Synfire chains as a neural mechanism for auditory grouping

Technical Report CS-99-11

November 1999

Stuart N Wrigley
s.wrigley@dcs.shef.ac.uk

Supervisor: Dr Guy J Brown
g.brown@dcs.shef.ac.uk

Speech and Hearing Research Group,
Department of Computer Science,
University of Sheffield

Contents

CHAPTER 1 *Introduction*

In typical situations, a mixture of sounds reach the ears. For example, a party with multiple concurrent conversations in the listener's vicinity, a musical recording or simply walking along a busy road. Despite this, the human listener can attend to a particular voice or instrument, implying they can separate the complex mixture.

Bregman (1990) has convincingly argued 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.

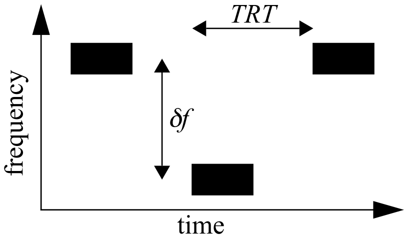
The perceptual grouping of sensory elements into streams can occur by two methods: *primitive grouping* and *schema-driven grouping*. Primitive grouping is data-driven whereas schema-driven grouping employs knowledge acquired through experience of varied acoustic environments. Bregman explains primitive grouping in terms of Gestalt principles of perceptual organisation (e.g. Koffka, 1936). 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*, determines likelihood of streaming. As presentation rate increases, tones of similar frequency group together.

Additional Gestalt grouping factors include *good continuation*: sounds which tend to change smoothly in frequency intensity and spatial location are likely to form a single stream; and *common fate* whereby elements which change in the same way at the same time tend to group together. Common fate properties include common onset/offset, common amplitude modulation (AM) and common frequency modulation (FM).

Attempts to create computer models that mimic auditory scene analysis has led to a new field of study known as computational auditory scene analysis (CASA). There has been work varying from the simple voice separation techniques of Denbigh and Zhao (1992) to the broader CASA research of Cooke (1993), Brown (1992) and Ellis (1996). However, such techniques are functional in approach: some form of time-frequency analysis generally followed by a high-level inference engine to group elements into perceptual streams.

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 1). Theories of perception are implied from experimental observations. Applying such mechanisms to figure 1, one can conclude that as δf decreases, it is more likely that the tones will be grouped together. Similarly, as *TRT* decreases, sequential tones will also be more likely to group.

Figure 1. Portion of a two tone streaming stimulus consisting of high-low-high pure tones.



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 primary objective of this study is to create a physiologically based account of auditory scene analysis. If such a model can be shown to produce data with a high correlation to psychoacoustic experiments, it would provide evidence that the model is indeed processing sound in a similar way to the human auditory system. In essence, the goal of this work is to generate insights into the nature of the auditory system and to improve the effectiveness of current CASA technology.

A long term objective of this field of study is to improve the performance of automatic speech recognition (ASR) systems. Most systems rely on the incoming speech having been pre-segregated or consisting of only one speaker. In a realistic environment, this is not possible and so the process requires automation. A successful computational auditory scene analysis implementation would produce a considerable improvement in current ASR technology.

Due to the scale of the problem, the work presented here will concentrate on modelling stream segregation by frequency proximity. The next section introduces some of the key terms associated with auditory scene analysis and will also discuss a number of contrasting approaches to producing a computational solution. A key stage of all computational models is the representation of the auditory periphery. Chapters 3 and 4 describe the physiology of the auditory periphery and the associated computational models. Chapter 5 describes two neuron models, one of which is used in the sniffier chain network described in chapter 6. Chapter 7 concludes the report.

CHAPTER 2

Literature Survey

Bregman's (1990) book Auditory Scene Analysis *drew together a wealth of perceptual information on how the auditory system is thought to separate multiple sounds into perceptual objects. During the past three decades, physiologists and computer modellers have sought to 'solve' the ASA problem using such information. Unfortunately, this task proved to be extremely difficult and the computer models produced have only had limited success. This chapter introduces the key elements of ASA and provides an overview of some of the proposed solutions.*

2.1. Auditory Scene Analysis

An understanding of the key principles involved in the processing of sound is required before the construction of a *computational* model of hearing. 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 2). 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 - for example in the perception of music. As mentioned in the introduction, the initial stage of auditory scene analysis

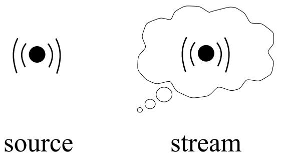


Figure 2. The relationship between a sound source and its mental perception - the stream.

is the decomposition of a sound into a collection of sensory elements - *segmentation*. The second stage of processing is stream formation and segregation. The mechanism by which these sensory elements are combined is termed grouping. *Primitive* grouping (bottom-up processing) encompasses the data-driven *simultaneous* and *sequential* perceptual organisations of sound. Simultaneous organisations correspond to grouping by sound source onset and offset. *Harmonicity* is also important in explaining how related sounds belong together, for example vocal tract sounds. In contrast, sequential organisations make use of continuity and proximity constraints across time.

Prior knowledge is also used to group sounds into streams. In the case of the cocktail party problem (Cherry, 1953) the listener has the task of attending to one conversation in the presence of many other voices and sounds. In this situation, grouping exploits *semantics* and *pragmatics*. The former allows the listener to analyse the sounds for meaning and direct her attention to the most interesting conversation. The practical knowledge of how language is used also enables a degree of prediction to aid the maintenance of the conversation stream. This use of experience and knowledge in the formation of streams is referred to as *schema-driven* grouping.

Both primitive and schema-driven grouping are concerned with combining individual sound elements into a perceptual stream. The issue of *how* grouping is implemented at the physiological level - the *binding problem* - has been the focus of much research by both physiologists and computer modellers.

2.2. Solutions to the binding problem

Even simple stimuli evoke highly fragmented and widely distributed responses in the auditory nervous system. Thus a particular acoustic stimulus will generate responses in a large number of spatially segregated neurons, each of which only encodes a small part of the acoustic object.

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 whole was explained by the supposed redundancy in neural circuits and averaging processes within the system. The advent of improved signal detection technology allowing physiologists to analyse the activity of single neurons dispelled this view. Neurons were no longer noisy indicators but the prime substrate of mental processes (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 increasing 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. 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.

Although conceived in the visual domain, such a theory can be applied to acoustic perception - with the same deficiencies. Singer (1993) discusses a selection of the limitations. First, cells at higher processing levels are often less selective than those at lower levels. Additionally, the upper levels of Barlow’s hierarchy correspond to particular 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. How would this cell indicate that it 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. Apart from cells that respond preferentially to faces, no other feature-specific cells have been found. Such hierarchies are unlikely to occur simply due to scale - it is not thought that there are enough neurons in the brain if all objects and all their possible views are to be each represented by one top-level neuron. Even if some more economical form of representation were to exist, no site has been found which is large enough to accommodate the ultimate site of convergence (see also Damasio, 1989). To exacerbate the problem, 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.

The alternative mechanism of grouping is based on the concept of an *assembly*: a large number of spatially distributed neurons. 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. This may be achieved by reciprocal connections between assembly members. Additionally, if the connections are dynamic, then the system can adapt its assembly structures and learn new objects.

2.2.1. Oscillatory solutions

It was proposed by von der Malsburg (1981; von der Malsburg and Schneider 1986; see also Milner, 1974) that the means of labelling different assemblies is by temporal synchronisation of the responses of assembly members. Their system used neural oscillations for expressing segregation. 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. 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.

Despite this success it was still of limited use. Their feature representations had no spectral relationship whereas stream segregation clearly depends relationships such as proximity in time and frequency - Gestalt grouping principles. A simple example of this relationship is *two tone streaming* (Bregman and Campbell, 1971; van Noorden, 1975). This demonstrates the trade-off between tone presentation rate and frequency separation. As presentation rate increases, the frequency difference between the tones required to generate two streams decreases.

The stream segregation occurring in figure 3 cannot be simulated by von der Malsburg and Schneider’s model.

Singer (1993) suggested that coherent oscillations in the visual cortex resulted from lateral connections within the cortex. Phillips and Singer (1997) re-iterated their belief in synchronisation as a neuro-physiological mechanism of grouping

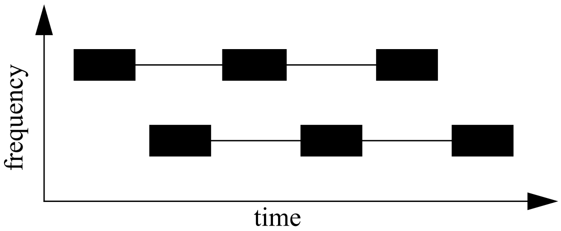


Figure 3. Spectrogram of six alternating tones. When stream segregation occurs, the high tone sequence and the low tone sequence form separate streams (indicated by the feint lines).

and also included the influence of contextual interaction. 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 network of relaxation oscillators with lateral excitation connections forming synchrony and a global inhibitor aiding desynchronisation. The global inhibitor receives excitation from each oscillator, and inhibits in turn each oscillator of the network. Once an 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. 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.

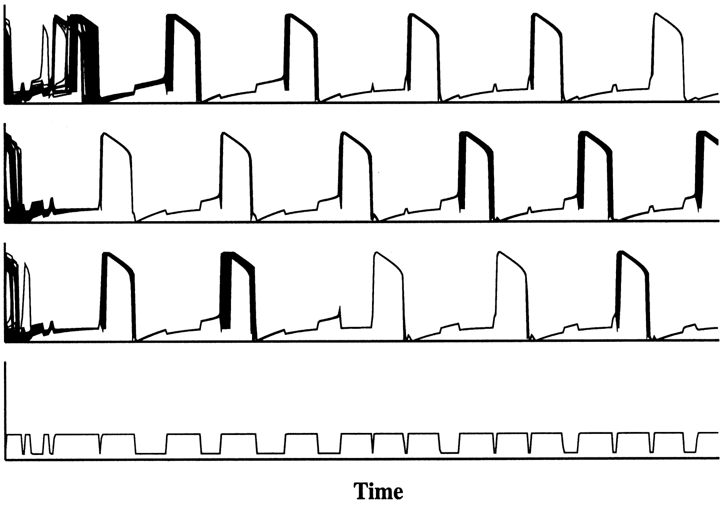


Figure 4. Temporal activities of the oscillator grid. The upper three traces show the combined temporal activities of the oscillator blocks representing the three streams. The bottom trace shows the temporal activity of the global inhibitor. Adapted from Wang (1996) Figure 5G.

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. Oscillators are connected by both permanent and dynamic weights. The permanent weighting between oscillators falls off exponentially with increasing distance. The dynamic weights change according to the degree of synchronisation in the network. 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.

Despite the dynamics being closely based on biological neurons and the network's ability to simulate streaming effects of repeated tones, Wang's oscillator model incorporates a number of unrealistic details. Most importantly is the use of a time-frequency grid on which to perform grouping. There is no physiological evidence for such extended delay lines; in fact they may be theoretically impossible. If Wang's topology is taken literally, the precise timing of responses required for grouping is unlikely to be preserved due to variability in synaptic processes (Abeles, 1991). However, the topology can also be considered to be an abstraction whereby each oscillator and each delay line represents a subnetwork such as a synfire chain (see chapter 6). In this case, loss of spike timing information would not occur. Additionally, at each time step, the continuous-time input is 'frozen' while the network oscillators achieve a stable state. This second time dimension (the oscillations in the segmentation process) exacerbates the time representation problem. Secondly, the input is sampled at 40ms intervals and at each time, the active oscillators are phase-randomised. In essence, the network produces a *snapshot* of the streams present at 40ms intervals. How such snapshots are integrated to give a time-varying estimate of stream content is not elaborated. Finally, Wang's model originates from his work in the field of *visual* object segregation. The usefulness of this analogue is dubious. In the visual domain, the temporal dimension can be regarded as separate from the spatial dimension. However, it is unlikely that such separation is possible in the auditory domain.

Although dealing with vowel recognition, the recognition aspect of Lui et al's (1994) 3 layer model can be considered to be a form of schema-driven grouping. The first level of the system encodes peaks in the linear prediction coefficients (LPC) input: a peak is represented by a group of active oscillators. The intermediate layer encodes the 'template' peak structure for each of the selection of vowels to be recognised in a manner similar to that of Wang *et al.* (1990). This form of associative memory consists of mutually connected oscillators with the coupling strengths determining the exact pattern to be represented. The use of reciprocal connections between the first 2 layers results in synchronised oscillations. The final layer then analyses this activity to produce a vowel category.

Desynchronisation is caused by inhibitory connections between next nearest neighbours in the intermediate layer.

The grouping mechanisms employed by Wang and Lui *et al.* have used lateral connections over a limited distance. This is useful for proximity grouping in the visual domain. It is less important in the auditory domain; in fact it is essential that features widely distributed across frequency can be grouped. In contrast to this approach, Brown and Cooke (1995) use global connectivity such as that used by von der Malsburg and Schneider. However, it does not produce oscillations by excitatory and inhibitory mechanisms as the above models do. The neural network model uses chaotic oscillators allowing a large number of groups to be represented. Unfortunately, the close match to human performance to two-tone streaming is overshadowed by the expensive cross-correlation process required to evaluate network synchronisation. In contrast, the model of Wang and colleagues requires only the application of a simple threshold.

In contrast to the above solutions, Baird (1996) implements a theory of attention and grouping based on adaptive synchronisation of 30-80 Hz oscillations. Rhythmic attention in audition (Jones, 1976) is modelled by coupled subsets of oscillatory associative memories analysing rhythmic frequencies of between 0.5 Hz and 10 Hz. Their activity, which is in the range of 30-80 Hz, is then integrated into the primary stream forming model. This model is a fast learning rule which reduces the coupling between frequency channels that do not exhibit the same activity at the same time. This reduction in coupling therefore reduces the synchrony between non-related channels and hence segregating channels which do not exhibit Gestalt *common fate*. Coupling gradually recovers between onsets, the rate of which can be adjusted to yield a qualitative match to van Noorden’s (1975) two tone streaming data. Close frequency channels tend to excite each other’s channel filters and so after stimulation of a particular channel subsequent stimuli of a non-rhythmic nature is captured due to the coupling change. However, with rhythmic stimuli, the expectancy system becomes an additional streaming factor. The oscillatory associative memories form a background (default) and a foreground stream. Suggested oscillatory frequencies are 35 Hz and 40 Hz respectively. Input conforming to the expected rhythm is synchronised with the attentional oscillators. However, the occurrence of a rhythmic mismatch causes the deviant activity to be boosted above the background frequency and is forced to synchronise with the attention stream thus modelling stimulus-driven attentional *pop out*.

