furthermore, there appears to be a degree of tolerance in which small changes in the perceived location of the source have little or no effect on the time course of streaming. However, for larger movements, a reset is observed and the magnitude of this reset may be proportional to shift distance. It would be interesting to conduct a further experiment to investigate how the streaming percept is affected by continuous, smooth, changes in the sources apparent location. It is issues such as these that must be investigated further before an accurate model of binaural attentional allocation can be developed.

Computational efficiency

In many computational models of perceptual systems, the amount of processing time required to produce a simulation is seen as one of the least important implementation factors. However, one of the motivations behind this thesis is the the production of an 'intelligent' hearing aid which can incorporate the attention effects described previously. If this is the case, the implementation must be capable of operating in real-time and with minimal processing lag. Unfortunately, our binaural model falls short of this criterion: approximately forty minutes of computation time is required for every second of acoustic input (2400 times real-time). However, this may be improved since the model lends itself to a parallel implementation. Every channel of the auditory front end can be processed independently and each node in the network could be implemented on a separate processor.

Grouping principles

Currently, the model only incorporates a small number of primitive grouping principles. For example, active grouping by common onset is not strictly implemented: the old-plus-new heuristic only actively segregates segments with asynchronous onsets. Furthermore, the model does not have mechanism to use common offset information. Another important grouping principle is that of continuity: the model must have the ability to represent a frequency 'trajectory' in order to track a potentially interrupted frequency glide.

In addition to primitive grouping rules, the model would also benefit from the inclusion of schema grouping principles and schema 'detectors' (see chapter 5). As we have discussed earlier in this chapter, schema grouping principles can be

incorporated within the existing neural oscillator framework relatively easily (e.g. Wang *et al.*, 1990; Liu *et al.*, 1994; Wang and Liu, 2002) and would allow commonly heard sounds to influence primitive grouping decisions such as in pattern completion. Schema 'detectors' are important since the conscious allocation of attention can be overridden by endogenous processing when a highly salient piece of information, such as one's name, is detected in one of the unattended groups (e.g. Moray, 1959).

8.5 Conclusions

This chapter has summarised the work presented in this thesis and has also identified a number of possible limitations of the model. The previous section has presented some directions for future work which would improve the explanatory power of the model.

Arguably the two most important areas for future work are investigation into the timecourse of binaural attentional allocation and the incorporation of schema-based influences at the endogenous processing stage. In the introduction to this thesis, we described how the importance of attention in analysing the auditory scene has been known for some time (e.g. Cherry, 1953; Spieth *et al.*, 1954). However, we also noted that it is only recently that psychophysical experiments have begun to look at how attention really influences our ability to perform ASA. Future research must continue to investigate the importance of attention in ASA if we hope to produce devices which attempt to mimic the human perception of sound.

Chapter 9. References

Abeles, M (1991). *Corticonics: Neural circuits of the Cerebral Cortex*. Cambridge University Press.

Aertsen, A, Diesmann, M and Gewaltig, M-O (1996). Propagation of synchronous spiking activity in feedforward neural networks. *Journal of Physiology* (Paris) **90** 243–247.

Alain, C and Woods, DL (1997). Attention modulates auditory pattern memory as indexed by event-related brain potentials. *Psychophysiology* **34** 534-546.

Anstis, S and Saida, S (1985). Adaptation to auditory streaming of frequency-modulated tones. *Journal of Experimental Psychology: Human Perception Performance* **11** 257-271.

Ashby, FG, Prinzmental, W, Ivry, R and Maddox, WT (1996). A formal theory of feature binding in object perception. *Psychological Review* **103** 165–192.

Assmann, PF (1996). Modeling the perception of concurrent vowels: Role of formant transitions. *Journal of the Acoustical Society of America* **100**(2) 1141-1152.

Assmann, PF and Summerfield, Q (1987). Perceptual segregation of concurrent vowels. *Journal of the Acoustical Society of America* **82** S120.

Assmann, PF and Summerfield, Q (1990). Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies. *Journal of the Acoustical Society of America* **88** 680-697.

Assmann, PF and Summerfield, Q (1994). The contribution of waveform interactions to the perception of concurrent vowels. *Journal of the Acoustical Society of America* **95**(1) 471-484.

Baird, B (1996). *A cortical network model of cognitive attentional streams, rhythmic expectation, and auditory stream segregation*. CPAM Technical Report 173-96, Dept of Mathematics, U.C.Berkeley, Berkeley, California.

Barker, J, Green, P and Cooke, MP (2001). Linking auditory scene analysis and robust ASR by missing data techniques. *Proceedings of WISP 2001*, Stratford-upon-Avon, UK

Barlow, HB (1972). Single units and sensation: A neuron doctrine for perceptual psychology? *Perception* **1** 371-394.

Barth, DS and MacDonald, KD (1996). Thalamic modulation of high-frequency oscillating potentials in auditory cortex. *Nature* **383** 78-81.

Beauvois, MW and Meddis, R (1991). A computer model of auditory stream segregation. *Quarterly Journal of Experimental Psychology* **43A**(3) 517-541.

Beauvois, MW and Meddis, R (1996). Computer simulation of auditory stream segregation in alternating-tone sequences. *Journal of the Acoustical Society of America* **99** 2270-2280.

Biederman, I (1987). Recognition by components: A theory of human image understanding. *Psychological Review* **94** 115-147.

Boer, E de and Jongh, HR de (1978). On cochlear encoding: potentialities and limitations of the reverse correlation technique. *Journal of the Acoustical Society of America* **63** 115-135.

Braitenberg, V (1978). Cell assemblies in the cerebral cortex. In *Lecture Notes in Biomathematics Volume 21 Theoretical Approaches in Complex Systems*, edited by R.Heim and G.Palm, Springer.

Brecht, M, Singer, W and Engel, AK (1998). Correlation analysis of corticotectal interactions in the cat visual system. *Journal of Neurophysiology* **79** 2394–2407.

Bregman, AS (1978). Auditory streaming is cumulative, *Journal of Experimental Psychology: Human Perception and Performance* **4** 380–387.

Bregman, AS (1990). *Auditory Scene Analysis. The Perceptual Organization of Sound*, MIT Press.

Bregman, AS and Campbell, J (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology* **89** 244-249.

Bregman, AS and Dannenbring, G (1973). The effects of continuity on auditory stream segregation. *Perceptual Psychophysics* **13** 308-312.

Bregman, AS and Pinker, S (1978). Auditory streaming and the building of timbre. *Canadian Journal of Psychology* **32**(1) 19-31.

Bregman, AS and Rudnicky, A (1975). Auditory segregation: Stream or streams? *Journal of Experimental Psychology: Human Perception and Performance* **1** 263-267.

Bregman, AS and Steiger, H (1980). Auditory streaming and vertical localisation: Interdependence of ‘what’ and ‘where’ decisions in audition. *Perception and Psychophysics* **28** 539-546.

Broadbent, DE (1958). *Perception and communication*. Pergamon Press, New York.

Broadbent, DE and Ladefoged, P (1957). On the fusion of sounds reaching different sense organs. *Journal of the Acoustical Society of America* **29** 708-710.

Brokx, JPL and Nooteboom, SG (1982). Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics* **10** 23-36.

Brown, GJ (1992). *Computational auditory scene analysis: A representational approach*, Doctoral thesis CS-92-22, Department of Computer Science, University. of Sheffield.

Brown, GJ and Cooke, MP (1994). Computational auditory scene analysis. *Computer Speech and Language* **8** 297-336.

Brown, G and Cooke M (1997). Temporal synchronisation in a neural oscillator model of primitive auditory stream segregation. In *Readings in Computational Auditory Scene Analysis*, edited by H.Okuno and D.Rosenthal, Lawrence Erlbaum.

Brown, GJ and Wang, DL (1996). *A neural oscillator model of concurrent vowel perception*. Technical Report CS-96-06, Department of Computer Science, University of Sheffield, UK.

Brown, GJ and Wang, DL (1997). Modelling the perceptual segregation of double vowels with a network of neural oscillators. *Neural Networks* **10**(9) 1547-1558.

Brown, GJ, Wang, DL and Barker, J (2001). A neural oscillator sound separator for missing data speech recognition. *Proceedings of IJCNN*, Washington DC, July 14-19.

Bruce, C, Desimone, R and Gross, C (1981). Visual properties of neurons in a polysensory area in the superior temporal sulcus of the macaque. *Journal of Neurophysiology* **46** 369–384.

Cariani, P, Delgutte, B and Tramo, M (1997). Neural representation of pitch through autocorrelation. *Proceedings of the Audio Engineering Society Meeting (AES)*, New York, September, 1997.