2.2.2. Evidence for oscillatory-based feature binding

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. Further work using the electroencephalogram (EEG) has 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 wide range of animals in which 40 Hz activity has been observed suggests that it is fundamental to neural processing.

2.2.3. Other solutions

In parallel to the development of oscillatory solutions, work has been conducted using a much more functional approach. 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 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 Beauvois and Meddis'. The model consists of two interacting arrays of neurons: a foreground array and a background array. 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 included in the model architecture, 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. In line with the work of Wang and Brown, McCabe and Denham finally suggest that the time constants required to simulate human perception were of a magnitude more consistent with cortical-based processing rather than peripheral-based as argued by Beauvois and Meddis (1991).

An alternative approach to explain the two tone streaming phenomena is demonstrated by Todd (1996). His physiologically-motivated model computes an amplitude modulation (AM) spectrum at each tonotopic frequency. From these, a cross-correlation matrix is calculated in which neighbourhoods of high correlation indicate temporal coherence. When streaming occurs distinct areas of low correlation are present. Frequency proximity grouping is simulated for stimuli which are sufficiently close in frequency have similar temporal characteristics. The mechanism can also account for temporal proximity grouping due to the interaction of AM harmonics. At low repetition rates the AM fundamental(s) may not be represented. In this case the cross-correlation process relies on the fundamental's harmonics, some of which may coincide, thus increasing the cross-correlation measure. At higher repetition rates, the repetition frequency and its harmonics are well separated which produces a lower cross-correlation measure.

2.3. Summary and discussion

It should be emphasised that while oscillatory activity and synchronisation often occur together, they do not depend on one another. Individual neurons can engage in oscillatory activity whilst not synchronised with other cells and similarly, cells can exhibit synchronisation without the presence of oscillations. Consider oscillatory activity favouring synchrony. The occurrence of an activity burst during oscillation predicts, with some degree of confidence, the occurrence of a subsequent activity burst. It is this predictability which is needed to synchronise spatially distant cell clusters with zero phase lag, despite the considerable delays in the coupling connections. Hence, oscillations may not carry stimulus information but be instrumental in the establishment of synchrony over large distances. Alternatively, oscillatory activity may simply be an emergent property of synchrony. An assembly of interconnected cells firing in synchrony will produce a burst of activity followed by a pause (due to cell refractoriness) followed by another burst. This burst-pause process is likely to be repeated a number of times, thus generating oscillations. Additionally, Abeles *et al.* (1994) have shown that synchronous transmission in synfire chains (Abeles, 1991) can generate oscillatory activity due to the interaction of excitatory and inhibitory feedback and not simply due to periodic cell activation (see later).

The existence of oscillations has been claimed to arise purely as a by-product of the experimental procedures and not from feature binding (Horikawa *et al.*, 1994). Many studies use anaesthetics which are known to stimulate rhythmic neural activity. However, it is unlikely that oscillations do not occur as a result of binding as such oscillations have also been recorded from awake animals (Singer, 1993).

The presence or absence of oscillatory activity neither proves nor disproves the presence of synchrony between spatially distant cells. Hence study of oscillations alone cannot elucidate the temporal code. Synchrony and its dependence on the stimulus must be used which can only be accurately assessed from simultaneous recording of multiple cells. Oscillations are a useful indicator of organised activity and can guide the search for synchronisation.

The simplest temporal code - synchronous firing - plays an important role in all of the models described in section 2.2.1. In an alternative temporal code, Hopfield (1995) proposes that the relative timing of spikes between cortical neurons can convey important information about sensory cues. The model neurons exhibit an oscillatory subthreshold variation of membrane potential. In the absence of input, no action potentials occur due to the subthreshold nature of the oscillation. When the combined input current and membrane potential exceed threshold, an action potential is elicited. The relative timing of the action potential relative to the oscillatory maximum is determined by the input current strength.

Hopfield suggests that if the logarithm of a sensory cue’s strength in encoded by some relative time advance in firing, then this information can be transferred quickly, in a scale-invariant form. There is currently very little data to support this temporal code and hence its applicability to the binding problem is yet to be seen.

The majority of models discussed here simulate a limited set of stimulus configurations - groups are formed on the basis of frequency and time proximity. A danger of this is that the models become overly adapted to solving one particular problem and cannot be extended to incorporate new features. Although it is currently highly unlikely that a single solution can explain all grouping cues, consideration must be paid to the extendability of a model. Ideally, models should simulate grouping by common amplitude modulation, common onset and offset, harmonicity, spatial location and timbre in addition to temporal and frequency proximity. Indeed, Bregman (1997) has commented, "We have so far concerned ourselves with models that attempt to solve the ASA problem directly. There is, however, another approach: trying to model the data that comes out of the perception laboratory. This is a dangerous mission and again requires a wide knowledge of ASA phenomenon. Without it, a researcher may invest a lot of effort to develop a model that offers a parsimonious account of a very limited subset of laboratory phenomenon. Consequently, while the model may be very parsimonious in accounting for a few perceptual effects, it may turn out to be so specific to that small set of phenomena that it is helpless when a wider range of laboratory effects has to be explained. Again, an early stage in the development of a model of this type should be to ask whether it is too narrowly focused."

A further inadequacy of current models is their representation of time. The models of von der Malsburg and Schneider (1986) and Liu *et al.* (1994) both use spectral inputs but do not allow responses at different times to be compared. As noted above, Wang's (1996) model fails to represent time in a physiologically plausible manner. In fact, his use of a pseudo-spectrogram with the time axis represented by delay lines may even be theoretically impossible. On a related issue, the manner in which parts of the pseudo-spectrogram are 'connected' is physiologically implausible. In Wang's model, input to an oscillator from another oscillator, no matter how distant in frequency or time, occurs instantaneously. However, time delays are inevitable in neuronal signal transmission. In an attempt to remedy this, Campbell and Wang (1996) included time delays in the inter-oscillator connections. Although this impaired the ability of the network to produce *perfect* synchronicities, it was still able to form synchronous groups.

Furthermore, Wang's model rapidly forms streams within *n* cycles for a stimulus containing *n* streams. Although such efficient synchronisation may be important for engineering applications, this is contrary to psychophysical evidence that stream segregation can take up to many seconds to appear (Anstis and Saida 1985). Other models (Brown and Cooke, 1997; Beauvois and Meddis, 1991, 1996; McCabe and Denham, 1997) successfully simulate the build up of streaming over time.

The segregation decision at a particular time instant should be based not only on the auditory information at that time but also the segregation decisions made in the recent past. To achieve this, a form of short-term memory is required. Horn and Usher (1992) present a model in which potentiation is used to sustain oscillations after the input is turned off. In this framework, the oscillator's threshold rises normally due to accommodation. However, when the stimulus ceases, the cell threshold decreases and falls below its resting level (potentiation). This causes oscillations to persist without external cell activation. Horn and Usher suggest that their model is also simulates the experimentally observed limited capacity of short term memory (7±2). Lisman and Idiart (1995) also show the 7±2 capacity of short term memory using nested oscillations similar to those recorded in the brain. Each memory is stored in a 40 Hz subcycle of a low frequency (5 - 12 Hz) oscillation.

Almost all input to the cortex passes through the thalamus. Crick (1984; Crick and Koch, 1990) has suggested that part of the thalamus (the thalamic reticular complex) may be involved in selective attention. The attentional *searchlight* is produced by rapid bursts of firing. When this activity synchronises with a group of neurons, that group becomes the attentional foreground and the remainder become the background. Although many researchers acknowledge the importance of an attentional searchlight, few have actually implemented one. For example, McCabe and Denham (1997) incorporate an attentive input into their model but concede that it is ‘not generally used’ and simply offers a way in which higher cognitive processes can influence the data-driven streaming process. Similarly, the attentional searchlight formed a component of the Brown and Cooke (1997) model but was not implemented in the computational simulation.

In summary, there are three areas which need to be addressed before satisfactory models of feature binding can be constructed. Firstly, and possibly most importantly, is the issue of time representation. A physiologically based representation is needed to allow comparisons of auditory activity at different times. Related to this is the role of short term memory. Segregation should be based on a contextual decision rather than being independently made thus allowing, for example, binding by temporal proximity. Finally, the role of attention and schema driven grouping has been the subject of little work by modellers.

CHAPTER 3

Auditory Periphery

Before a computational model of acoustic feature binding such as those discussed in the previous chapter can be produced, a detailed understanding of the physiology of the auditory periphery is essential. For example, it has been argued that when we listen to a complex tone, it is easier to 'hear out' the fundamental and lower partials than it is to hear out higher partials (Plomp, 1964). This is explained by the fact that harmonics are linearly spaced in frequency whereas the mapping of frequency onto the basilar membrane is logarithmic. Hence, lower harmonics are spaced further apart and so have a higher perceptual 'resolution'. In terms of auditory grouping, this phenomenon is seen in the relative ease by which a preceding tone can capture lower harmonics of a complex in comparison to higher harmonics.

3.1. The auditory periphery

The auditory periphery, which extends as far as the auditory nerve, can be divided into three compartments: the external, middle, and inner ear (figure 5). Brief descriptions of these structures follow. However, more detailed treatment can be found in Pickles (1988).

3.1.1. The External Ear

The external ear comprises the pinna and the external auditory meatus (duct or canal, some 2.7cm long). Sound waves are funnelled by the pinna into the meatus to impinge on the elastic tympanic membrane that separates the external and middle ear compartments. The tympanic membrane is vibratile and held under tension. The effect of the outer ear on the incoming sound has been analysed from two approaches. One is the property of pressure gain and the other is sound localisation.

3.1.1.1. Pressure gain

The external auditory meatus acts as a resonator (similar to an organ pipe) with a resonance of approximately 2 to 7kHz. The resonant frequency of an oscillating system is that frequency at which a minimum energy input is required to maintain the oscillation, i.e. the system is maximally sensitive at that frequency. This enhancement property of the external auditory meatus serves to ensure reliable transmission of the major sound frequency components of normal speech.

The convolutions and cavities of the pinna, concha and meatus combine to increase the sound pressure of some frequencies and decrease the sound pressure of other frequencies at the tympanic membrane. Figure 6 shows the average pressure gain (in decibels) in man provided by the outer ear over a range of frequencies. The functions in figure 6 are called transfer functions.

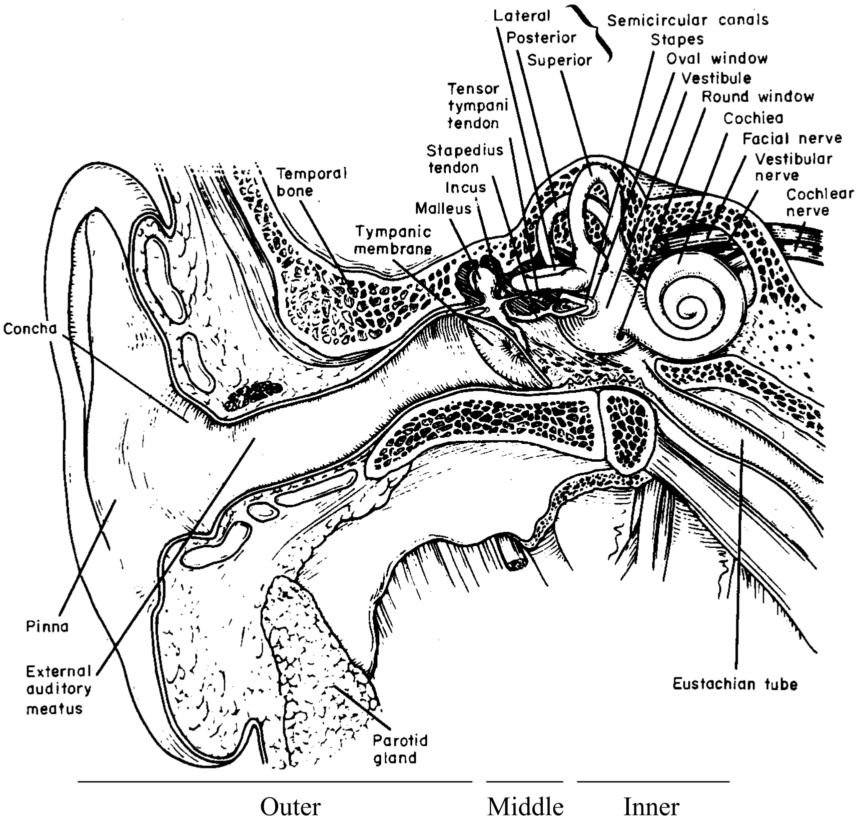


Figure 5. Anatomy of the ear showing the three compartments. From Pickles (1988).

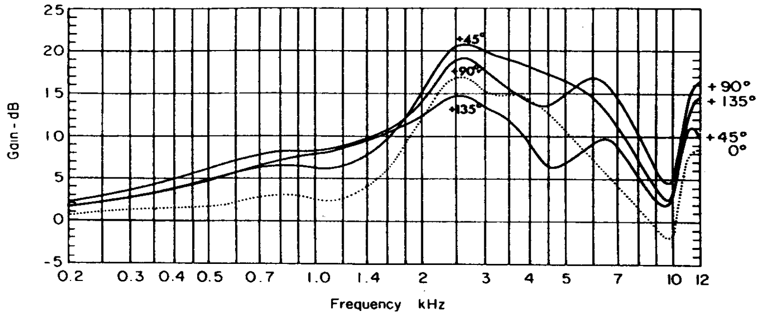


Figure 6. Average pressure gain of the ear in man. Pressure gain at the eardrum with reference to free field is plotted as a function of frequency. Zero degrees is straight ahead and positive angles are ipsilateral to the ear. From Pickles (1988).

3.1.2. The Middle Ear

The middle ear space is a gas pocket, closed to the outside world except for the Eustachian tube which opens into the pharynx behind and to one side of the tongue. Normally this tube is closed, which prevents one from being 'deafened' by the sound of one's own breathing and voice. This tube opens intermittently (for example, during yawning) to allow pressure equilibration between the external and middle ear environments.

Mechanical impedance can be defined as the total resistance of an object or substance to movement. The middle ear acts as an impedance matching, or energy-coupling, device. Its purpose is to transfer, without significant loss, sound vibrations in the air (tympanic membrane) to vibrations in the much denser, liquid medium of the inner ear. This is accomplished via a chain of three ossicles (bones) which are interposed between the tympanic membrane and the membrane of the oval window: namely the malleus (hammer), incus (anvil) and stapes (stirrup).

The first two ossicles are joined relatively rigidly so that when the tympanic membrane is deflected, the force is transferred to the stapes. The stapes is attached to the oval window which is a flexible membrane in the wall of the cochlea.