Cariani, P (1999). Temporal coding of periodicity pitch in the auditory system: An Overview. *Neural Plasticity* **6**(4) 147-172

Carlyon, RP, Cusack, R, Foxton, JM and Robertson, IH (2001). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance* **27**(1) 115-127.

Carr, CE and Konishi, M (1988). Axonal delay lines for time measurement in the owl's brainstem. *Proceedings of the National Academy of Sciences of USA* **85** 8311-8315.

Cave, KR and Wolfe, JM (1990). Modeling the role of parallel processing in visual search. *Cognitive Psychology* **22** 225–271.

Cherry, EC (1953). Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustical Society of America* **25** 975-979.

Cheveigné, A de (1993). Separation of concurrent harmonic sounds: fundamental frequency estimation and a time-domain cancellation model of auditory processing. *Journal of the Acoustical Society of America* **93**(6) 3271-3290.

Churchland, PS and Sejnowski, TJ (1992). *The Computational Brain*. MIT Press.

Ciocca, V and Bregman, AS (1987). Perceived continuity of gliding and steady-state tones through interrupting noise. *Perception and Psychophysics* **42**(5) 476-484.

Ciocca, V and Darwin, CJ (1993). Effects of onset asynchrony on pitch perception: adaptation or grouping? *Journal of the Acoustical Society of America* **93**(5) 2870-2878.

Crick, F (1984). Function of the thalamic reticular complex: The searchlight hypothesis. *Proceedings of the National Academy of Sciences of USA* **81** 4586-4590.

Cohen, A and Ivry, RB (1989). Illusory conjunctions inside and outside the focus of attention. *Journal of Experimental Psychology: Human Perception and Performance* **15** 650–663.

Cooke, MP (1991/1993). *Modelling auditory processing and organisation.* Cambridge University Press.

Cooke, M, Brown, GJ, Crawford, M and Green, P (1993). Computational auditory scene analysis: listening to several things at once. Endeavour, **17**(4), 186-190.

Corteen, RS and Wood, B (1972). Autonomic responses to shock-associated words in an unattended channel. *Journal of Experimental Psychology* **94** 308-313.

Cusack, R and Carlyon, RP (2001). Personal communication.

Damasio, AR (1985). Disorders of complex visual processing: agnosia, achromatopsia, Balint's syndrome, and related difficulties of orientation and construction. In *Principles of Behavioural Neurology*, edited by M.M.Mesulam, Davis.

Damasio, AR (1989). The brain binds entities and events by multiregional activation from convergence zones. *Neural Computation* **1** 123-132.

Darwin, CJ (1984). Perceiving vowels in the presence of another sound: constraints on formant perception. *Journal of the Acoustical Society of America* **76** 1636-1647.

Darwin, CJ and Bethell-Fox, CE (1977). Pitch continuity and speech source attribution. *Journal of Experimental Psychology: Human Perception and Performance* **3** 665-672.

Darwin, CJ and Ciocca, V (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.

Darwin, CJ and Hukin, RW (1999). Auditory objects of attention: the role of interaural time-differences in attention to speech. *Journal of Experimental Psychology: Human Perception and Performance* **25** 617-629.

Darwin, CJ and Sutherland, NS (1984). Grouping frequency components of vowels: when is a harmonic not a harmonic? *Quarterly Journal of Experimental Psychology* **36A** 193-208.

Darwin, CJ, Hukin, RW and Al-Khatib, BY (1995). Grouping in pitch perception: Evidence for sequential constraints. *Journal of the Acoustical Society of America* **98**(2) 880-885.

Darwin, CJ, Pattison, H and Gardner, RB (1989). Vowel quality changes produced by surrounding tone sequences. *Perception and Psychophysics* **45** 333-342.

Dawson, ME and Schell, AM (1982). Electrodermal responses to attended and nonattended significant stimuli during dichotic listening. *Journal of Experimental Psychology: Human Perception and Performance* **8** 315-324.

DeCharms, RC, Blake, DT and Merzenich, MM (1998). Optimizing sound features for cortical neurons. *Science* **280** 1439–1443.

Denbigh, PN and Zhao, J (1992). Pitch extraction and separation of overlapping speech. *Speech Communication* **11**(2-3) 119-125.

Desimone, R and Duncan, J (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience* **18** 193–222.

Deutsch, D (1974). An auditory illusion. *Journal of the Acoustical Society of America* **55** 518-519.

Deutsch, D (1975). Two-channel listening to musical scales. *Journal of the Acoustical Society of America* **57** 1156-1160.

Deutsch, D (1986). Auditory pattern recognition. In *Handbook of Perception and Performance* **2**, edited by K.Boff, L.Kaufman and J.Thomas. Wiley.

Deutsch, D and Roll, P (1976). Separate 'what' and 'where' decision mechanisms in processing a dichotic tonal sequence. *Journal of Experimental Psychology: Human Perception and Performance* **2** 23-29.

Deutsch, JA and Deutsch, D (1963). Attention: Some theoretical considerations. *Psychological Review* **70** 80-90.

DeValois, RL and DeValois, KK (1988). *Spatial Vision*. Oxford University Press.

Diesmann, M, Gewaltig, M-O and Aertsen, A (1997). Cortical synfire activity - a two dimensional state space analysis. In *From Membrane to Mind: Proceedings of the 25th Gottinger Neurobiology Conference*, edited by H.Wassle and N.Elsner. Thieme-Verlag.

Donnelly, N, Humphreys, GW and Riddoch, MJ (1991). Parallel computation of primitive shape descriptions. *Journal of Experimental Psychology: Human Perception and Performance* **17** 561-570.

Doupe, AJ (1997). Song- and order-selective neurons in the songbird anterior forebrain and their emergence during vocal development. *Journal of Neuroscience* **17** 1147–1167.

Dowling, WJ (1973). Rhythmic groups and subjective chunks in memory for melodies. *Perception and Psychophysics* **14** 37-40.

Dubnowski, JJ, Schafer, RW and Rabiner, LR (1975). Real time digital hardware pitch detector. *IEEE Transactions on Acoustics, Speech and Signal Processing* **24** 2-8.

Duncan, J (1980). The locus of interference in the perception of simultaneous stimuli. *Psychological Review* **87** 272-300.

Duncan, J (1984). Selective attention and the organisation of visual information. *Journal of Experimental Psychology: General* **113** 501-517.

Duncan, J, Martens, S and Ward, R (1997). Restricted attentional capacity within but not between sensory modalities. *Nature* **387** 808-810.

Edelman, GM (1978). Group selection and phasic reentrant signaling: a theory of higher brain function. In *The Mindful Brain: Cortical Organization and the Group-Selective Theory of Higher Brain Function*, edited by G.M.Edelman and V.B.Mountcastle. MIT Press.

Eggermont, JJ (2001). Between sound and perception: reviewing the search for a neural code. *Hearing Research* **157**(1-2) 1-42.

Elder, J and Zucker, S (1993). The effect of contour closure on the rapid discrimination of two-dimensional shapes. *Vision Research* **33** 981-991.

Ellis, DPW (1996). *Prediction-driven computational auditory scene analysis*, Doctoral thesis, MIT Department of Electrical Engineering and Computer Science.

Engel, AK, König, P, Kreiter, AK, Chillen, TB and Singer, W (1992). Temporal coding in the visual cortex: new vista on integration in the nervous system. *Trends in Neurosciences* **15** 218-225.

Engel, AK, Roelfsema, PR, Fries, P, Brecht, M and Singer, W (1997). Role of the temporal domain for response selection and perceptual binding. *Cerebral Cortex* **7** 571-582.

Engelmore, R and Morgan, T (1988). *Blackboard Systems*. Addison-Wesley, Reading, MA.

Eriksen, CW and Webb, JM (1989). Shifting of attentional focus within and about a visual display. *Perception & Psychophysics* **45** 175-183.

Erman, LD, Hayes-Roth, F, Lesser, VR and Reddy, DR (1980). The HEARSAY-II speech-understanding system: integrating knowledge to resolve uncertainty. *Computing Surveys* **12** 213-253.

Fant, G (1960). *Acoustic theory of speech production*. Mouton, The Hague.

Feng, AS and Ratnam, R (2000). Neural basis of hearing in real-world situations. *Annual Review of Psychology* **51** 699–725.

FitzHugh, R (1961). Impulses and physiological states in models of nerve membrane. *Biophysical Journal* **1** 445-466.

Gabor, D (1946). Theory of communication. *Journal of the Institution of Electrical Engineers* **93** 429–457.

Gallant, JL, Connor, CE, Rakshit, S, Lewis, JW and Van Essen, DC (1996). Neural responses to polar, hyperbolic, and cartesian gratings in area V4 of the macaque monkey. *Journal of Neurophysiology* **76** 2718–2739.

Gerald, H, Tomlinson, BE and Gibson, PH (1980). Cell counts in human cerebral cortex in normal adults throughout life using an image analysing computer. *Journal of Neurology* **46** 113-136.

Gerstein, GL, Bedenbaugh, P and Aertsen, MH (1989). Neuronal assemblies. *IEEE Transactions on Biomedical Engineering* **36** 4–14.

Gibson, JR and Maunsell, JHR (1997). The sensory modality specificity of neural activity related to memory in visual cortex. *Journal of Neurophysiology* **78** 1263–1275.

Glasberg, BR and Moore, BCJ (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research* **47** 103–138.

Godsmark, D and Brown, GJ (1999). A blackboard architecture for computational auditory scene analysis. *Speech Communication* **27** 351-366.

Goldstein, JL (1967). Auditory nonlinearity. *Journal of the Acoustical Society of America* **41** 676-689.

Goldstein, JL (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.

Goldstone, RL (1998). Perceptual learning. *Annual Review of Psychology* **49** 585–612.

Graham, N (1989). *Visual Pattern Analyzers*. Oxford University Press.

Gray, CM and Singer, W (1989). Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proceedings of the National Academy of Science USA* **86** 1698–1702.

Gray, CM, Koenig, P, Engel, AK and Singer, W (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* **338** 334–337.

Greenberg, GZ and Larkin, WD (1968). Frequency-response characteristic of auditory observers detecting signals of a single frequency in noise: The probe-signal method. *Journal of the Acoustical Society of America* **44** 1513-1523.

Green, DM (1960). Psychoacoustics and detection theory, *Journal of the Acoustical Society of America* **32** 1189-1203.

Green, DM (1961). Detection of auditory sinusoids of uncertain frequency, *Journal of the Acoustical Society of America* **33** 904-911.

Green, DM and Swets, JA (1966). *Signal Detection Theory and Psychophysics*, Wiley.

Griffith, JS (1963). On the stability of brain-like structures. *Biophysical Journal* **3** 299-308.

Hafter, ER, Schlauch, RS and Tang, J (1993). Attending to auditory filters that were not stimulated directly. *Journal of the Acoustical Society of America* **94** 743-747.

Hartmann, WM and Johnson, D (1991). Stream segregation and peripheral channeling. *Music Perception* **9**(2) 155-184.

Hebb, DO (1949). *The organization of behavior*. New York: Wiley & Sons.

Hewitt, MJ and Meddis, R (1991). An evaluation of eight computer models of mammalian inner hair cell function, *Journal of the Acoustical Society of America* **90**(2) 904-917.

Hodgkin, AL and Huxley, AF (1952). A quantitative description of membrane current and its application to conduction and excitation in the nerve. *Journal of Physiology* **117** 500-544.

Holdsworth, J, Nimmo-Smith, I, Patterson, R and Rice, P (1988). Implementing a gammatone filter bank. Annex C of the *SVOS Final Report* (part A: The auditory filter bank), MRC Applied Psychology Unit, Cambridge, UK.

Horikawa, J, Tanahashi, A and Suga, N (1994). Afterdischarges in the auditory cortex of the moustached bat - no oscillatory discharges for binding auditory information. *Hearing Research* **76** 45-52.

Hubel, DH (1988). *Eye, brain, and vision*. Scientific American Library. New York: Freeman and Company.

Hubel, D and Wiesel, T (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology* **28** 229–289.

Hummel, JE and Biederman, I (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review* **99** 480–517.

ISO. Normal equal-loudness level contours (ISO 226), International Organization for Standardization.

Itti, L and Koch, C (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research* **40** 1489–1506.

James, W (1890/1950). *The Principles of Psychology,* Volume 1. Dover.

Johnson, DM and Hafter, ER (1980). Uncertain-frequency detection: Cuing and condition of observation. *Perceptual Psychophysics* **28** 143-149.

Joliot, M, Ribary, U and Llinás, R (1994). Human oscillatory brain activity near 40 Hz coexists with cognitive temporal binding. *Proceedings of the National Academy of Sciences of the USA* **91** 11748-51.

Jones, MR (1976). Time, our lost dimension: Toward a new theory of perception, attention and memory. *Psychological Review* **83** 323-355.

Jones, MR, Kidd, G and Wetzel, R (1981). Evidence for rhythmic attention. *Journal of Experimental Psychology: Human Perception and Performance* **7** 1059-1073.

Jones, MR, Maser, DJ and Kidd, GR (1978). Rate and structure in memory for auditory patterns. *Memory and Cognition* **6** 246-258.

Jonides, J and Yantis, S (1988). Uniqueness of abrupt visual onset in capturing attention. *Perception & Psychophysics* **43** 346-354.

Kahneman, D, Treisman, A and Gibbs, B (1992). The reviewing of object files: object-specific integration of information. *Cognitive Psychology* **24** 175–219.

Kastner, S, De Weerd, P, Desimone, R and Ungerleider, LG (1998). Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. *Science* **282** 108–111.

Kastner, S, Pinsk, MA, De Weerd, P, Desimone, R and Ungerleider, LG (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron* **22** 751–761.

Keil, K, Müller, MM, Ray, WJ, Gruber, T and Elbert, T (1999). Human gamma band activity and perception of a Gestalt. *Journal of Neuroscience* **19** 7152–7161.

Kemp, DT (1978). Stimulated acoustic emissions from within the human auditory system. *Journal of the Acoustical Society of America* **64** 1386-1391.

Kilgard, MP and Merzenich, MM (1998). Plasticity of temporal information processing in the primary auditory cortex. *Nature Neuroscience* **1** 727–731.

Kim, DO, Molnar, CE and Matthews, JW (1980). Cochlear mechanics: nonlinear behaviour in two-tone responses as reflected in cochlear-nerve-fibre responses and in ear-canal sound pressure. *Journal of the Acoustical Society of America* **67** 1704-1721.

Kobatake, E and Tanaka, K (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology* **71** 856–857.

Kobatake, E, Wang, G and Tanaka, K (1998). Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. *Journal of Neurophysiology* **80** 324–330.

Koch, C and Ullman, S (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology* **4** 219–227.

Koffka, K (1936). *Principles of Gestalt psychology.* Harcourt and Brace, New York.

Konen, W and von der Malsburg, C (1993). Learning to generalize from single examples in the dynamic link architecture. *Neural Computation* **5** 719–735.

Konishi, M, Kahashi, TT, Wagner, H, Sullivan, W and Carr, CE (1988). Neurophysiological and anatomical substrates of sound localisation in the owl. In *Auditory Function*, edited by G.M.Edelman, W.E.Gall and W.M.Cowan (721-745), Wiley.

Kovacs, I and Julesz, B (1993). A closed curve is much more than an incomplete one: effect of closure in figure-ground segmentation. *Proceedings of the National Academy of Science USA* **90** 7495-7497.

Lapicque, L (1907). Recherches quantitatives sur l'excitation electrique des nerfs traitee comme une polarization. *Journal de Physiologie et Pathologie General* **9** 620-635.

Large, EW and Jones, MR (1999). The dynamics of attending: How we track time varying events. *Psychological Review* **106**(1) 119-159.

Lesser, VR, Nawab, SH and Klassner, FI (1995). IPUS: An architecture for the integrated processing and understanding of signals. *Artificial Intelligence* **77** 129-171.

Lettvin JY, (1995). J.Y.Lettvin on grandmother cells. In *The Cognitive Neurosciences*, edited by M.S.Gazzaniga (434-435). MIT Press.

Liberman, AM (1982). On finding that speech is special. *American Psychologist* **37** 148-167.

Licklider, JCR (1951). A duplex theory of pitch perception. *Experientia* **7**(4) 128-134.

Licklider JCR (1959). Three auditory theories. In *Psychology: A Study of a Science*, edited by S.Koch, McGraw-Hill.

Liu, F, Yamaguchi, Y and Shimizu, H (1994). Flexible vowel recognition by the generation of dynamic coherence in oscillator neural networks: speaker-independent vowel recognition. *Biological Cybernetics* **7** 105-114.

Livingstone, MS (1996). Oscillatory firing and interneuronal correlations in squirrel monkey striate cortex. *Journal of Neurophysiology* **75** 2467–2485.

Livingstone, MS and Hubel, DH (1988). Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science* **240** 740-749.