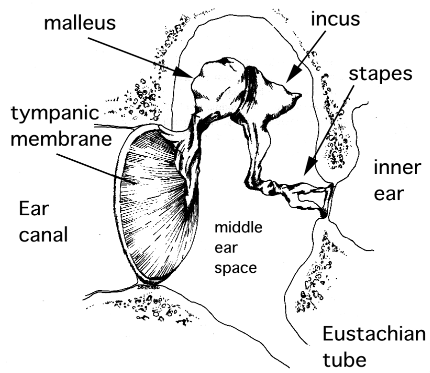


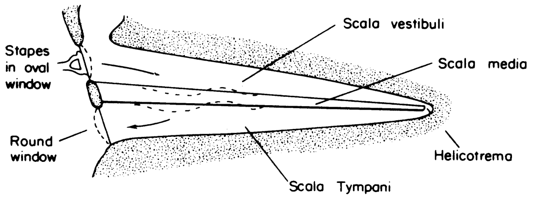
Figure 7. The ossicles which are interposed between the tympanic membrane and the membrane of the oval window. From WWW (1999).

This ossicular chain amplifies the sound pressure it conveys by two means:
1. By a mechanical lever arm action;
2. By pressure amplification: the force at the tympanic membrane is transferred to the much smaller oval window.

The total pressure gain in the middle ear is approximately 22. This ensures efficient transfer of sound energy from the air to the much denser (and therefore more resistant) liquid medium in the inner ear.

The three middle ear ossicles form a vibrating system, having elastic and inertial components. Consequently, they have (as a vibrating system), a resonant or natural frequency. For the ossicles this frequency range is about 500 to 2,000 Hz. Thus the combined resonant frequencies of the external ear (2,000 - 5,000) and the middle ear (500 - 2,000), largely explain the high sensitivity of the average ear between 500 to 5,000 Hz. It should be noted that there are two small muscles in the inner ear (the tensor tympani and the stapedius) that are reflexly activated (contracted) by very loud sounds (greater than 80dB above threshold), which function to reduce the amplification generated via this system and prevent the inner ear structures from being over loaded.

Figure 8. The unwound cochlea. From Pickles (1988).



3.1.3. The Inner Ear

The inner ear, or *cochlea*, is a coiled passage in the temporal bone of the head (it is shown uncoiled in figure 8). Structurally the cochlea is subdivided into three components or ducts (the Scala Vestibuli, the Scala Media, and the Scala Tympani) separated by two membranes (Reisner's membrane and the *basilar membrane* respectively). The Scala Vestibuli and Scala Tympani both contain perilymph which is similar in composition to extracellular fluid; while the Scala Media contains endolymph which is similar to intracellular fluid.

The cochlea contains the structures which translate sound vibrations into electrical neural signals. This mechanism is found in the organ of Corti which is located on top of the basilar membrane within the Scala Media. At the end of the cochlea, closest to the middle ear cavity, the basilar membrane is relatively stiff and narrow. The membrane becomes more elastic and wider as it extends throughout the cochlea towards the apex.

When pressure waves push on the tympanic membrane, the chain of ossicles, in turn, push the stapes against the oval window membrane. Next, the pressure on the oval window produces a wave of pressure in the liquid filled inner scala vestibuli. Most of this pressure wave is transmitted to the elastic basilar membrane. Since the fluids of the inner ear are incompressible, the pressure variations set up at the oval window will be further transmitted to the round window membrane which acts as a pressure release valve.

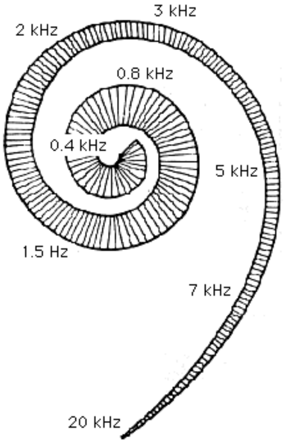


Figure 9. Frequency place coding on the Basilar Membrane.

The stiff portion of the membrane closest to the middle ear cavity (base) vibrates immediately in response to pressure changes transmitted to the oval window. The vibrations from the base then travel along the basilar membrane toward its apex (the wide end) - a *travelling wave* is formed. However, the position of maximal displacement of the travelling wave varies with sound frequency. The properties of the membrane nearest the oval window (base) are such that it resonates optimally (under goes the largest deformation) with high frequency tones; the more distant (wider) regions of the membrane (near the apex) vibrate maximally in response to low frequency sounds. Thus, the frequencies of incoming sound waves are 'sorted' along the basilar membrane: each frequency has its characteristic place (figure 9). Note, however that very low frequencies (less than 200Hz) are compressed on to a relatively limited section at the apical end of the membrane.

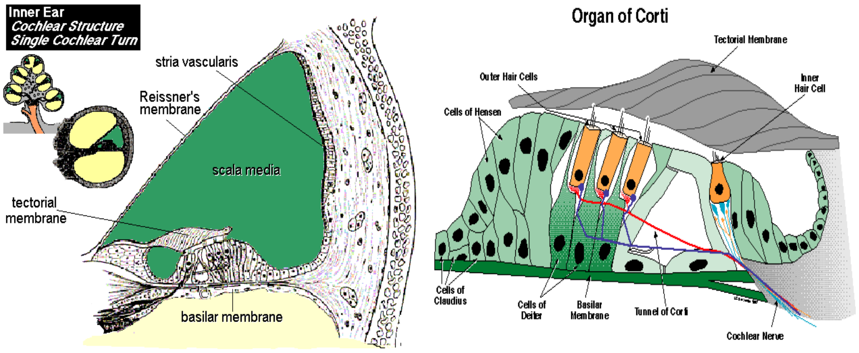


Figure 10. Cross-section through a cochlear tube showing the Basilar Membrane (left) and Organ of Corti (right).

The organ of Corti, which contains the ciliated receptor cells, extends from the base to the apex of the cochlea. The base of each hair cell is attached to the flexible basilar membrane, while its cilia are firmly attached at the ends to the tectorial membrane (a structure which forms a roof over the basilar membrane). The groups of hair cells are arranged in rows of *inner* and *outer* hair cells the functional significance of which will be further discussed below.

The actual transduction process (change from mechanical to electrical energy) at the receptor cell level is well understood. Where the displacement of the basilar membrane is a maximum, the stimulation of the receptors (hair cells) which sit upon the membrane is greatest. The mechanism for this is shown in figure 10, which represents a cross-section through the cochlear tube. As described above, the base of each hair cell is attached to the flexible basilar membrane, while its cilia are firmly attached at the ends to the rigid tectorial membrane. Consequently, when a given section of the basilar membrane is displaced by sound waves, this arrangement imposes a shearing (or bending) force on the cilia, which in turn, causes a receptor potential in the cells. This mechanism is extremely efficient, since each individual hair cell itself is also tuned to generate its maximum receptor potential in response to a shearing force occurring at the frequency which corresponds to its position on the basilar membrane.

The outer and inner hair cells also perform different functions. Outer hair cells are *active*. When the basilar membrane vibrates, outer hair cell stereocilia are deflected causing K^+ ions to move into the cells. This causes the outer hair cells to contract and lengthen as the basilar membrane vibrates which feeds extra movement into the basilar membrane, making the vibrations bigger - a *positive feedback loop* (figure 11).

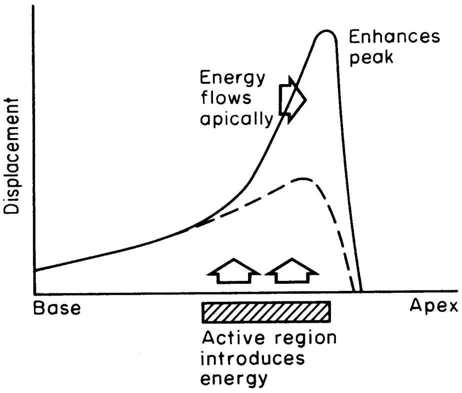


Figure 11. Basilar membrane response enhancement by outer hair cells. From Pickles (1988).

The important consequence of this enhancement is that without outer hair cells, the auditory system would be about 40 dB less sensitive to sounds. Outer hair cells also sharpen frequency selectivity because they increase basilar membrane vibrations.

Inner hair cells are, on the other hand, *passive*. As in the outer hair cells, when the basilar membrane vibrates, inner hair cell stereocilia are deflected causing K^+ ions move into the cells. This causes the release a neurotransmitter onto the auditory nerve fibres at their base which stimulates the nerve fibres and causes action potentials. The increased vibration of the basilar membrane produced by the outer hair cells, results in the inner hair cells moving more. The inner hair cells passively turn the vibrations of the basilar membrane into action potentials.

In the transduction process, the louder the sound, the greater the amplitude of basilar membrane vibration at a given location, the larger the bending of the cilia, the greater the receptor potential, the more transmitter release, and the higher the action potential frequency in the sensory nerve fibres.

Inner hair cells may provide sharp tonotopic pitch discrimination, while outer hair cells (many of which are converging from a large area of basilar membrane upon a single afferent fibre) may provide more broadly tuned auditory sensations. In addition, hair cells receive (centrally originating) efferent innervation which may reduce or suppress hair cell excitation.

3.1.4. The Auditory Nerve

In man, approximately 30,000 nerve fibres, whose cell bodies are contained within the Spiral Ganglion, form the direct connection between the cochlea and the cochlear nucleus. 95% of these fibres are directed to the inner hair cells and only 5% receive information from the outer hair cells.

As is shown in figure 12, the auditory nerve fibres connect to the inner hair cell closest to the fibre’s point of entry to the cochlea. In contrast, the fibres connecting to outer hair cells travel basally before terminating. Each inner hair cell fibre connects to one and only one cell whereas the fibres connecting to outer hair cells branch and connect to up to ten cells. About 20 fibres connect to each inner hair cell but only 6 connect to outer hair cells.

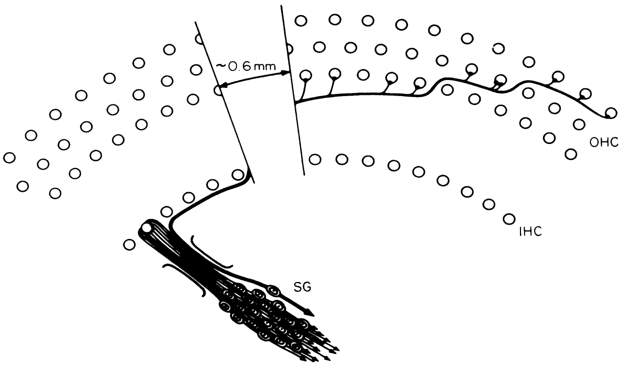


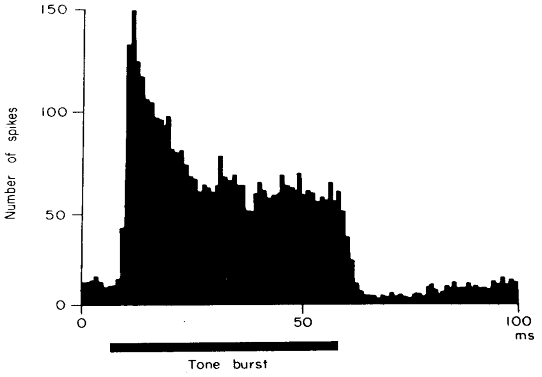
Figure 12. The majority (95%) of auditory nerve fibres connect with the inner hair cells. From Pickles (1988).

In the absence of stimulation, nerve fibres discharge at their *spontaneous rate*. When stimulated, the nerve fibres continue to fire at their spontaneous rate unless the stimulus intensity exceeds the nerve fibre's firing threshold. Above threshold, the firing rate increases almost linearly with intensity until such a level is reached that the nerve fibre does not increase its rate of firing when the stimulus intensity is increased. The nerve fibre is said to be *saturated*.

3.1.4.1. Frequency selectivity

Fibres are responsive to single tones and, if presented in isolation, the tones are always excitatory. The standard way of showing these responses is the post-stimulus time histogram (PSTH). This is built up by presenting the stimulus many times and for each action potential that occurs, incrementing the count for the bin corresponding to the time after the beginning of the stimulus. A tone burst causes a sharp onset response which rapidly decays over the first 10-20ms and then more slowly to a steady state over a period of 20-100ms. This property is known as *adaptation* and can be seen in figure 13. At stimulus offset, firing activity falls below the spontaneous firing rate. After a brief recovery period, the firing rate then returns to the spontaneous rate.

Figure 13. Response of an auditory nerve fibre to the presentation of a tone burst. The initial sharp onset and subsequent decay is clear. From Pickles (1988).



The nerve fibres can also be characterised by their firing threshold with respect to frequency. The intensity of the tone burst is adjusted until an increase above the spontaneous firing rate is just detectable. This is then repeated for a range of stimulus frequencies until a *tuning curve* is built up (figure 14). These curves exhibit a low threshold at a specific frequency - the fibre is highly sensitive at this specific frequency. This frequency is called the fibre's *best* or *characteristic frequency* (BF or CF).

The variation in curve shape is also visible in figure 14. Low frequency fibres are broadly symmetrical but at higher frequencies the curves become asymmetric with a sharp characteristic frequency trough and a long tail extending to the lower frequencies (The slight increase in sensitivity in this tail at approximately 1kHz is due to the power enhancement of the middle ear). A single auditory nerve fibre, therefore, behaves as a non-linear asymmetric bandpass filter. The frequency selectivity of the fibres is almost certainly derived from the basilar membrane and hair cell frequency selectivity.

Work by Liberman (1982) has shown that the population of auditory nerve fibres can be split into three broad groups based upon their spontaneous firing rate and also their associated firing threshold. Fibres with high spontaneous rates (greater than 18 spikes per second) have low thresholds and fibres with low spontaneous (less than 0.5 spikes per second) rates have high thresholds. Fibres with intermediate spontaneous firing rates have intermediate threshold levels. This distribution can be seen in figure 15.