Logothetis, NK and Pauls, J (1995). Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cerebal Cortex* **5** 270–288.

Luck, SJ, Chelazzi, L, Hillyard, SA and Desimone, R (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *Journal of Neurophysiology* **77** 24–42.

Lyon, RF (1996). The all-pole gammatone filter and auditory models. In *Forum Acusticum '96*, Antwerp, Belgium.

Mack, A, Tang, B, Tuma, Regina and Kahn, S (1992). Perceptual Organisation and Attention. *Cognitive Psychology* **24** 475-501.

MacKay, DM (1973). Visual stability and voluntary eye movements. In *Handbook of Sensory Physiology* **8/3**, edited by R.Jung. Springer-Verlag.

Maes, P (ed., 1991). *Designing autonomous agents: theory and practice from biology to engineering and back*. MIT Press.

Mann, VA and Liberman, AM (1983). Some differences between phonetic and auditory modes of perception. *Cognition* **14** 211-235.

Marr, D (1982). *Vision*. W. H. Freeman and Company.

McCabe, SL and Denham, MJ (1997). A model of auditory streaming. *Journal of the Acoustical Society of America* **101**(3) 1611-1621.

McKeown, JD and Patterson, RD (1995). The time course of auditory segregation: Concurrent vowels that vary in duration. *Journal of the Acoustical Society of America* **98** 1866-1877.

Meddis, R (1986). Simulation of mechanical to neural transduction in the auditory receptor, *Journal of the Acoustical Society of America* **79**(3) 702-711.

Meddis, R (1988). Simulation of auditory-neural transduction: Further studies, *Journal of the Acoustical Society of America* **83**(3) 1056-1063.

Meddis, R and Hewitt, M (1991a). Virtual pitch and phase sensitivity of a computer model of the auditory periphery: I. pitch identification. *Journal of the Acoustical Society of America* **89**(6) 2866-2882.

Meddis, R and Hewitt, M (1991b). Virtual pitch and phase sensitivity of a computer model of the auditory periphery: II. phase sensitivity. *Journal of the Acoustical Society of America* **89**(6) 2883-2894.

Meddis, R and Hewitt, M (1992). Modelling the identification of concurrent vowels with different fundamental frequencies. *Journal of the Acoustical Society of America* **91** 233-245.

Meddis, R and O'Mard, L (1997). A unitary model of pitch perception. *Journal of the Acoustical Society of America* **102**(3) 1811-1820.

Melara, RD and Marks, LE (1990). Perceptual primacy of dimensions: Support for a model of dimension interaction. *Journal of Experimental Psychology: Human Perception and Performance* **16** 398-414.

Mellinger, DK (1991). *Event formation and separation in musical sound*. Doctoral thesis, Stanford University.

Miller, GA and Licklider, JCR (1950). Intelligibility of interrupted speech. *Journal of the Acoustical Society of America* **22** 167-173.

Milner, PM (1974). A model for visual shape recognition. *Psychological Review* **81** 521-535.

Minsky, M (1986). *Society of Minds*. Simon and Schuster, Inc.

Mondor, TA and Bregman, AS (1994). Allocating attention to frequency regions. *Perception & Psychophysics* **56**(3) 268-276.

Mondor, TA and Zatorre, RJ (1995). Shifting and Focusing Auditory Spatial Attention. *Journal of Experimental Psychology: Human Perception and Performance* **21**(2) 387-409.

Mondor, TA, Zatorre, RJ and Terrio, NA (1998). Constraints on the Selection of Auditory Information. *Journal of Experimental Psychology: Human Perception and Performance* **24**(1) 66-79.

Moore, BCJ (1997). *An introduction to the psychology of hearing - 4th edition*. Academic Press.

Moore, BCJ and Glasberg, BR (1983). Suggested formulae for calculating auditory filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America* **74** 750-753.

Moore, CM and Egeth, H (1997). Perception Without Attention: Evidence of Grouping Under Conditions of Inattention. *Journal of Experimental Psychology: Human Perception and Performance* **23** 339-352.

Moore, BCJ and Rosen, SM (1979). Tune recognition with reduced pitch and interval information. *Quarterly Journal of Experimental Psychology* **31** 229-240.

Moore, BCJ, Glasberg, BR and Peters, RW (1985). Relative dominance of individual partials in determining the pitch of complex tones. *Journal of the Acoustical Society of America* **77** 1853-1860.

Moore, BCJ, Glasberg, BR and Peters, RW (1986). Thresholds for hearing mistuned partials as separate tones in harmonic complexes. *Journal of the Acoustical Society of America* **80** 479-483.

Moran, J and Desimone, R (1985). Selective attention gates visual processing in the extrastriate cortex. *Science* **229** 782–784.

Moray, N (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology* **11** 56-60.

Müller, HJ and Rabbitt, PMA (1989). Reflexive and voluntary orienting of visual attention: Time course of activation and resistance to interruption. *Journal of Experimental Psychology: Human Perception and Performance* **15** 315-330.

Nagumo, J, Arimoto, S and Yoshizawa, S (1962). An active pulse transmission line simulating nerve axon. *Proceedings of the Institute of Radio Engineers* **50** 2061-2070.

Nakatani, T, Okuno, H and Kawabata, T (1994). Auditory stream segregation in auditory scene analysis with a multi-agent system. *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Aug. 1994, Seattle, 100-107.

Nakatani, T, Okuno H, Goto M and Ito T (1998). Multiagent based binaural sound stream segregation, in *Computational Auditory Scene Analysis*, edited by D.F.Rosenthal and H.Okuno, Mahwah, NJ: Lawrence Erlbaum, pp. 195–214.

Nakayama, K and Mackeben, M (1989). Sustained and transient components of focal visual attention. *Vision Research* **29** 1631-1647.

Newstead, SE and Dennis, I (1979). Lexical and grammatical processing of unshadowed messages: A re-examination of the MacKay effect. *Quarterly Journal of Experimental Psychology* **31** 477-488.

Niebur, E and Koch, C (1996). Control of selective visual attention. In *Advances in neural information processing systems*, edited by D.Touretzky, M.C.Mozer and M.E.Hasselmo. MIT Press. 802–808.

Niebur, E, Koch, C and Rosin, C (1993). An oscillation-based model for the neuronal basis of attention. *Vision Research* **33** 2789–2802.

Norman, DA (1968). Toward a theory of memory and attention. *Psychological Review* **75** 522-536.

Nothdurft, HC (1993). Faces and facial expression do not pop-out. *Perception* **22** 1287–1298.

O'Regan, JK, Rensink, RA and Clark, JJ (1999). Change blindness as a result of ‘mudsplashes’. *Nature* **398** 34.

Olshausen, BA, van Essen, DC and Anderson, CH (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience* **13** 4700-4719.

Oswald, I, Taylor, AM and Treisman, M (1960). Discriminative responses to stimulation during human sleep. *Brain* **83** 440-453.

Palm, G (1981). Towards a theory of cell assemblies. *Biological Cybernetics* **39** 181–194.

Palm, G (1990). Cell assemblies as a guideline for brain research. *Concepts in Neuroscience* **1** 133–137.

Palmer, AR (1987). Physiology of the cochlear nerve and cochlear nucleus. *British Medical Bulletin* **43**(4) 838-855.

Palomäki, KJ, Brown, GJ and Wang, DL (2001). A binaural model for missing data speech recognition in noisy and reverberant conditions. *Proceedings of the workshop on consistent and reliable cues for sound analysis (CRAC)*, Aalborg, September 2 2001.

Parsons, TW (1976). Separation of speech from interfering speech by means of harmonic selection. *Journal of the Acoustical Society of America* **60**(4) 911-918.

Patterson, RD and Moore, BCJ (1986). Auditory filters and excitation patterns as representations of frequency resolution. In *Frequency Selectivity in Hearing*, edited by B.C.J.Moore, Academic Press, 123-177.

Patterson, RD, Nimmo-Smith, I, Holdsworth, J and Rice, P (1988). *APU Report 2341: An Efficient Auditory Filterbank Based on the Gammatone Function*. Applied Psychology Unit, University of Cambridge, UK.

Perrett, D and Oram, M (1993). Neurophysiology of shape processing. *Image and Vision Computing* **11** 317–333.

Pickles, JO (1988). *An Introduction to the Physiology of Hearing*, 2nd Edition. Academic Press.

Plomp, R (1965). Detectability thresholds for combination tones. *Journal of the Acoustical Society of America* **37** 1110-1123.

Posner, MI (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology* **32** 3-25.

Prinzmetal, W and Keysar, B (1989). Functional theory of illusory conjunctions and neon colors. *Journal of Experimental Psychology: General* **118** 165–190.

Pulvermüller, F, Birbaumer, N, Lutzenberger, W and Mohr, B (1997). High frequency brain activity: its possible role in attention, perception and language processing. *Progress in Neurobiology* **52** 427–445.

Rand, TC (1974). Dichotic release from masking for speech. *Journal of the Acoustical Society of America* **55** 678-680.

Reynolds, JH and Desimone, R (1999). The role of neural mechanisms of attention in solving the binding problem. *Neuron* **24** 19–29.

Reynolds, J, Chelazzi, L and Desimone, R (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *Journal of Neuroscience* **19** 1736–1753.

Rhodes, G (1987). Auditory attention and the representation of spatial information. *Perception & Psychophysics* **42** 1-14.

Riesenhuber, M and Poggio, T (1999). Are cortical models really bound by the 'binding problem'? *Neuron* **24** 87–93.

Roberts, B and Moore, BCJ (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** 2922-2932.

Robles, L, Ruggero, MA and Rich, NC (1991). Two-tone distortion in the basilar membrane of the cochlea. *Nature* **349** 413-414.

Rock, I and Palmer, S (1990). The Legacy of Gestalt Psychology. *Scientific American* **263**, 48-61.

Roskies, AL (1999). The binding problem. *Neuron* **24** 7-9.

Sakai, K and Miyashita, Y (1991). Neural organization for the longterm memory of paired associates. *Nature* **354** 152–155.

Sakai, K and Miyashita, Y (1994). Neuronal tuning to learned complex forms in vision. *Neuroreport* **5** 829–832.

Scheffers, MTM (1983). *Sifting vowels: auditory pitch analysis and sound segregation*. Ph.D. thesis, Groningen University, NL.

Schlauch, RS and Hafter, ER (1991). Listening bandwidths and frequency uncertainty in pure-tone signal detection. *Journal of the Acoustical Society of America* **90** 1332-1339.

Schouten, JF (1940). The residue and the mechanism of hearing. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen* **43** 991-999.

Schouten, JF (1970). The residue revisited. In *Frequency Analysis and Periodicity Detection in Hearing*, edited by R.Plomp and G.F.Smoorenburg, Sijthoff.

Schouten, JF, Ritsma, RJ and Cardozo, BL (1962). Pitch of the residue. *Journal of the Acoustical Society of America* **34** 1418-1424.

Seidemann, E and Newsome, WT (1999). Effect of spatial attention on the responses of area MT neurons. *Journal of Neurophysiology* **81** 1783–1794.

Shadlen, MN and Newsome, WT (1998). The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *Journal of Neuroscience* **18** 3870–3896.

Sherrington, CS (1941). *Man on His Nature*, Cambridge University Press, Cambridge.

Simons, DJ and Levin, DT (1997). Change blindness. *Trends in Cognitive Sciences* **1** 261–268.

Singer, W (1993). Synchronisation of cortical activity and its putative role in information processing and learning. *Annual Review of Physiology* **55** 349-74.

Singer, W and Gray, CM (1995). Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience* **18** 555-586.

Slaney, M (1991). *Release notes for version 2.2 of Lyon's cochlear model*. Apple Computer Technical Report, Apple Computer Inc.

Slaney, M and Lyon, (1993). On the importance of time - a temporal representation of sound. In *Visual Representations of Speech Signals* (95-116), edited by M.Cooke, S.Beet and M.Crawford, Wiley.

Sondhi, MM (1968). New methods of pitch extraction. *IEEE Transactions on Audio and Electroacoustics* **16** 262-268.

Spence, CJ and Driver, J (1994). Covert Spatial Orienting in Audition: Exogenous and Endogenous Mechanisms. *Journal of Experimental Psychology: Human Perception and Performance* **20**(3) 555-574.

Spieth, W, Curtis, JF and Webster, JC (1954). Responding to one of two simultaneous messages. *Journal of the Acoustical Society of America* **26**(3) 391-396.

Summerfield, Q, Lea, A and Marshall, D (1990). Modelling auditory scene analysis: Strategies for source segregation using autocorrelograms. *Proceedings of the Institute of Acoustics* **12**(10) 507-514.

Sussman, E, Ritter, W and Vaughan Jr, HG (1998). Attention affects the organisation of auditory input associated with the mismatch negativity system. *Brain Research* **789** 130-138.

Sussman, E, Ritter, W and Vaughan Jr, HG (1999). An investigation of the auditory streaming effect using event-related brain potentials. *Psychophysiology* **36** 22-34.

Sussman, E and Winkler, I (2001). Dynamic sensory updating in the auditory system. *Cognitive Brain Research* **12** 431–439.

Suzuki, S and Cavanagh, P (1995). Facial organization blocks access to low-level features: an object inferiority effect. *Journal of Experimental Psychology: Human Perception and Performance* **21** 901–913.

Tallon-Baudry, C, Bertrand, O, Peronnet, F and Pernier, J (1998). Induced gamma band activity during the delay of a visual short-term memory task in humans. *Journal of Neuroscience* **18** 4244–4254.

Tanaka, K (1996). Inferotemporal cortex and object vision: stimulus selectivity and columnar organization. *Annual Review of Neuroscience* **19** 109-139.

Terman, D and Wang, DL (1995). Global competition and local cooperation in a network of neural oscillators. *Physica D* **81** 148–176.

Tononi, G and Edelman, GM (1998). Consciousness and complexity. *Science* **282** 1846–1851.

Tononi, G, Srinivasan, R, Russell, DP and Edelman, GM (1998). Investigating neural correlates of conscious perception by frequency-tagged neuromagnetic responses. *Proceedings of the National Academy of Sciences USA* **95** 3198–3203.

Treisman, A (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology* **77** 533-546.

Treisman, A (1964a). Monitoring and storage of irrelevant messages in selective attention. *Journal of Verbal Learning and Verbal Behavior* **3** 449-459

Treisman, A (1964b). Selective attention in man. *British Medical Bulletin* **20** 12-16.

Treisman, A (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London Series B - Biological Sciences* **353** 1295–1306.

Treisman, A (1992). Perceiving and reperceiving objects. *American Psychologist* **47** 862–875.

Treisman, A and Gelade, G (1980). A feature integration theory of attention. *Cognitive Psychology* **12** 97-136.

Treisman, A and Gormican, S (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review* **95** 15-48.

Treisman, A and Schmidt, H (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology* **14** 107-141.

Trejo, LJ, Ryan-Jones, DL and Kramer, AF (1995). Attentional modulation of the mismatch negativity elicited by frequency differences between binaurally presented tone bursts. *Psychophysiology* **32** 319-328.

Treue, S and Maunsell, JHR (1996). Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature* **382** 539–541.

Tsal, Y (1989). Do illusory conjunctions support feature integration theory? A critical review of theory and findings. *Journal of Experimental Psychology: Human Perception and Performance* **15** 394–400.

Tuckwell, HC (1988). *Introduction to Theoretical Neurobiology.* Cambridge University Press.

van der Pol, B (1926). On relaxation oscillations. *Philosophical Magazine* **2**(11) 978–992.

van Essen, DC and Gallant, JL (1994). Neural mechanisms of form and motion processing in the primate visual system. *Neuron* **13** 1–10.

van Essen, DC and Maunsell, JHR (1983). Hierarchical organization and functional streams in the visual cortex. *Trends in Neurosciences* **6** 370-375.

van Noorden, LPAS (1975). *Temporal coherence in the perception of tone sequences*. Doctoral thesis, Institute for Perceptual Research, Eindhoven, NL.

von der Malsburg, C (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik* **14** 85-100.

von der Malsburg, C (1981). *The correlation theory of brain function*. Internal report 81-2, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany.

von der Malsburg, C (1986). Am I thinking assemblies? In *Proceedings of the Trieste Meeting on Brain Theory*, edited by G.Palm and A.Aertsen. Springer.

von der Malsburg, C (1988). Pattern recognition by labeled graph matching. *Neural Networks* **1** 141–148.

von der Malsburg, C (1995). Binding in models of perception and brain function. *Current Opinion in Neurobiology* **5** 520–526.

von der Malsburg, C (1999). The what and why of binding: the modeler's perspective. *Neuron* **24** 95-104.

von der Malsburg, C and Schneider, W (1986). A neural cocktail-party processor. *Biological Cybernetics* **54** 29-40.

Wallis, G and Rolls, E (1997). A model of invariant object recognition in the visual system. *Progress in Neurobiology* **51** 167–294.