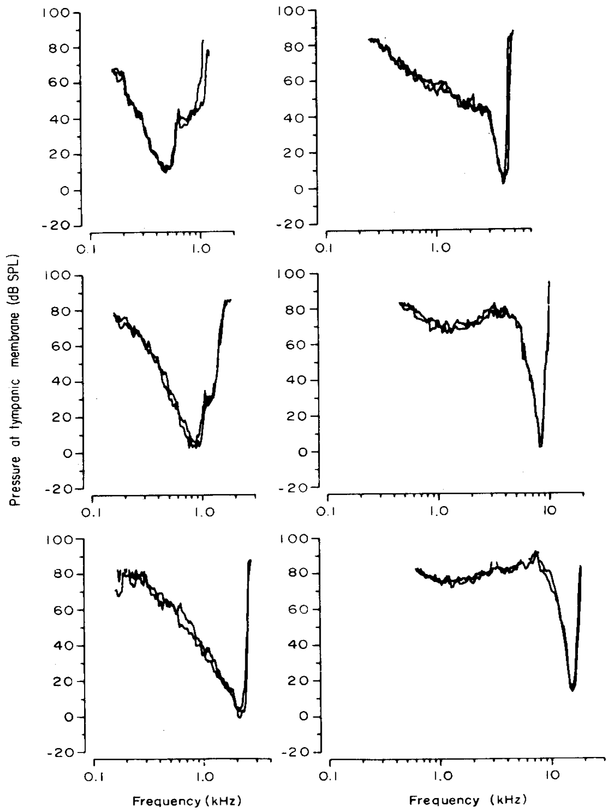


Figure 14. Tuning curves for six different frequency ranges. In each plot, the responses from two fibres of similar characteristic frequency and threshold are shown. From Pickles (1988).

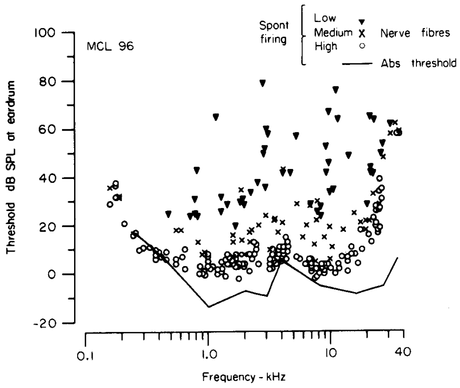


Figure 15. Distribution of low, intermediate and high spontaneous firing rate fibres and their associated firing threshold values. From Pickles (1988).

An alternative method of measuring firing rate as a function of intensity which produces rate-intensity functions (figure 16). The functions are sigmoidal in shape and have a dynamic range of 20 - 50dB.

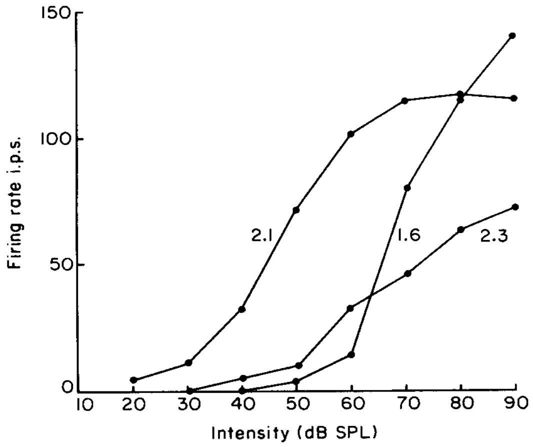


Figure 16. Rate-intensity functions for one auditory nerve fibre (CF 2.1kHz) at different frequencies. From Pickles (1988).

Rate-intensity curves display the firing rate - intensity combinations for constant frequency. Similarly, the combination of intensity and frequency can be displayed for constant firing rates. These iso-rate tuning curves (figure 17) show that the frequency selectivity of the fibre generally improves as a higher firing rate (and hence higher intensity) is used, although it later deteriorates as the fibre saturates.

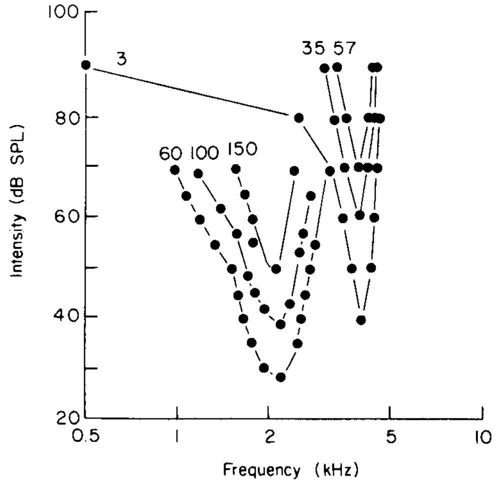


Figure 17. Tuning curves constructed at different firing rates for two fibres of differing centre frequency. From Pickles (1988).

3.1.4.2. Phase locking

At frequencies above 5kHz, the auditory nerve fibre fires with equal probability in every part of the stimulating waveform cycle. Below this frequency, the firing of the nerve fibre is locked to a particular phase of the stimulating waveform. Although the fibre may not fire every period, when it does fire, it will do so only in one phase of the stimulus. This characteristic occurs because the inner hair cells only initiate nerve firings during the upward deflection of the basilar membrane.

Phase locking can be shown using a period histogram. It is created by plotting the occurrence in time of every auditory nerve spike but resetting the time axis every period. It is evident from such period histograms that the response of the fibre is a half-wave rectified version of the stimulating waveform. Figure 18 demonstrates the half-wave rectification property and its preservation as intensity increases.

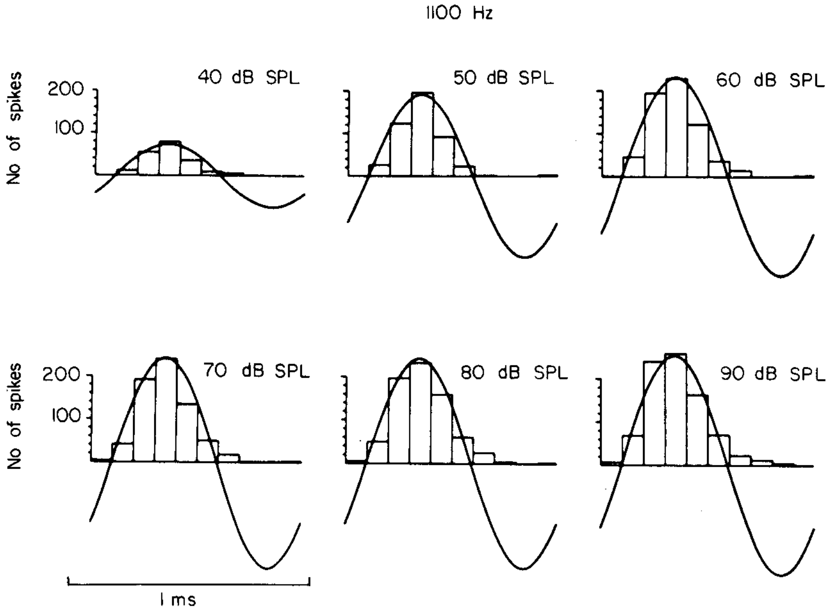


Figure 18. Phase locking preservation at increasing intensities. Note that the firing is saturated above 70dB but the phase locking remains unaffected. From Pickles (1988).

The loss of phase locking as the stimulating waveform frequency approaches 5kHz can be seen in figure 19.

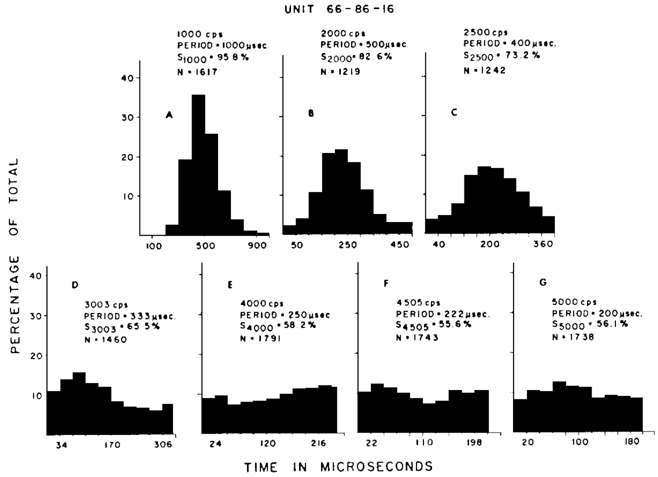


Figure 19. Period histograms demonstrating the loss of phase locking as the stimulating frequency approaches 5kHz. From Rose *et al.* (1967).

The distribution of intervals between successive auditory nerve events - an interspike interval histogram - is sharply polymodal. As can be seen from figure 20, the peak of each partial distribution is very close to an integer multiple of the waveform period and the population is always larger than the following one. This serves to reinforce the fact the fibre will not fire every period but when it does fire, it will do so only in one phase of the stimulus.

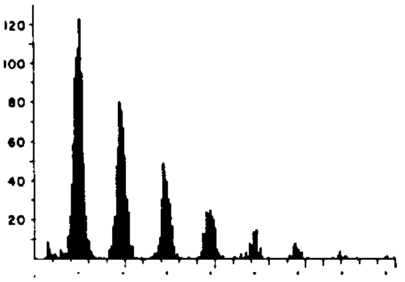


Figure 20. Interspike interval histogram. Dots below the abscissa indicate integral values of the stimulating tone period. Adapted from Rose *et al.* (1967) fig 1.

3.2. Summary

The auditory periphery consists of three main areas: the outer, middle and inner ears. Sound travels down the auditory canal and causes the tympanic membrane to vibrate. The ossicles then transfer this energy to the cochlea. The combination of the outer and middle ear resonances explain the increased hearing sensitivity in the 500 - 5000 Hz range. Sound energy at different frequencies is converted to mechanical motion of the basilar membrane which in turn stimulates the activity of hair cells in contact with the membrane. This activity is transmitted to the brain via the spiral ganglion. The next chapter will introduce a number of computational models which simulate this process.

CHAPTER 4

Auditory Periphery Model

As discussed in the previous chapter, a detailed understanding of the auditory periphery can allow a number of perceptual phenomenon to be explained. Similarly, if computational models are to explain as wide a range of perceptual experiences as possible, they must incorporate an accurate simulation of this peripheral processing. This chapter describes a number computational solutions which model the various stages of the auditory periphery.

4.1. Introduction

The auditory periphery can be divided up into four main functional areas for the purposes of computational modelling: outer and middle ear resonances, basilar membrane response, inner hair cell transduction and auditory nerve spike generation. The models presented here are existing models which are in close agreement with the experimental data.

4.2. Outer and middle ear resonances

The resonances of the outer and middle ear are essentially linear for low to medium intensity sounds and can be modelled using a simple high-pass filter of the form

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

where *x[t]* is the amplitude of the input at time *t*.

Another means of simulating the outer and middle ear resonances is to use a hearing threshold curve as a weighting function. Figure 21 shows the gain across frequency of a hearing threshold curve detailed by Fay (1988).

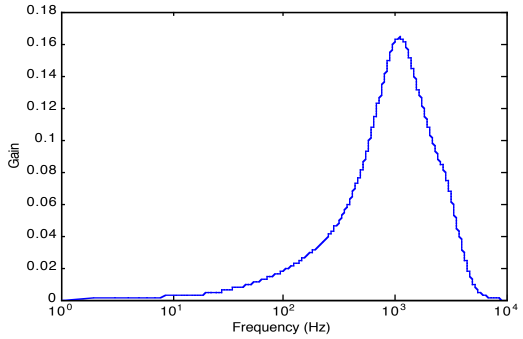


Figure 21. Gain across frequency of the Fay hearing threshold curve.

The work presented in this report uses a single frequency channel of the basilar membrane and so no relative tuning of the frequency channels is required.

4.3. Basilar membrane filtering

The frequency selectivity of the basilar membrane is modelled by a gammatone filterbank in which the output of each filter represents the frequency response of the membrane at a specific position. Any filter can be completely characterised by its response to a brief click - the impulse response. The filterbank is based on an analytical approximation to physiological measurements of auditory nerve impulse responses obtained by the *reverse correlation* technique of de Boer and de Jongh (1978). The gammatone filter of order *n* and centre frequency f_0 Hz is given by

$$gt[t] = t^{n-1}e^{-2\pi bt}\cos(2\pi f_0 t + \phi)u[t] \qquad (2)$$

where ϕ represents the phase, *b* is related to the bandwidth and *u[t]* is the unit step (Heaviside) function

$$u[t] = \begin{cases} 1 & t \ge 0 \\ 0 & t < 0 \end{cases} \qquad (3)$$

The name *gammatone* comes from the fact that the envelope of the filter impulse response (figure 22) is the statistical *gamma* function and the fine structure of the impulse response is a *tone* of frequency f_0 and phase ϕ.

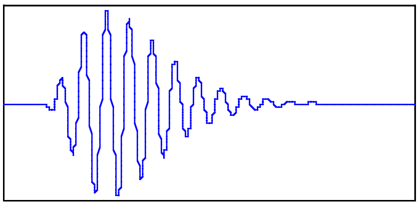


Figure 22. Impulse response of the gammatone filter.

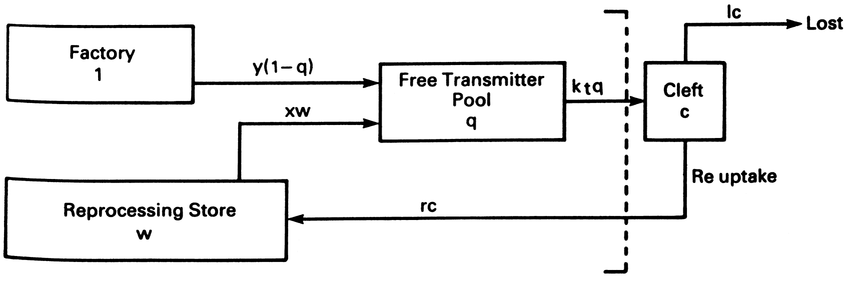
Although the gammatone filter is linear and cannot simulate any non-linearities and also has a symmetrical magnitude response, its amplitude characteristic exhibits a very good fit to the *roex(p)* function commonly used to represent the magnitude characteristic of the human auditory filter shapes (Patterson and Moore, 1986). This property justifies its use to model auditory frequency selectivity.

4.4. Inner hair cell transduction

Within the cochlea, the movement of the basilar membrane is converted in to electrical signals by the inner hair cells located within the organ of Corti. As noted above, this leads to properties such as phase locking, adaptation and saturation.

There has been extensive work conducted on creating a computational model that will explain the non-linearities that occur at the junction between hair cells and the auditory nerve (Meddis, 1988; Schroeder and Hall, 1974). The model used here is the multiple-reservoir scheme proposed by Meddis (1986, 1988). In a review of eight hair cell transduction models, Hewitt and Meddis (1991) concluded that their model exhibited the closest fit to physiological data and was also computationaly efficient. When presented with the response of the basilar membrane from the gammatone filter, the model returns the probability of a spike occurring in the auditory nerve.

The model can be understood in terms of the production, movement and dissipation of transmitter substance in the vicinity of the hair cell-auditory nerve fibre synapse (figure 23).



$$\frac{dq}{dt} = y(1-q(t)) + xw(t) - k(t)q(t)$$

$$\frac{dc}{dt} = k(t)q(t) - lc(t) - rc(t)$$

$$\frac{dw}{dt} = rc(t) - xw(t)$$

Figure 23. Flow diagram for transmitter substance and differential equations defining the model. From Meddis (1986) model B Fig 10.

The model parameters are based on those described in (Meddis, 1988) with only small number of changes to improve the model's match to experimental data (see table 1).

Table 1: Inner hair cell transduction model parameters.

Parameter	Meddis (1988) value	New value
A	5	2
B	300	300
g	1000	2000
y	11.11	8
l	1250	2500
r	16667	6580
x	250	66.31

4.5. Auditory nerve spike generation

The aim of this modelling work is to produce auditory nerve spikes for use in modelling higher level brain processes. This final stage of the periphery model converts the probabilistic output of the inner hair cell model into discharge times based upon a process proposed by Carney (1993). The spike generator is a Poisson process which takes as its input the Meddis model output and combines terms for both absolute and relative refractory periods. After an absolute refractory period of 0.75ms, the effect of the refractoriness, also called the *discharge-history* effect, gradually decays to zero over a period of approximately 40-50ms. The time course of the history effect is given by

$$H(t) = H_{max}(c_0 e^{-t-t_l-R_A/s_0} + c_1 e^{-t-t_l-R_A/s_1}) \qquad (4)$$

for $(t\text{-}t_l) \geq R_A$

$$H(t) = 0 \qquad (5)$$

for $(t\text{-}t_l) < R_A$

where $t\text{-}t_l$ is the time interval since the previous spike and H_{max} determines the maximum threshold increase due to a previous discharge.

The discharge history effect, *H*, is shown in figure 24.

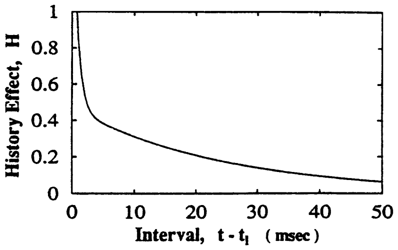


Figure 24. Discharge history effect showing absolute (0.75ms) and relative (40-50ms) refractory periods. From Carney (1993).

Given the discharge history effect, *H*, the instantaneous spiking rate of the AN fibre is modified from that of the Meddis model output (s_k) to

$$r_k = s_k - H(t) \qquad (6)$$

The firing decision is made by the firing probability $T(r_k)$, where *T* is the sampling period. For each sampling period, a random number q_k, uniformly distributed between 0 and 1, is produced by a standard random number generator. If $T(r_k) \geq q_k$, a spike is generated; otherwise no spike is generated. The spiking decision is used to update *H(t)* and the simulation for spike generation proceeds until the input stimulus to the model terminates.

The hair cell output shown in figure 25 (left panel) is the probability of a spike being generated. The spike generation process involves a certain degree of randomness and so the output in terms of auditory nerve discharge times will vary slightly for each presentation. Therefore, to obtain an accurate description of this response, a post-stimulus time histogram (PSTH) is produced from a number of presentations to the spike generator (see section 3.1.4.1). The PSTH for the tone used to produce the probabilistic hair cell response is also shown in figure 25 (right panel). The PSTH exhibits the fundamental characteristics of the experimental PSTH (figure 13): a sharp onset response which drops rapidly over the first 10-20 ms and then more slowly. The recovery period after tone offset is evident.

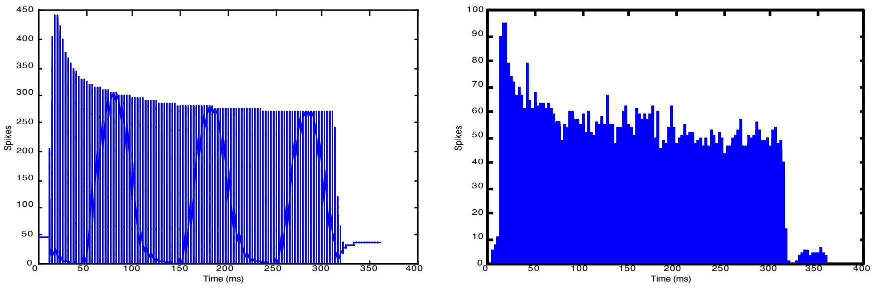


Figure 25. Probabilistic spike output (left) and post-stimulus time histogram (400 presentations) for generated spikes (right) of the inner hair cell transduction model in response to a 300ms 70dB 1kHz tone starting at 10ms.

The firing rate as a function of intensity can be used to show how the hair cell response varies with intensity. The rate-intensity curve of the model is shown in figure 26. The function has the expected sigmoidal shape (c.f. figure 16) and saturates at an acceptable 50dB above the threshold.

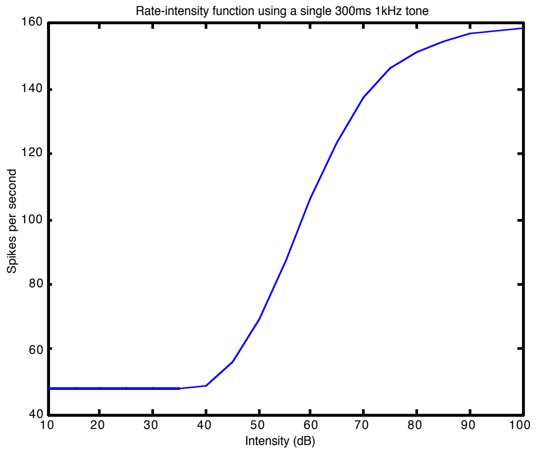


Figure 26. Rate-intensity function of the periphery model (c.f. figure 16).

Similarly, the combination of intensity and frequency can be displayed for constant firing rates (figure 27). These iso-rate tuning curves show that the frequency selectivity of the fibre generally improves as a higher firing rate is used. Although the basic shape of these tuning curves agrees well with the observed data, the nature of the gammatone filter is evident in the symmetrical tuning curves it produces.

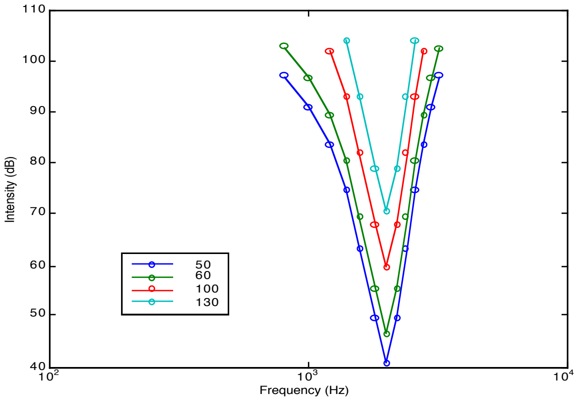


Figure 27. Tuning curves for a single auditory nerve of centre frequency 2kHz (c.f. figure 17). Legend shows spiking rate (per second).

As explained in section 3.1.4.2, auditory nerve fibres exhibit phase locking: the nerve fibre firing is locked to a particular phase of the stimulating waveform. A period histogram can be used to demonstrate the half-wave rectification of the stimulating waveform. A second method of evaluating the extent of phase-locking is by calculating the vector strength of each period histogram as used by Goldberg and Brown (1969). The vector strength is a normalised estimate of the probability of firing at a particular phase in the stimulating waveform. The vector strength r is given by

$$r = \frac{\sqrt{\left[\sum_{k=0}^{K-1} R_k \cos 2\pi(k/K)\right]^2 + \left[\sum_{k=0}^{K-1} R_k \sin 2\pi(k/K)\right]^2}}{\sum_{k=0}^{K-1} R_k} \qquad (7)$$

where K is the number of bins in the period histogram and R_k is the magnitude of the k th bin. When the spike discharge is uniformly distributed across the histogram, r equals zero. In the extreme case of perfect phase locking, r equals 1.

Figure 28 shows a period histogram for the 300ms 1kHz tone with preservation of phase locking at varying intensities. Figure 29 shows the associated vector strength function which confirms the improvement of phase locking with increasing intensity. It is also evident that the strength of the phase locking begins to decrease at high intensities. On examination of the period histograms in figure 28, this is due to a broadening of the histogram peak: increased response to the other phase of the stimulating waveform. Although the strength of the phase locking deteriorates, it is still significantly phase locked and this behaviour is just evident in the physiological period histograms of figure 18.

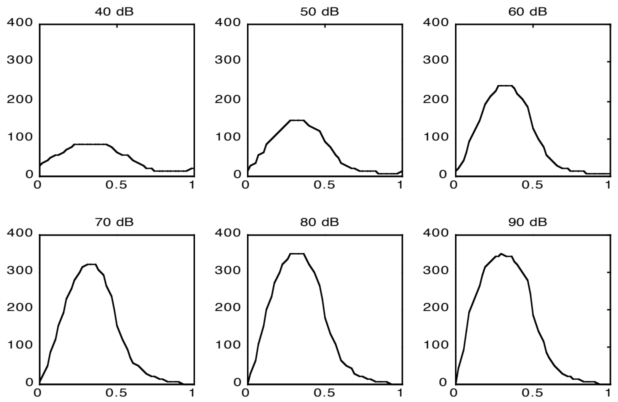


Figure 28. Period histogram of the inner hair cell transduction model in response to a 300ms 1Khz tone starting at 10ms. (c.f. figure 18). Centre frequency of 1kHz.

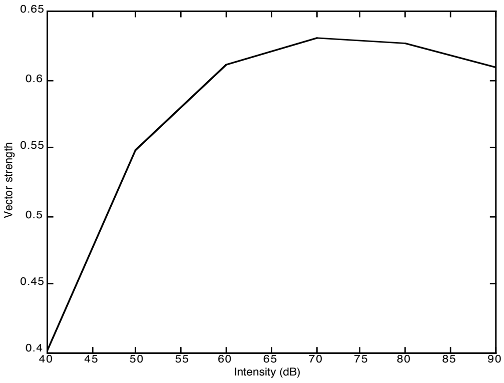


Figure 29. Vector strength with increasing intensity.

Figure 30 shows that as the centre frequency of the fibre approaches 5 kHz, phase locking deteriorates although this is not as pronounced as would be expected from physiological data (figure 19).

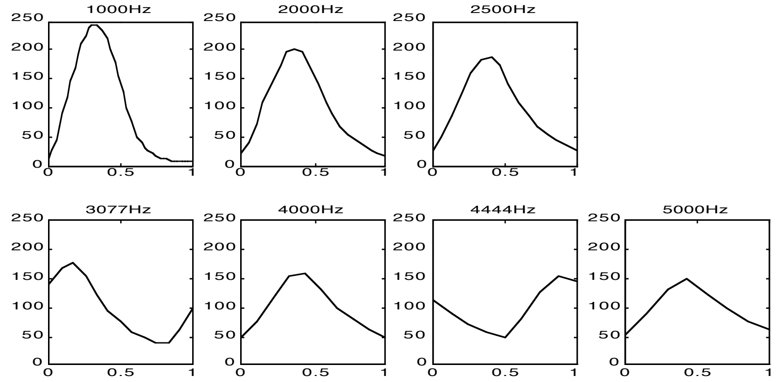


Figure 30. Reduction in phase locking with increasing frequency. (c.f. figure 19).

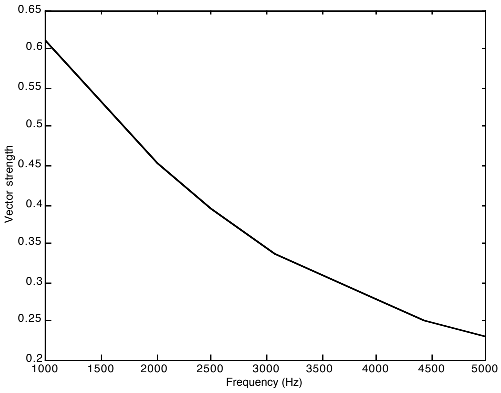


Figure 31. Reduction in vector strength with increasing frequency.

As with the probabilistic output of the Meddis model, period histograms can be produced to confirm that phase-locking properties are preserved after the spike generation process. As can be seen in figure 32, phase-locking is indeed preserved.

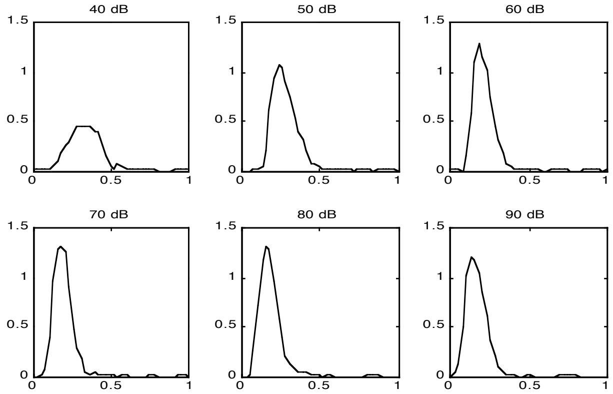


Figure 32. Period histogram of auditory nerve firing output. Centre frequency of 1kHz.

Figure 33, shows the distribution of intervals between successive auditory nerve events. This shows that the model is correctly simulating phase-locking: responses only occur in a particular phase of the stimulus. This is evident by the event intervals being clustered around integer multiples of the stimulating tone period.

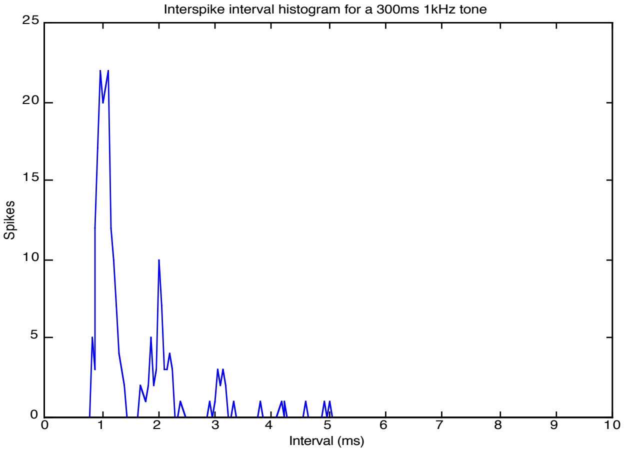


Figure 33. Interspike interval histogram for a 70dB 300ms 1kHz tone. (c.f. figure 20).

4.6. Summary

This chapter has discussed the four main modelling areas involved in simulating the behaviour of the auditory periphery. The Fay (1988) weighting function provides a good approximation to the outer and middle ear resonances. Simulation of place coding is achieved using the gammatone filter which exhibits a very good fit to the human auditory filter shapes. The Meddis (1986, 1988) hair cell model in combination with a spike generation process based on one proposed by Carney (1993) is the final stage of the computational model. This not only provides a good approximation to experimental PSTH shapes but also captures the phenomenon of hair cell phase locking.