Wang, DL (1993). Modeling global synchrony in the visual cortex by locally coupled neural oscillators. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (1058-1063). Hillsdale, NJ: Erlbaum.

Wang, DL (1995). Emergent synchrony in locally coupled neural oscillators. *IEEE Transactions on Neural Networks* **6** 941-948.

Wang, DL (1996). Primitive auditory segregation based on oscillatory correlation. *Cognitive Science* **20** 409-456.

Wang, DL (1999). Object selection based on oscillatory correlation. *Neural Networks* **12** 579–592.

Wang, DL (2000). On connectedness: A solution based on oscillatory correlation. *Neural Computation* **12** 131-139.

Wang, DL (2001). Personal communication.

Wang, DL and Brown, GJ (1999). Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Transactions on Neural Networks* **10** 684-697.

Wang, DL and Liu, X (2002). Scene analysis by integrating primitive segmentation and associative memory. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* **32** 254-268.

Wang, DL and Terman, D (1995). Locally excitatory globally inhibitory oscillator networks. *IEEE Transactions on Neural Networks* **6** 283-286.

Wang, D, Buhmann, J and von der Malsburg, C (1990). Pattern segmentation in associative memory. *Neural Computation* **2** 94-106.

Warren, RM and Warren, RP (1970). Auditory illusions and confusions. *Scientific American* **223**(12) 30-36.

Warren, RM, Obusek, CJ and Ackroff, JM (1972). Auditory induction: Perceptual synthesis of absent sounds. *Science* **176** 1149-1151.

Wessel, DL (1978). *Timbre space as a musical control structure*. Institut de Recherche et Coordination Acoustique/Musique (Ircam) Report 12/78.

Weintraub, M (1985). *A theory and computational model of auditory monaural sound separation*. Doctoral thesis, Deptartment of Electrical Engineering, Stanford University.

Williams, SM, Green, PD and Nicolson, RI (1990). Streamer: mapping the auditory scene. *Proceedings of the Institute of Acoustics* **12**(10) 567-575.

Winkler, I and Czigler, I (1998). Mismatch negativity: deviance detection or the maintenance of the 'standard'. *NeuroReport* **9** 3809–3813.

Winslow, JT, Parr, LA and Davis M (2002). Acoustic startle, prepulse inhibition, and fear-potentiated startle measured in rhesus monkeys. *Biological Psychiatry* **51**(11) 859-866.

Wolfe, JM (1996). Extending Guided Search: why Guided Search needs a preattentive 'item map'. In *Converging Operations in the Study of Visual Selective Attention*, edited by A.Kramer, G.H.Cole and G.D.Logan. American Psychological Association.

Wolfe, JM and Bennett, S (1997). Preattentive object files: shapeless bundles of basic features. *Vision Research* **37** 25–44.

Wolfe, JM and Cave, KR (1999). The psychophysical evidence for a binding problem in human vision. *Neuron* **24** 11-17.

Wolfe, JM, Cave, KR and Franzel, SL (1989). Guided Search: an alternative to the Feature Integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance* **15** 419–433.

Wrigley, SN (1999). *Synfire chains as a neural mechanism for auditory grouping*. Technical Report CS-99-11, Department of Computer Science, University of Sheffield, UK.

Wrigley, SN and Brown, GJ (2000). Synfire chains as a neural mechanism for auditory grouping. *British Journal of Audiology* **34**(2) 116-117.

Wrigley, SN and Brown, GJ (2001). A neural oscillator model of auditory attention. *Lecture Notes in Computer Science* **2130** 1163-1170.

Wrigley, SN and Brown, GJ (2002). A neural oscillator model of auditory selective attention. In *Advances in Neural Information Processing Systems 14*, edited by T.G.Dietterich, S.Becker and Z.Ghahramani, MIT Press.

Yeshurun, Y and Carrasco, M (1999). Spatial attention improves performance in spatial resolution tasks. *Vision Research* **39** 293–306.

Yin, TCT and Chan, JCK (1988). Neural mechanisms underlying interaural time sensitivity to tones and noise. In *Auditory Function*, edited by G.M.Edelman, W.E.Gall and W.M.Cowan (721-745), Wiley.

Yu, AC and Margoliash, D (1996). Temporal hierarchical control of singing in birds. *Science* **273** 1871–1875.

Zatorre, RJ, Mondor, TA and Evans, AC (1999). Auditory Attention to Space and Frequency Activates Similar Cerebral Systems. *NeuroImage* **10** 544-554.

Appendix A. Computational Model Parameters

Parameter	Description	Typical Value
n	Gammatone filter order	4
s	Size of the difference of gaussians (DOG) kernel	5
w	DOG weighting	0.8
σ	Size of the gaussian kernel in the DOG function	2
τ_{max}	Maximum autocorrelation lag	20 ms
P	Autocorrelation window size	25 ms
N	Number of channels in the correlogram	128
K	Steepness of the sigmoid in the squash function	50
ε	Oscillator parameter	0.4
γ	Oscillator parameter	6.0
β	Oscillator parameter	0.1
I_{low}	'Off' oscillator input	-5.0
I_{high}	'On' oscillator input	0.2
W_z	Global inhibitor weighting	0.7
W_{ik}	Weight between oscillators i and k	-
-	Internode excitatory weight within segments	1.0
-	Internode excitatory weight between segments	5.0
θ_s	Segment membership threshold	0.3
θ_t	Segment tonal threshold	0.7
θ_n	Segment noise threshold	0.2
θ_x	Oscillator activity influence (squash) factor	-0.5
θ_z	Global inhibitor activity influence (squash) factor	0.1

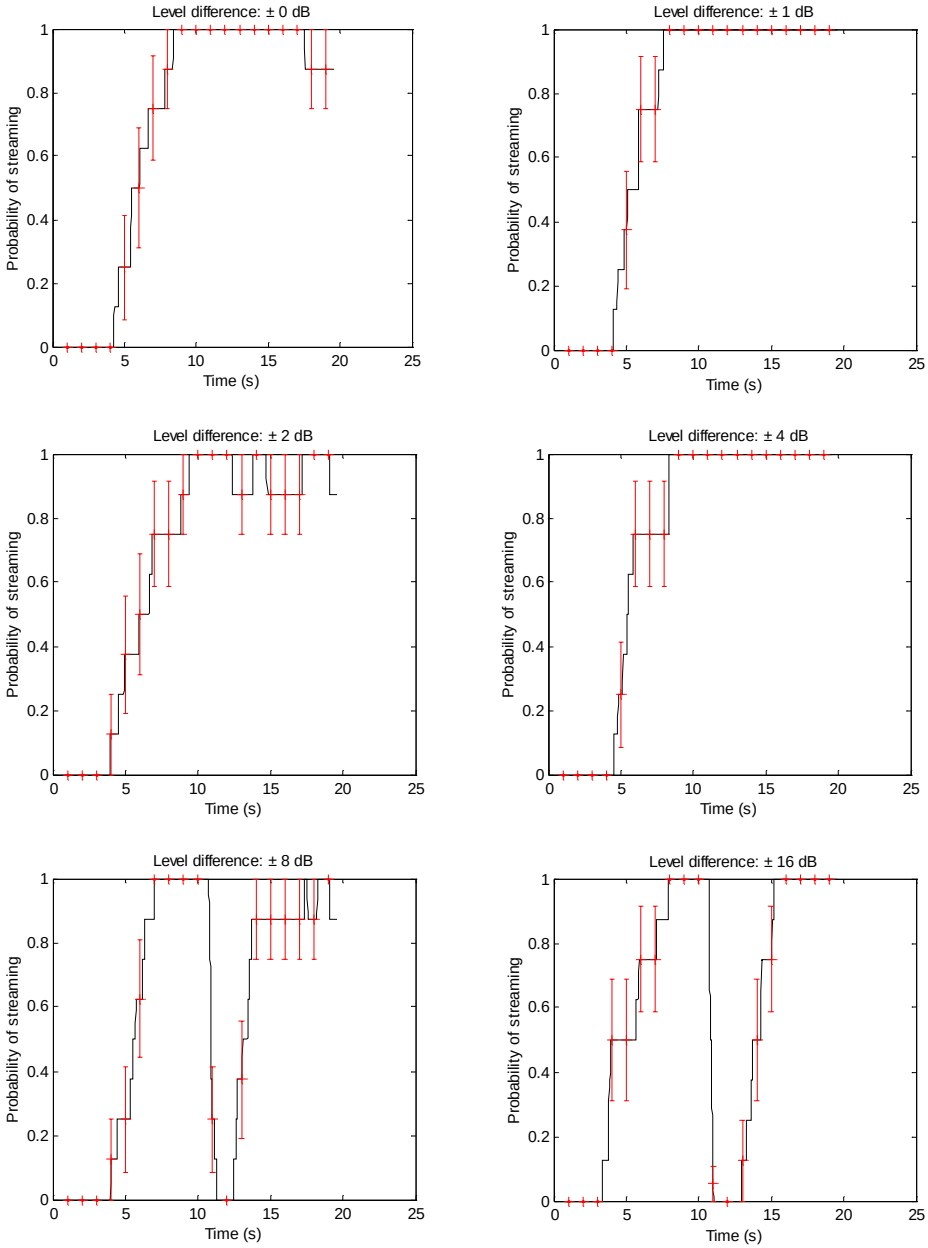
Parameter	Description	Typical Value
θ_{clip}	Normalised centre clipping value	0.3
θ_c	Energy : autocorrelation value threshold	0.65
c_B	Age tracker parameter (decay)	5
d_B	Age tracker parameter (rise)	0.001
g_B	Age tracker parameter (gain)	3
θ_a	Age tracker threshold	0.1
θ_α	Normalising factor for instantaneous envelope	3
c_L	Attentional leaky integrator parameter (decay)	1
d_L	Age tracker parameter (rise)	0.0005
g_L	Age tracker parameter (gain)	3
θ_{ALI}	Threshold above which oscillator array can influence ALI	0.2
σ_{ALI}	Width of the attentional interest peak	6
max_{Ak}	Maximum value of the attentional interest vector	1
min_{Ak}	Minimum value of the attentional interest vector	0.05