Such cochlear nerve activity is then transmitted to neurons within the brain. The next chapter discusses the behaviour of a 'typical' neuron and describes two models of differing complexity which simulate neuron performance.

CHAPTER 5

Neuron Models

The neuron is the basic processing substrate of the animal brain. In mammals, all regions of the neocortex contain approximately 800 million synapses, 4 km of axons and 0.5 km of dendrites per cubic mm; neural densities are as high as 10,000 cells per cubic mm in deep layers of the human motor cortex (Abeles et al., 1994). This gives an indication of the sheer scale of the animal brain. This chapter introduces the key attributes of the neuron, such as the action potential, in terms of biochemical changes that occur within the cell. Bearing in mind the network sizes required, the second half of the chapter looks at two computational models of increasing simplicity.

5.1. Neuron attributes

5.1.1. The equilibrium potential

Most cell membranes exhibit a potential difference, with the inside of the cell being negative in relation to the exterior. This is denoted by using a minus sign. For example, the membrane resting potential of a stellate cell is approximately -65mV. This potential difference is due to the varying concentrations of ions in the extracellular and intracellular fluids.

The ions with most influence are Na^+, K^+ and Cl^-. The extracellular concentration of Na^+ and Cl^- ions is in the order of ten times that of the intracellular concentration. This ratio is reversed for K^+ ions. Cell membranes are virtually impermeable to protein and organic anions. However, they are moderately permeable to Na^+ and more freely permeable to K^+ and Cl^-. In fact, up to 500 times more permeable. These ions do not simply cross the membrane via pores but through specific protein ion channels. Each ion has its own set of channels. In the soma, these channels (and hence the amount of ion diffusion) are voltage-dependent according to the level of the membrane potential.

The varying ionic concentrations and charges in the extracellular and intracellular fluids leads to two types of diffusion gradient: the concentration gradient and the electrical gradient. Chloride ions, as stated above, have a higher extracellular concentration and so tend to diffuse along the concentration gradient into the cell. However, the cell interior is negatively charged (the cell resting potential) and so the negatively charged chloride ions experience electrical repulsion along the electrical gradient out of the cell. Equilibrium occurs at the point at which Cl^- efflux and influx are equal. The situation for K^+ is similar but reversed. The concentration gradient is directed out of the cell and the electrical gradient is directed into the cell. Na^+ diffusion is different again. As is expected, the concentration gradient is directed into the cell, as for chloride ions. However, the electrical gradient is also directed into the cell. As neither of the Na^+ or K^+ ionic transfers are at equilibrium, it would be expected that Na^+ would continue to enter the cell and K^+ would leave. This does not happen and the intracellular concentrations of these two cations remain constant. This is due to the presence of active transport of ions against one or both of its diffusion gradients.

5.1.2. The action potential

A slight decrease in resting potential leads to increased K^+ influx and Cl^- efflux in order to restore the cell's membrane resting potential. Due to the fact that the ion channels act in a voltage-dependent manner, a unique change occurs in the cell once depolarization exceeds approximately 7mV. At this point, Na^+ permeability begins to rise and continues to do so as the level of depolarization reaches the firing level. Once the firing level is reached, Na^+ permeability is so great that the influx of sodium cations swamps the replolarizing process and a runaway depolarization occurs: a spike potential. The Na^+ ions attempt to reach their equilibrium potential of +60mV. However, the Na^+ permeability is short-lived. Na^+ permeability tends towards its resting level during the rising phase of the spike potential and is even decreased during the falling phase. In addition to the change in permeability, the electrical gradient begins to work against the influx of sodium ions once the cell enters the overshoot phase of the spike potential and the interior of the cell becomes positively charged.

In the same manner, K^+ permeability also increases in a voltage-dependent manner, although the onset is slightly later than that for Na^+. K^+ permeability reaches a maximum during the falling phase of the spike potential. As the intracellular concentration of K^+ is much higher than the extracellular concentration, K^+ defuses out of the cell. The net transfer of positive charge out of the cell completes repolarization. Na^+ efflux hyperpolarises the cell until equilibrium is restored.

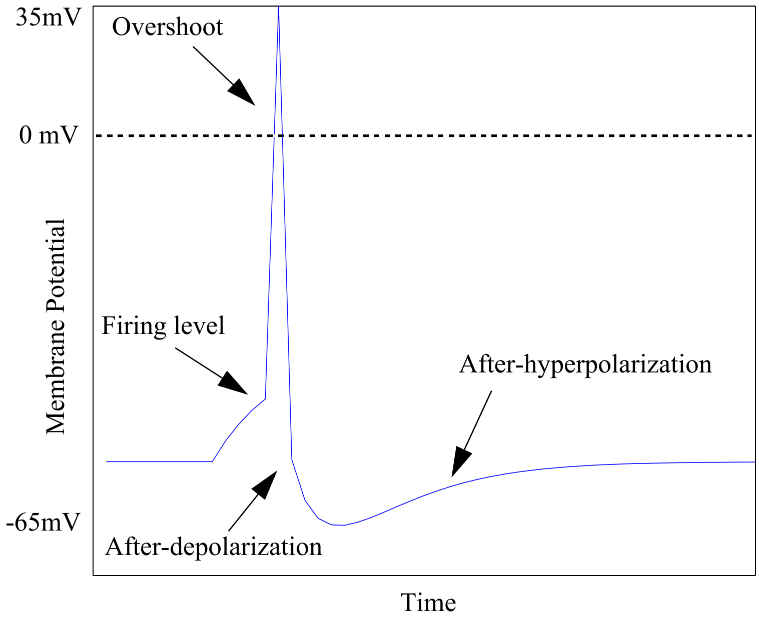


Figure 34. Typical cell action potential. The proportions of this diagram have been intentionally distorted to highlight the various phases of the action potential.

Although not as important as Na^+, K^+ and Cl^-, it is worth mentioning the role of Ca^{2+} in the action potential. As with Na^+, Ca^{2+} electrical and concentration gradients are directed into the cell. It is thought Ca^{2+} enters the cell via the Na^+ ion channels although in much smaller amounts. The early phase of Ca^{2+} influx is blocked by the poison tetrodotoxin (TTX) which blocks Na^+ channels without affecting the K^+ channels. Later Ca^{2+} influx is thought to occur via different voltage-sensitive pathways. Not only does Ca^{2+} play an important role in the secretion of synaptic transmitters, it also aids the depolarization of the cell prior to the spike potential.

5.2. MacGregor point neuron model (ptnrn10)

The MacGregor point neuron model (MacGregor, 1987 p458) produces relatively realistic firing properties for a neuron with basic accommodation properties. In fact, figure 34 was produced by the model. The model describes the simplified processes that occur in a neuron in terms of the cell's membrane potential *E*, potassium conductance G_k and its firing threshold *Th*. The cell behaviour is described by three linked differential equations:

$$\frac{dE}{dt} = \frac{-E + V + G_k(E_k - E)}{\tau_{mem}} \tag{8}$$

$$\frac{dTh}{dt} = \frac{Th_0 - Th + cE}{\tau_{Th}} \tag{9}$$

$$s = \begin{cases} 1 & E \geq Th \\ 0 & E < Th \end{cases} \tag{10}$$

$$\frac{dG_k}{dt} = \frac{-G_k + bs}{\tau_{G_k}} \tag{11}$$

$$P = E + s(50 - E) \tag{12}$$

Variable	Function	Typical Value	Units
E	Transmembrane potential.	-	mV
V	Stimulating voltage from an applied electrode as a function of time.	-	mV
G_k	Potassium conductance above resting level.	-	S
E_k	Equilibrium potential of the potassium conductance.	-10	mV
τ_{mem}	Membrane time constant.	5-11	ms
Th	Time-varying firing threshold.	-	mV
Th_0	Resting threshold of the cell.	10-20	mV
c	Accommodation constant.	0-1	-
τ_{Th}	Accommodation time constant.	20-25	ms
s	Spiking variable.	0 or 1	-
b	Potassium conductance rise constant.	4	nS
τ_{Gk}	Potassium conductance decay time constant.	3-10	ms
P	Cell output.	-	mV

Table 2. Point neuron model equation variables.

The important aspect of the equations is that they describe the rate of change of a particular variable. Therefore, the description will be made in terms of rate of change.

At equilibrium, the rate of change of each variable will be zero. On application of a steady current, the rate of change of *E* will increase due to the *V* term in equation 11. The rate of this change is determined by the value of τ_{mem}, the membrane time constant. The smaller this value, the faster the rate of change of the cell membrane potential. The increase is rapid at first due to the relatively large difference between the cell membrane potential of the applied potential. However, as this differences reduces, so too does the rate of change. This exponential rise in cell membrane potential can been seen in figure 36.

As the membrane potential *E* increases, the threshold rate of change begins to increase due to the *cE* term in equation 9. Once again, the rate of this change is regulated by a time constant. In this case, τ_{Th}. However, the threshold should not continue to increase in proportion to the value of *E*. As the value of the threshold increases above its resting level Th_0, the rate of change of *Th* is encouraged to diminish due to the Th_0-Th term. The model used to create figure 36 included a high level of accommodation and so threshold rise is evident (see also figure 35).

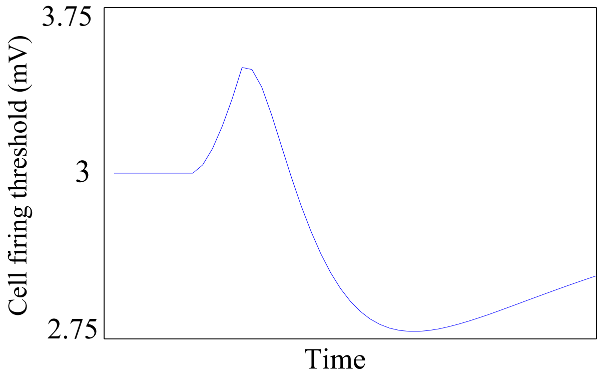


Figure 35. Cell firing threshold variation during a single action potential.
In this example, Th_0 was 3mV and *c* was 0.5.

If the membrane potential exceeds the cell threshold, a spike potential is generated (denoted by assigning 1 to the value of *s*).

It is known from the function of real cells that shortly after a spike potential is initiated, the level of potassium conductance increases. The use of *bs* in equation 11 ensures that the potassium conductance increases. As shown in figure 36 the rate of decay back to its equilibrium value is exponential with a time constant τ_{Gk}.

The increased level of G_k causes E to drop due to the $G_k(E_k\text{-}E)$ term in equation 8. Note that E, the membrane potential, is not the same as the output potential, P. While G_k is in an elevated state, the transmembrane potential E rapidly drops towards its equilibrium value. However, as G_k diminishes, the applied current to the cell again begins to dominate and causes E to rise. E surpasses the threshold Th and a second action potential is elicited causing another rise in G_k. As G_k drops the potential E goes back up, and if the firing threshold has risen to a level higher than the equilibrium potential of the cell associated with the value of the applied current, no further firing occurs. The model used in figure 36 has strong accommodation and so the threshold increase is evident. This type of response illustrates on-response to a steady step current.

In figure 34, the applied current has also been removed by the falling phase of the spike potential and so the reduction of E overshoots the cell equilibrium potential and causes hyperpolarization which slowly returns to the resting potential.

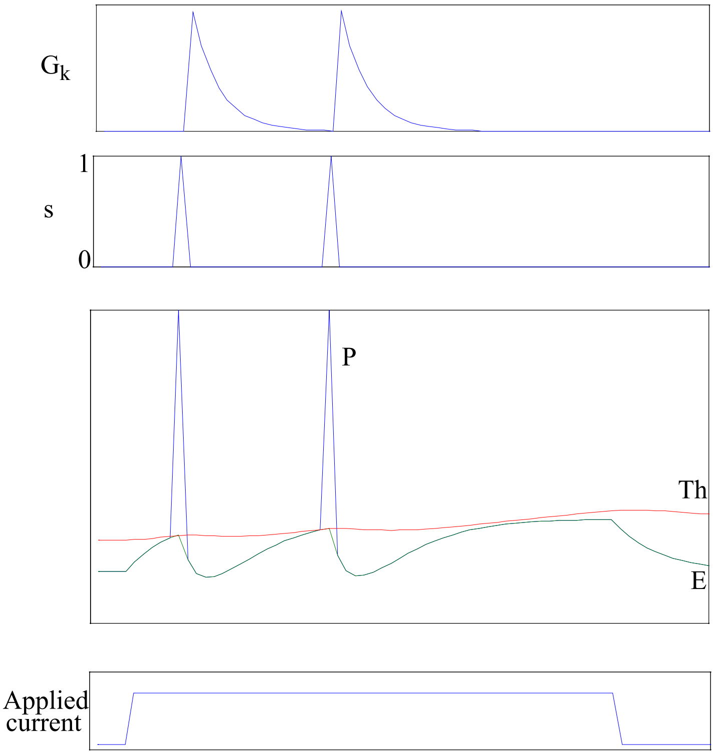


Figure 36. Behaviour of the state variables for repetitive firing with strong adaptation.

5.3. Integrate and fire neuron model

The MacGregor point neuron model represents a useful simplification of the Hodgkin and Huxley (1952) model but is still computationally expensive if large networks are to be simulated. When considering the simple network described in chapter 6, the computational expense of this neuron model becomes unacceptable. The integrate and fire neuron (*leaky* neuron) model is simplistic - the behaviour of the cell is determined by two properties: the membrane potential and the non-adaptive membrane threshold. Input to the cell is in the form of spikes which are represented as binary values.

$$E_t = E_{t-1}e^{\frac{-T_s}{\mu}} + E_r\left(1 - e^{\frac{-T_s}{\mu}}\right) + H_t \qquad (13)$$

where

$$H_t = \begin{cases} V & \text{if spike input occurs at time } t \\ 0 & \text{otherwise} \end{cases}$$

and $E_t = E_{ref} \quad \textit{if} \quad E_t \geq E_{thresh}$

$$O_t = \begin{cases} 1 & \text{if } E_t \geq E_{thresh} \\ 0 & \text{otherwise} \end{cases}$$

Table 3. Integrate and fire neuron model variables and their typical values.

Variable	Function	Typical Value	Units
E_t	Membrane potential	-	mV
E_r	Resting membrane potential	-60	mV
E_{thresh}	Threshold potential	-50	mV
E_{ref}	Refractory potential	-70	mV
V	Input potential (per spike)	8	mV
μ	Decay time constant	10	ms
T_s	Sampling period	-	ms
T_{ref}	Refractory period	3	ms
O_t	Cell output	-	mA

At equilibrium, the cell membrane potential rests at E_r. On receiving a spike, the cell membrane potential is raised by V. If the new value of E_t is below the cell threshold, E_{thresh}, no action potential is generated. Provided no further spike input is received, E_t decays back towards E_r with a time constant of μ.

If E_t exceeds the cell threshold, an action potential occurs (signified by O_t being set to 1) and the cell enters a period of absolute refractoriness. During this period, the cell membrane potential is fixed at E_{ref} for a period of T_{ref}. E_{ref} is typically below the cell resting potential. Subsequently, the cell enters a period of relative refractoriness during which the cell membrane potential decays back to the resting potential with a time constant of μ.

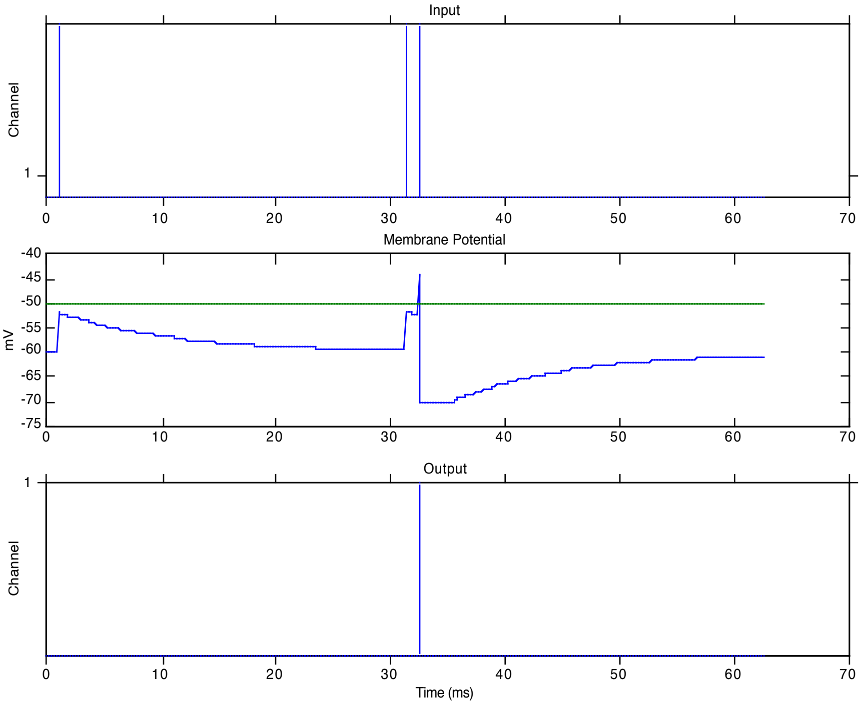


Figure 37. Behaviour of the integrate and fire neuron in response to a train of three spikes. In the middle panel, the cell threshold is shown in green.

5.4. Summary

Two models of neuron behaviour have been described. Despite the accuracy if MacGregor's model and its close relation to the chemical transfers involved real neurons, it is still too computationally expensive for any network of more than a few cells. The integrate and fire neuron provides an extremely simple mechanism to simulate neuron behaviour while maintaining a large degree of accuracy.

The next chapter describes how the integrate and fire neuron model is incorporated into the synfire chain network.

CHAPTER 6

Synfire Chain Network

The synfire chain network is a form of two-dimensional neural network whose connections are organised in a feedforward manner. The cell model used for the network is the integrate and fire cell described in the previous chapter. This chapter describes the key terms associated with synfire transmission and then describes in more detail the network topology and output assessment criterion. Finally, the ability of the network to bind features according to frequency proximity is considered.

6.1. Synchronous transmission

The long delays encountered in reaction-time experiments can be explained by the hypothesised structure of the processing substrate. Long delays are accounted for by assuming that information processing occurs in a serial manner: spikes travel from one neuron to another along a chain of *n* neurons (figure 38a).

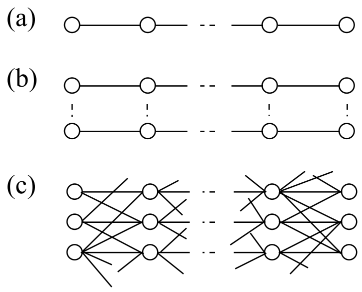


Figure 38. Alternative connections between neurons. (a) serial; (b) parallel serial; (c) diverging/converging. After Abeles (1991), figure 6.1.1.

However, this assumption is flawed. 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, Tomlinson and Gibson, 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 38b) provides this. Unfortunately, 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 38c) simply incorporates redundancy while limiting the number of neurons required.

Abeles (1991) contends that information transmission in the cortex is likely to occur between sets of neurons connected by such diverging and converging pathways. There are two possible mechanisms of transmission: synchronous and asynchronous. In asynchronous transmission, cells of the ‘sending’ node begin to fire at a high rate. Due to spatial and temporal summation, this causes cells of the ‘receiving’ node to fire. Synchronous transmission relies on the cells of the sending node firing in synchrony. The receiving node cells experience a synchronised volley of spikes causing them to also fire in synchrony. This form of transmission assumes that the synapses are strong enough to ensure synchronous firing and that there is sufficient allowance for the jitter in spike timings.

A special case of synchronous transmission is that of *synfire* transmission. 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.

6.2. Synfire chain network

The role of frequency proximity plays an important part in auditory feature binding and has been the subject of a number of perceptual experiments. In addition to the two tone streaming phenomenon documented by van Noorden (1975), Bregman and

Pinker (1978) used a similar experiment to demonstrate the influence of frequency proximity. The experiment consisted of a repeating cycle formed by a pure tone A and a two pure-tone complex of B and C (see figure 39). An important result from the study was that the frequency proximity of the tones A and B influenced how the sequence was perceived. The closer A and B were in frequency to one another, the greater the likelihood of them forming a stream separate from tone C.

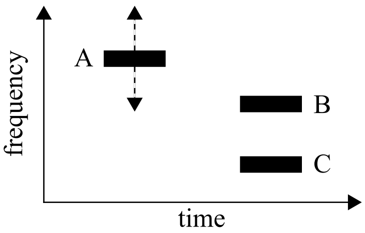


Figure 39. Stimulus used by Bregman and Pinker (1978).

Vicario (1982) has also reported an effect consistent with frequency proximity grouping. In his study, Vicario used the stimulus shown in figure 40. A pure tone M was sounded for a short period and was then joined by another two tones H and L. After the tones H and L ceased, tone M continued for a short period.

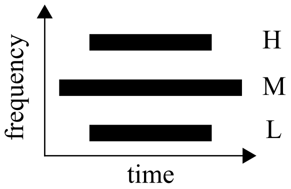


Figure 40. Stimulus used by Vicario (1982).

Vicario found that the perception of tone M as a distinct entity was not reduced by increasing the number of tones it has to pass through. He concluded that it was only the local proximity of tones H and L that made tone M harder to hear as they became closer in frequency.

It is this form of feature binding that the synfire chain network aims to simulate: as the frequency separation of two stimuli increases, the likelihood of binding should decrease. Thus, as separation increases, the synchronisation of the two centre frequencies should decrease (figure 41).

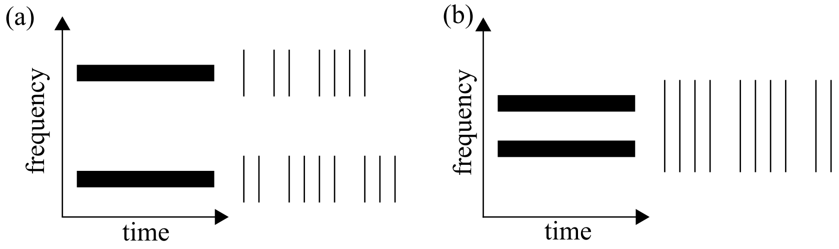


Figure 41. Desired synfire chain network response to spectrally distant stimuli (a) and spectrally close stimuli (b).

6.2.1. Network topology

The synfire chain was 5 neurons long and received input from 50 frequency channels (figure 42). It should be pointed out that although the input is referred to in terms of frequency channels, the current network does not possess an auditory periphery model.

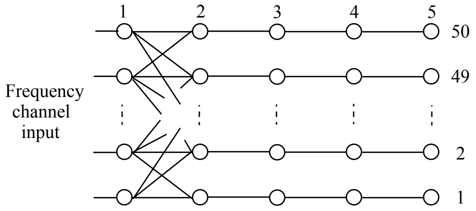


Figure 42. The synfire chain network used to simulate frequency proximity grouping. The synfire chain was 5 neurons long and received input from 50 frequency channels. For reasons of clarity, only connections in the first synfire link have been shown.

Activation of different centre frequencies is represented by random spike trains with an interspike interval of no less than 1 ms (absolute refractoriness). Synapse strengths vary with distance in a gaussian fashion (figure 43).

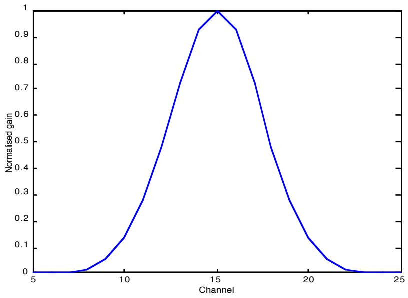


Figure 43. Input synapse strengths for cell in channel 15.

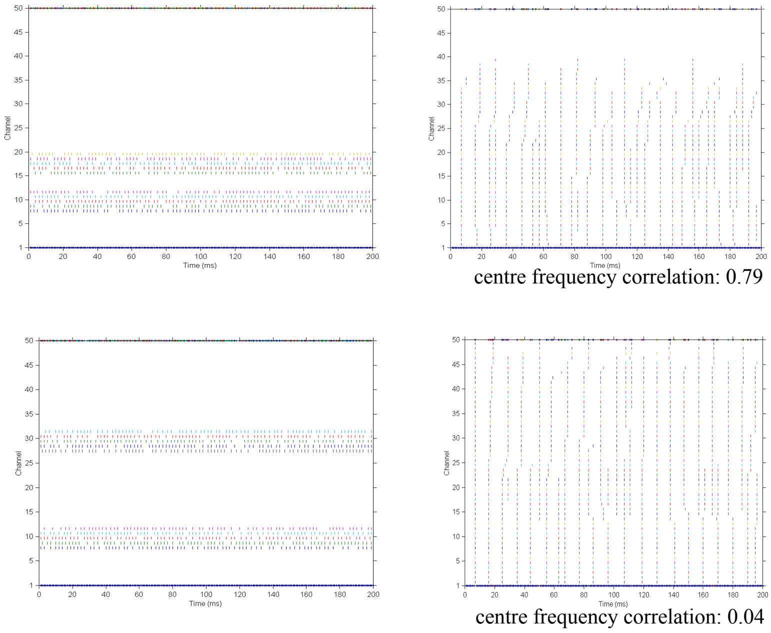
6.2.2. Grouping by frequency proximity

The nature of the local feedforward connections give rise to grouping by frequency proximity; spatially close inputs give rise to synchronised activity, whereas spatially distant inputs do not. The degree of synchronisation between centre frequencies is assessed by the correlation of their outputs *X* and *Y*,

$$C(X, Y) = \frac{\sum x(t)y(t)}{\sqrt{\sum x^2(t)y^2(t)}} \qquad x(t) = X(t) - \langle X \rangle \qquad (14)$$

where $\langle X \rangle$ is the mean of *X(t)*.

In the experiment, two centre frequencies were stimulated with random spike trains. Each activated frequency band had a width of 5 channels (see figure 44). This experiment was repeated for a range of 20 channel separations. Additionally, the experiment was performed with a number of connectivities ranging from 3 to 15 connections on each side of the centre frequency.



centre frequency correlation: 0.79

centre frequency correlation: 0.04

Figure 44. Example inputs to the synfire chain network. Two bands of random spike input are present with a frequency separation of 8 (top) and 20 (bottom) channels. An increase in centre frequency separation reduces centre frequency channel synchronisation.

As expected, as frequency separation increased, the correlation and hence synchronisation between the outputs of the centre frequencies would decrease and increased connectivity delayed the decrease in synchronisation (figure 45). It is evident in the network's response to the 20 channel separation example (figure 44) that a number of intermediate channels are synchronised to both frequency bands. In order to determine the synchronicity between the two bands, only the correlation between the two centre frequencies is calculated. Physiologically, this can be thought of as peripheral activity weighting the channel binding decision.

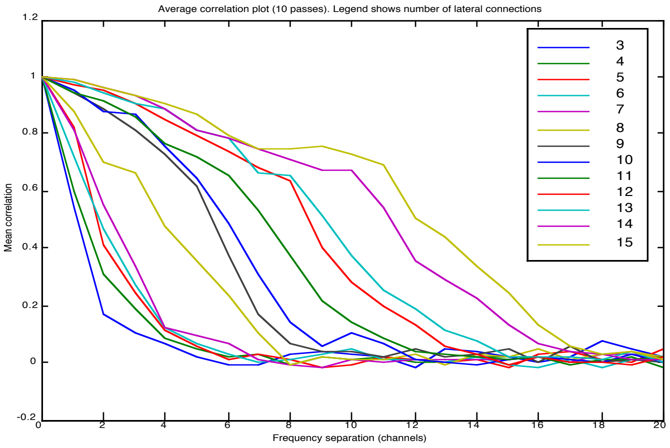


Figure 45. Varying centre frequency synchronisation as frequency separation increases.

6.3. Summary

The model shows behaviour that is consistent with grouping by frequency proximity. However, to fully model the perception of alternating tone sequences, the model must incorporate grouping by temporal proximity as well. There is no mechanism to perform temporal grouping in the current model. This will be discussed in the next chapter.

CHAPTER 7

Conclusions

As discussed in chapter 2, primitive grouping encompasses the data-driven simultaneous and sequential perceptual organisations of sound. Simultaneous organisations correspond to grouping by sound source onset and offset, harmonicity and frequency proximity. In contrast, sequential organisations make use of continuity and proximity constraints across time.

Despite the goal of simulating two tone streaming effects, the synfire chain model described in chapter 6 cannot perform sequential grouping. The next stage of development needs to concentrate on implementing a form of short-term memory (STM). Grossberg (1996) suggested a form of channel *resonance* in which channel stimulation produces a build of activity which continues after the stimulus has ceased. Over time, the resonance decays away.