Appendix B. Psychophysical Experiment Subject Responses

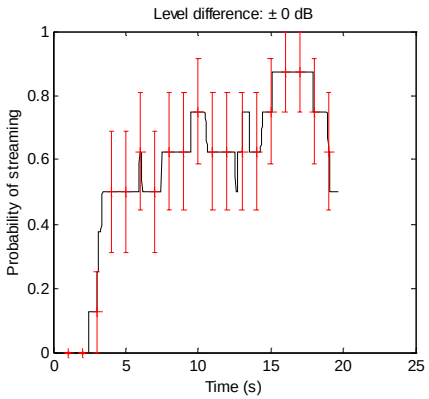
B.1 Introduction

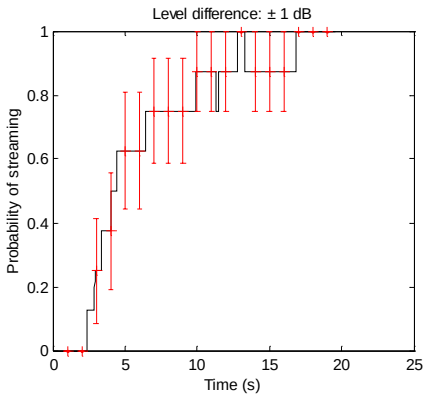
In section 7.5, we presented a psychophysical experiment in which the effect of source movement on the time course of auditory streaming was investigated. It was hypothesised that a movement of the source, and hence a movement of the attentional focus, would result in an 'reset' of the attentional buildup responsible for the streaming percept. This source movement was simulated by applying an interaural intensity difference to the binaural signal. In order to determine if a minimum movement was required to illicit a reset response, a range of movement distances were simulated. 5 subjects participated in the study and each was presented with 48 stimuli (6 intensity differences x 8 repetitions). The results shown in chapter 7 figure 18 were produced by averaging the responses to each intensity difference across subjects. In this appendix, the average of each subject's responses to each intensity difference is presented and error bars indicating ± 1 standard error are displayed at one second intervals. Finally, the same diagrams as shown in chapter 7 figure 18 are presented but with the addition of error bars indicating ± 1 standard error.