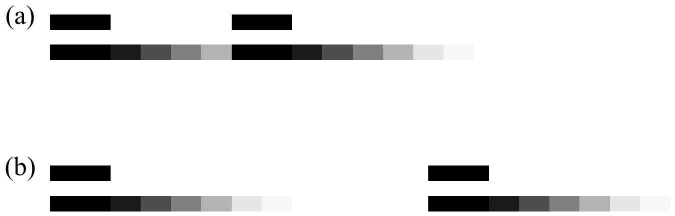


Figure 46. Channel resonance grouping. Top trace shows channel stimulus; bottom trace shows channel activity. (a) two stimuli are grouped temporally. (b) two stimuli are not grouped temporally.

Temporal proximity grouping could be achieved by using channel resonance to link successive stimuli: if a second tone occurs within some time period of the first tone, the later tone ‘accesses’ the resonance and the two stimuli are grouped (figure 46a). If the second stimulus occurs after the resonance has decayed away, the two stimuli are not grouped temporally (figure 46b).

An alternative STM mechanism is the use of time-delayed feedback connections. Such connections would increase the channel’s mean activity thus increasing its ability to group a subsequent stimulus in a similar manner to Grossberg's resonances. These two methods will be investigated, along with others, in future work.

The implementation of short term memory is related to the representation of time within the network. Previous work by von der Malsburg and Schneider (1986) and Liu *et al.* (1994) both use spectral inputs but do not allow responses at different times to be compared.

Related to this, is the time course of stream formation and segregation. For example, it may take up to ten seconds for a sequence of alternating tones to segregate into two perceptual streams (Anstis and Saida 1985). Other models (Brown and Cooke, 1997; Beauvois and Meddis, 1991, 1996; McCabe and Denham, 1997) successfully simulate the build up of streaming over time. If the network presented here is to achieve its objective of being physiologically plausible, it too must be able to simulate this build up.

The role of attention in auditory grouping has been acknowledged by previous modellers (e.g. McCabe and Denham, 1997; Brown and Cooke, 1997) but has not been implemented. Crick (1984; Crick and Koch, 1990) has suggested that part of the thalamus (the thalamic reticular complex) may be involved in selective attention. Until recently, it was thought stream formation such as that involved in the two tone streaming phenomenon was passive in nature: streaming occurred whether attended to or not. Attention was considered useful only in guiding a particular stream into the attentive 'foreground'. However, recent work by Carlyon *et al.* (1999) suggests that attention does indeed play an important role in stream formation.

In Carlyon's experiment, a 21s sequence of A and B pure tones alternating in an ABA-ABA sequence was presented to the left ear. In the 'baseline' condition, 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. After 10s the subjects switched to the streaming task. In the 'one-task-with-distractor' condition the noise bursts were presented to the right ear as in the two-task condition, but

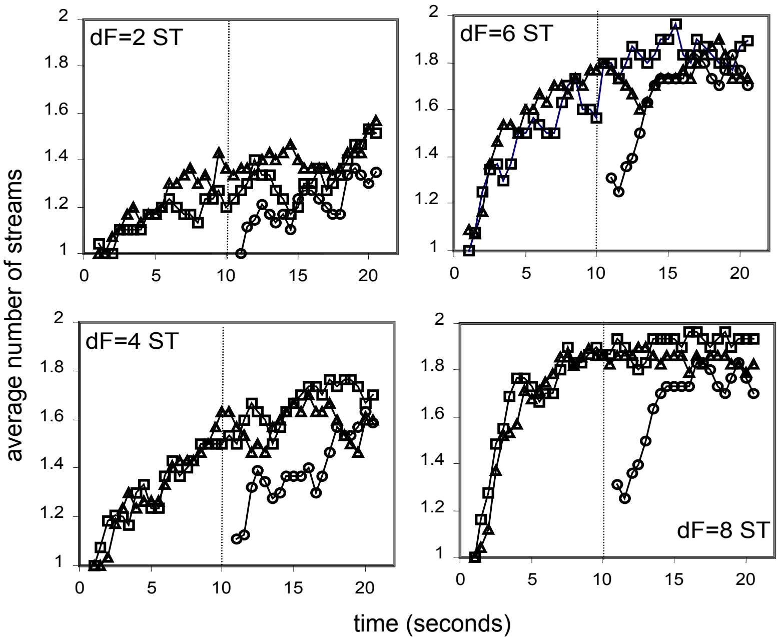


Figure 47. Build up of streaming over time for four frequency differences. Scores are averaged across listeners and repetitions for the baseline (triangles), two-task (circles), and one-task-with-distractor (squares) conditions. From Carlyon (1999) figure 3.

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 is similar to that at the beginning of the baseline sequence - in the absence of attention, streaming had not built up.

The findings of Carlyon *et al.* suggest that attention is crucial for the build-up of auditory streaming. The implications of this will be studied further in future work.

In summary, the nature of the synfire network described elegantly simulates feature binding based on frequency proximity. In order to incorporate a physiologically plausible short term memory, the representation of time within the network needs to be addressed. In addition to this, the role of attention in stream formation needs to be reviewed.

CHAPTER 8

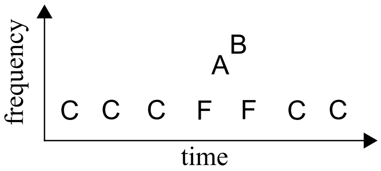
Future Work

The next stages of modelling involve incorporating a mechanism that will allow sequential grouping to be performed: a form of short term memory. In addition to this, recent work has called into question the role of auditory attention in stream segregation. Both of these areas will be studied with the intention of producing a new model of auditory stream segregation.

8.1. Auditory attention

When thinking about how auditory features are grouped, it is natural to enquire how such arguably primitive mechanisms relate to other aspects of perception such as attention. 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 pitch - 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 48).

Figure 48. Tone sequence used by Bregman and Rudnicky (1975).



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: stream segregation without attention.

Recent work by Carlyon *et al.* (1999) brings this theory into question. Carlyon *et al.* performed a number of experiments in which listeners were presented with a different stimulus to each ear: a repeating tone sequence and a repeating noise burst sequence. The results show that when listeners concentrated on describing the nature of the noise bursts, stream segregation of the tone sequence did not occur. Furthermore, a second experiment required listeners to assess the nature of the tones which made up the sequence - ‘fast’ or ‘slow’ amplitude modulation - in order show that the lack of steam segregation in the first experiment was not a result of attending to a different ear. Again, stream segregation did not occur in the presence of the attended auditory task.

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.”*

In summary, evidence produced by Carylon *et al.* (1999) suggests that the long standing role of attention may be incorrect. Instead of being a mechanism for bringing an existing stream to the perceptual foreground, it may be an essential part of stream segregation. Therefore, in order to produce an accurate simulation of stream segregation, the mechanism of attention needs to be reviewed.

8.2. Short term memory

Anstis and Saida (1985) have shown that streaming decisions are not instantaneous. When assessing the time course of the stream formation and segregation process, listeners demonstrate that it can take up to ten seconds for a sequence of alternating tones to segregate into two different perceptual streams. The initial state of streaming is always temporal coherence: only one stream exists. This is then altered over a period of time. Beauvois and Meddis (1991, 1996) implement this by incorporating a random bias at every time step into their model which stimulates increased segregation over time. In order to create a complete model, this build up of streaming must be included.

As discussed in the previous chapter, a form of short term memory is also required to account for sequential grouping. A suggested mechanism for this process is that of resonance (Grossberg, 1996) which allows the activity within a channel to be sustained for a period of time. A subsequent tone which occurs within the time period is able to 'access' the resonance and be grouped with the previous tone (figure 46). However, this solution does not fully explain some aspects of the auditory induction phenomenon.

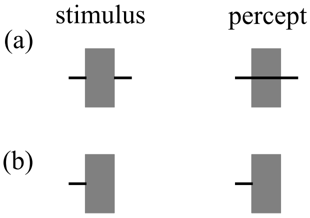


Figure 49. Auditory induction on tone-noise sequences.

Figure 49 shows two situations in which a tone is followed by a broadband noise burst (Miller and Licklider, 1950). When the tone-noise sequence is immediately followed by a second tone of the same frequency as the first, a single tone is perceived to travel through the noise burst (figure 49a). This can be easily explained by Grossberg's resonance theory: the second tone 'accesses' the first tone's resonance. However, the second example cannot be explained in this manner. If the tone-noise sequence is not followed by a second tone, the first tone is not perceived to continue through the noise burst (figure 49b). The resonance of the first tone will still occur and continue activity in that channel though the noise burst. Grossberg's theory does not explain how this resonance is suppressed.

For the model to provide a plausible explanation of stream segregation using a form of short term memory and auditory attention, it must take into account the factors described above. This work is a computational modelling study with the eventual goal of producing a model which will increase our understanding of the role of auditory attention and also cast light on the behaviour of short term memory.

8.3. Time plan

	1999	2000				2001		
	Oct - Dec	Jan - Mar	Apr - Jun	Jul - Sep	Oct - Dec	Jan - Mar	Apr - Jun	Jul - Sep
Write up							■	■
Conferences*	▌[1]			▌[2]	▌[3]			▌[2]
Other work	DERA	DERA						
STM Research	Modelling	Modelling	Modelling Simulation	Simulation				
Attention Research				Modelling	Modelling	Modelling Simulation	Simulation	

* I shall attend the Neural Information Processing Systems (NIPS) Conference in November / December 1999[1]. It is envisaged that I will attend the next two annual British Society of Audiology Short Papers meetings[2] and also NIPS 2000[3].

The feasibility study for the Defence, Evaluation and Research Agency (DERA) aims to apply CASA techniques to signal-to-noise enhancement problems involving sonar signals. Work on this part time project will be completed at the end of February 2000.

The remaining two years are split evenly between research on short term memory and auditory attention. Although these two topics have been scheduled to run sequentially, it is expected that a certain amount of overlap will occur due to possible interaction between the two fields of study. Each work package will involve a detailed literature review of physiological data and perceptual experiments followed by the construction of a computational model. It is expected that such models will be based upon the synfire chain network described in this report, although the use of relaxation oscillators (e.g. Wang, 1996) has not been ruled out. Each work package will end with an evaluation of the computational model.

Four months has been allocated to writing the final thesis. I plan to submit by 1 October 2001.

CHAPTER 9

References

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

Abeles, M, Prut, Y, Bergman, H and Vaadia, E (1994). Synchronisation in neuronal transmission and its importance for information processing. *Progress in Brain Research* **102** 395-404.

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

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.

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.

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.

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

Bregman AS (1997). Psychological data and computational ASA. In *Readings in Computational Auditory Scene Analysis*, edited by H.Okuno and D.Rosenthal, Lawrence Erlbaum.

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

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

Brown, GJ (1992). *Computational auditory scene analysis: A representational approach*, Ph.D. thesis CS-92-22, CS dept., Univ. of Sheffield.

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 (1999). Timing is of the essence: Neural oscillator models of auditory grouping. In *Listening to Speech*, edited by S.Greenberg and W.Ainsworth, Oxford University Press, in press.

Campbell, SR and Wang, DL (1996). *Relaxation oscillators with time delay coupling*. CIS-Technical Report #47, Ohio State University, Ohio.

Carlyon, RP, Cusack, R, Foxton, JM and Robertson, IH (1999). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance* (submitted).

Carney, LH (1993). A model for the responses of low-frequency auditory-nerve fibers in cat, *Journal of the Acoustical Society of America* **93**(1) 401-417.

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.

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

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

Crick, F and Koch, C (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences* **2** 263-275.

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

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

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

Fay, RR (1988). Comparative psychoacoustics, *Hearing Research* **34**(3) 295-305.

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.

Goldberg, JM and Brown, PB (1969). Responses of binaural neurons of dog superior olivary complex to dichotic tonal stimulation: Some physiological mechanisms of sound localization, *Journal of Neurophysiology* **32** 940-958.

Grossberg, S (1996). Pitch-based streaming in auditory perception. In *Creative Networks*, edited by N.Griffith and P.Todd, MIT Press, MA.

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.

Hopfield, JJ (1995). Pattern recognition computation using action potential timing for stimulus representation. *Nature* **376** 33-36.

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

Horn, D and Usher, M (1992). Oscillatory model of short term memory. In *Advances in Neural Information Processing Systems 4*, edited by J.E.Moody, S.J.Hanson and R.PLippmann.

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.

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

Liberman, MC (1982). Single neuron labelling in the cat auditory nerve, *Science* **216** 1239-1240.

Lisman, JE and Idiart, MAP (1995). Storage of 7 ± 2 short-term memories in oscillatory subcycles. *Science* **267** 1512-1515.

Lui, 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.

MacGregor, RJ (1987). *Neural and Brain Modeling*. Academic Press.

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

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.

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.

Patterson, RD and Moore, BCJ (1986). Auditory filters and excitation patterns as representations of frequency resolution, in *Frequency Selectivity in Hearing* (ed. BCJ Moore). Academic Press, 123-177.

Phillips, WA and Singer, W (1997). In search of common foundations for cortical computation. *Behavioral and Brain Sciences* **20** 657-722.

Pickles, JO (1988). *An Introduction to the Physiology of Hearing*, 2nd Edition. Academic Press.

Plomp, R (1964). The ear as a frequency analyzer. *Journal of the Acoustical Society of America* **36** 1628-1636.

Rose, JE, Brugge, JF, Anderson, DJ and Hind, JE (1967). Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey, *Journal of Neurophysiology* **30** 769-793.

Schroeder, MR and Hall, JL (1974). Model for the mechanical to neural transduction in the auditory receptor. *Journal of the Acoustical Society of America* **55**(5) 1055-1060.

Sherrington, CS (1941). *Man on His Nature*, Cambridge University Press, Cambridge.

Singer, W (1993). Synchronisation of cortical activity and its putative role in information processing and learning. *Annual Review of Physiology* **55** 349-374.

Todd, N (1996). An auditory cortical theory of primitive auditory grouping. *Network: Computational in Neural Systems* **7** 349-356.

van Noorden, LPAS (1975). *Temporal coherence in the perception of tone sequences*. Doctoral thesis, Institute for Perceptual Research, Eindhoven, NL.

Vicario, G (1982). Some observations in the auditory field. In *Organization and Representation in Perception*, edited by J.Beck, Erlbaum.

von der Malsburg (1981) *The correlation theory of brain function*. Internal report 81-2, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany.

von der Malsburg, C and Schneider, W (1986). A neural cocktail-party processor. *Biological Cybernetics* **54** 29-40.

Wang, DL (1996). Primitive auditory segregation based on oscillatory correlation. *Cognitive Science* **20** 409-456.

Wang, DL, Buhmann, J and von der Malsburg, C (1990). Pattern segmentation in associative memory. *Neural Computation* **2** 94-106.

Wang, DL and Brown, GJ (1999). Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Transactions on Neural Networks* **10** 684-697.

WWW (1999). `http://weber.u.washington.edu/~otoweb/middle_ear.html`