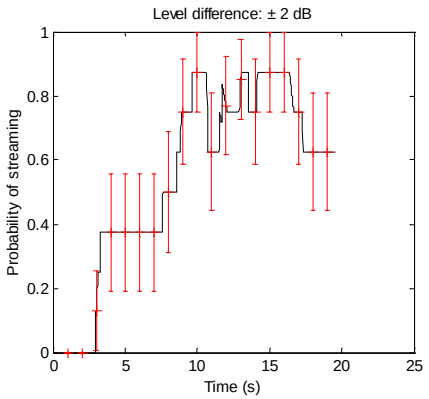
B.2 Subject GB

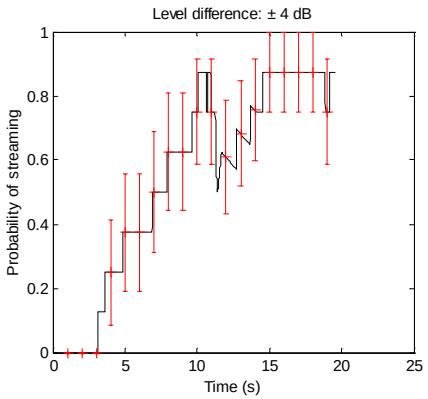


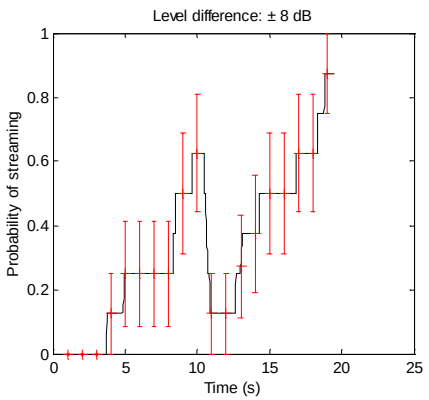
B.3 Subject JE

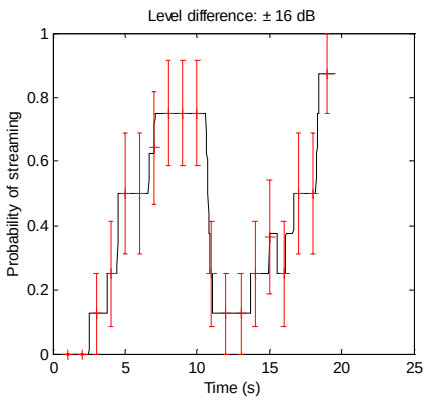




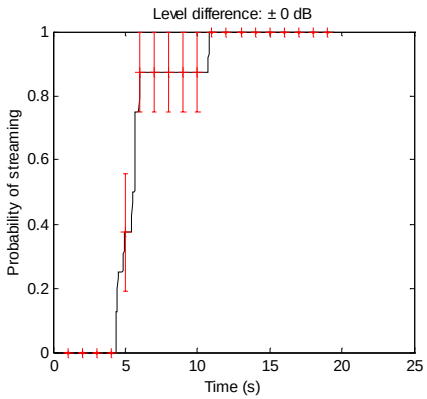


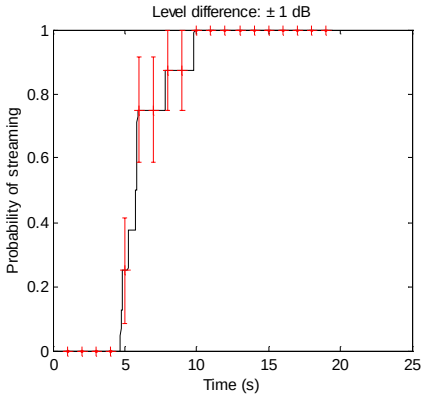


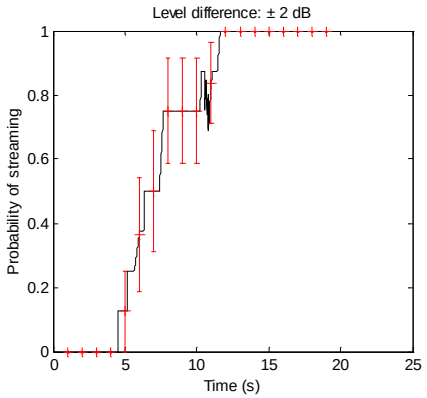


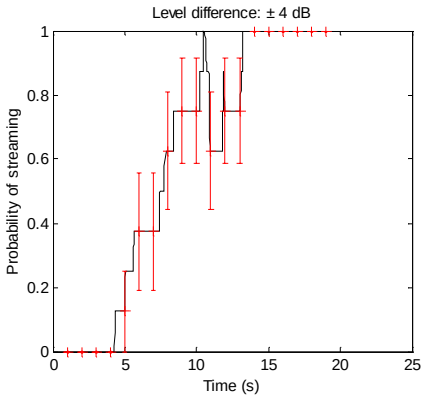


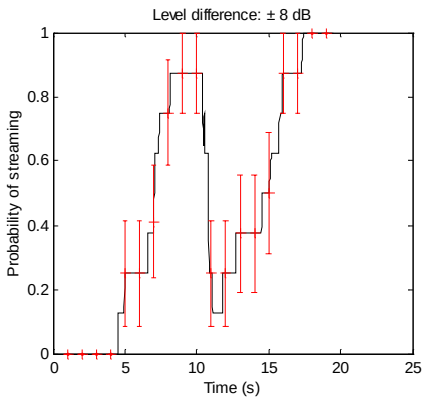
B.4 Subject SC

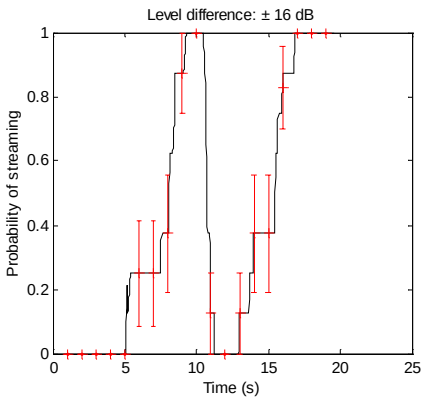




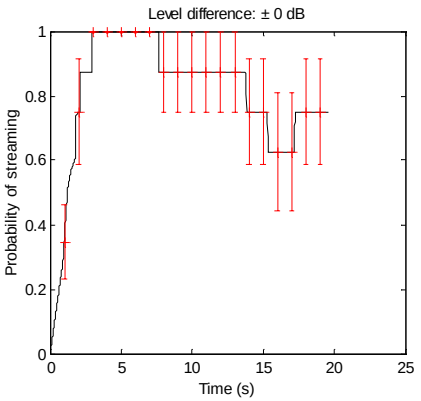


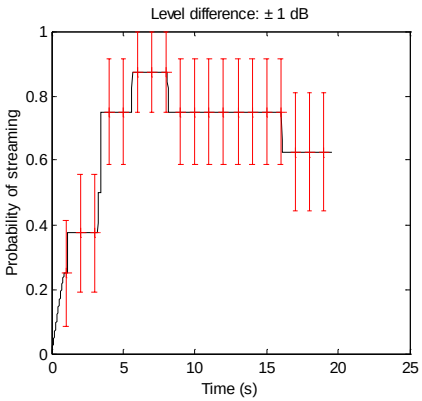


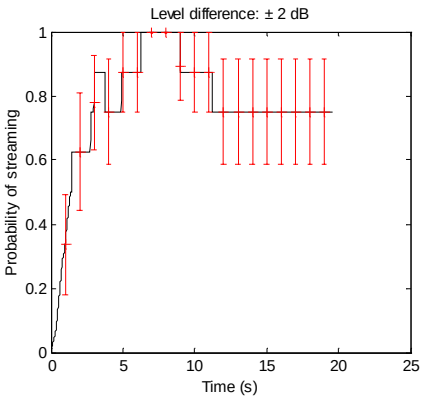


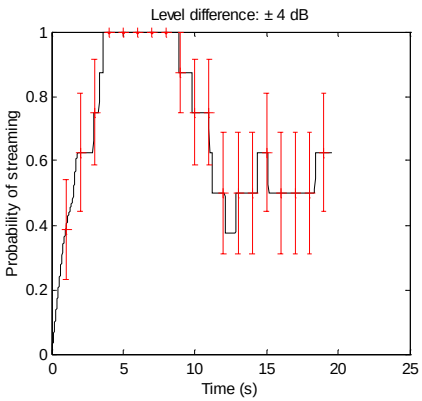


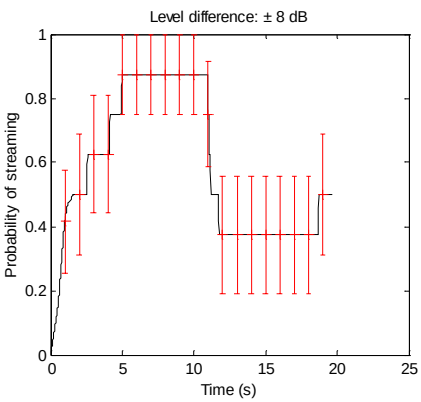
B.5 Subject SM

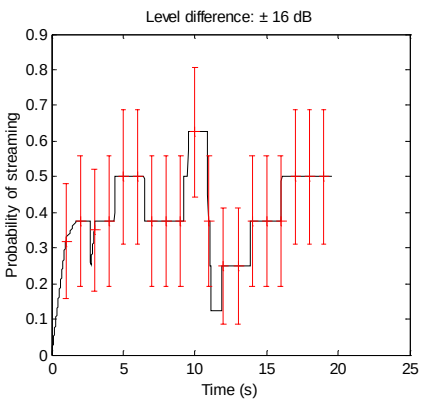




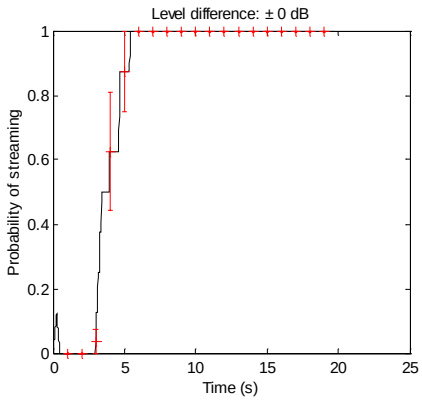


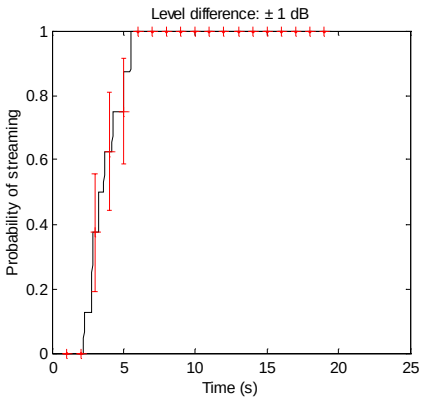


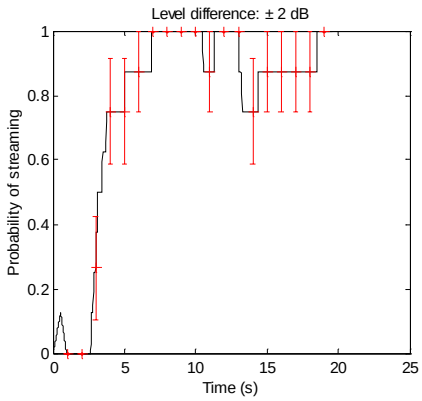


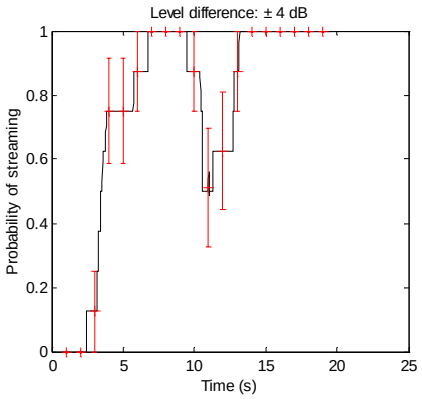


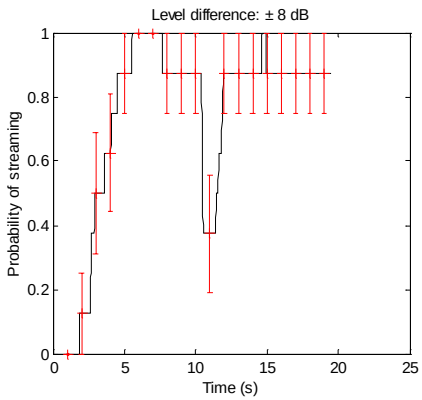
B.6 Subject ST

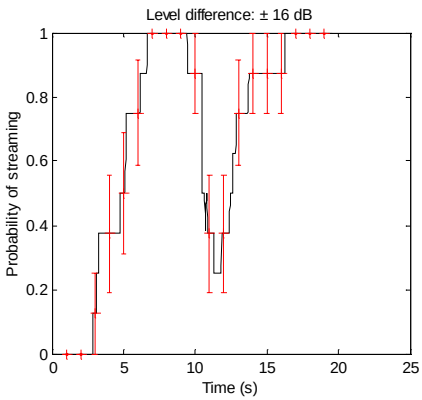












B.7 All subjects