Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Efficient Coding of Sensory Stimuli.

Garrett Greene

A thesis submitted to the University of Dublin, Trinity College
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

School of Mathematics,
Trinity College Dublin

April 21, 2011

Declaration

I, the undersigned, (known as the candidate) declare that:

1. this work has not been submitted as an exercise for a degree at this University or at any other University *and*

2. this is entirely the candidate's own work and all work by others which is discussed is duly referenced and acknowledged *and*

3. the candidate agrees that the Library may lend or copy the thesis upon request and that this permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement.

Summary

An important goal of mathematical neuroscience is to understand the coding principles governing the behaviour of sensory systems under stimulation. Here, we investigate the theory of efficient coding in neural sensory systems using linear models of neuronal response, as well as network models of neuronal interactions in sensory cortex.

The response of an auditory neuron to a stimulus is commonly described by its Spectro-Temporal Receptive Field (STRF), a linear kernel which gives a prediction of the neural firing rate. We describe a novel method of calculating STRFs which gives improved predictive accuracy over the standard algorithm. Furthermore, this method clarifies the STRF calculation, and avoids the use of approximations employed in the established method. We then use this STRF formulation to calculate an optimal sparse code for natural stimuli.

Sparse Coding is a sensory coding strategy which represents a trade-off between metabolic efficiency and representational accuracy. We compute an optimal sparse code for an ensemble of natural sounds consisting of zebra finch song recordings. Using our simplified STRF formulation, we obtain a set of predicted STRF-like kernels which allow an accurate sparse coding of zebra finch song. These are compared to STRFs from the Field L region of the zebra finch auditory pathway. The sparse kernels and the receptive fields, though differing in some respects, display several significant similarities, which are described by computing quantative properties such as the seperability index and Q-factor. These findings imply that Field L neurons are specifically adapted to sparsely encode birdsong and supports the idea that sparsification may be an important element of early sensory processing.

Finally, we demonstrate that sparse coding may arise naturally in sensory systems through neuronal interactions. We construct a simple network model of sensory neurons which generates accurate sparse representations of novel stimuli. This network learns a sparse code for a stimulus ensemble in an unsupervised manner,

and thus suggests a biologically plausible mechanism whereby sparse coding may be implemented in primary sensory areas.

Acknowledgements

I would like to thank my parents, Eileen and Brendan, for constant support in everything I do, and my siblings Barry and Ciara for help and advice, and for setting a good example. My supervisor Conor Houghton has provided wonderful guidance and insight without which I would never have got this far. I must also thank all the members of the Mathematical Neuroscience Lab, in particular James, for putting up with my mess, and Peter, for providing an unfailing supply of tea. Finally, I have to thank Clare, for always helping, and for constantly prodding me until I did some work.

Contents

Chapter 1

Introduction

Neurons are the principle active components of the nervous system and are believed to perform most of the cognitive and computational functions of the brain. For this reason, a proper understanding of the behaviour of individual neurons and their response to stimuli is essential to building a bottom-up understanding of brain function. However, at the risk of understatement, the brain is a complex system, and in its computational abilities is far greater than the proverbial sum of its parts. To try and explain brain function, it is therefore necessary to study neural systems at many levels, from the molecular biology of individual cell membranes, through cortical networks, up to emergent behaviour at the level of brain regions.

The work presented here focuses on the level of individual neurons and small scale networks of neurons in the sensory cortex. Arguably, it is on this scale of local networks that the brain is most amenable to mathematical modelling, being small enough to to take into account the dynamics of individual cells but large enough to perform recognisable computational functions, such as signal filtering and decomposition. At this level it is possible to understand how mechanisms of neural computation and information transmission can arise from the biophysical characteristics of neuronal networks.

1.1 The Neuron

Neurons are found in all but the simplest forms of animal life and are thought to be responsible for practically all motor and sensory functions. A neuron is a single cell which can be electrically excited by the exchange of ions through its outer membrane. Neurons can transmit electrical signals in the form of voltage spikes, known as action potentials, which result from non-linear feedback effects in the action of ion channels in the cell membrane.

There exists a significant difference in electrical potential between the interior of a neuron and the outside environment. This is a result of differing concentrations of sodium and potassium ions, as well as small quantities of other ion species such as calcium. These differences create a voltage across the cell membrane, known as the membrane potential, with a typical resting value of around -70mV.

Small changes in this membrane potential can cause voltage-dependent ion gates in the cell wall to open. If a depolarizing (positive) change occurs, sodium gates open, allowing a flow of depolarizing Na^+ ions to enter the cell, while potassium gates allow an opposing current of K^+ ions to leak out of the cell. If the fluctuation in potential is small, the potassium leak current quickly exceeds the sodium current, and the membrane potential returns to it's resting value. However, the voltage dependency of ion gates is highly non-linear: if the initial depolarization is sufficient to raise the membrane potential above a certain threshold value - typically around -55mV - many more sodium gates are opened, creating a runaway non-linear excitation, in which the membrane potential rapidly rises to values as high as $+30$mV, before just as rapidly falling again due to a similar non-linear increase in the potassium current. After a brief oscillatory period, the membrane potential quickly returns to its resting value of -70mV. The creation of a voltage spike in one patch of cell membrane stimulates the opening of ion channels in neighbouring patches, and so the spike, or action potential, propogates along the length of the neuron.

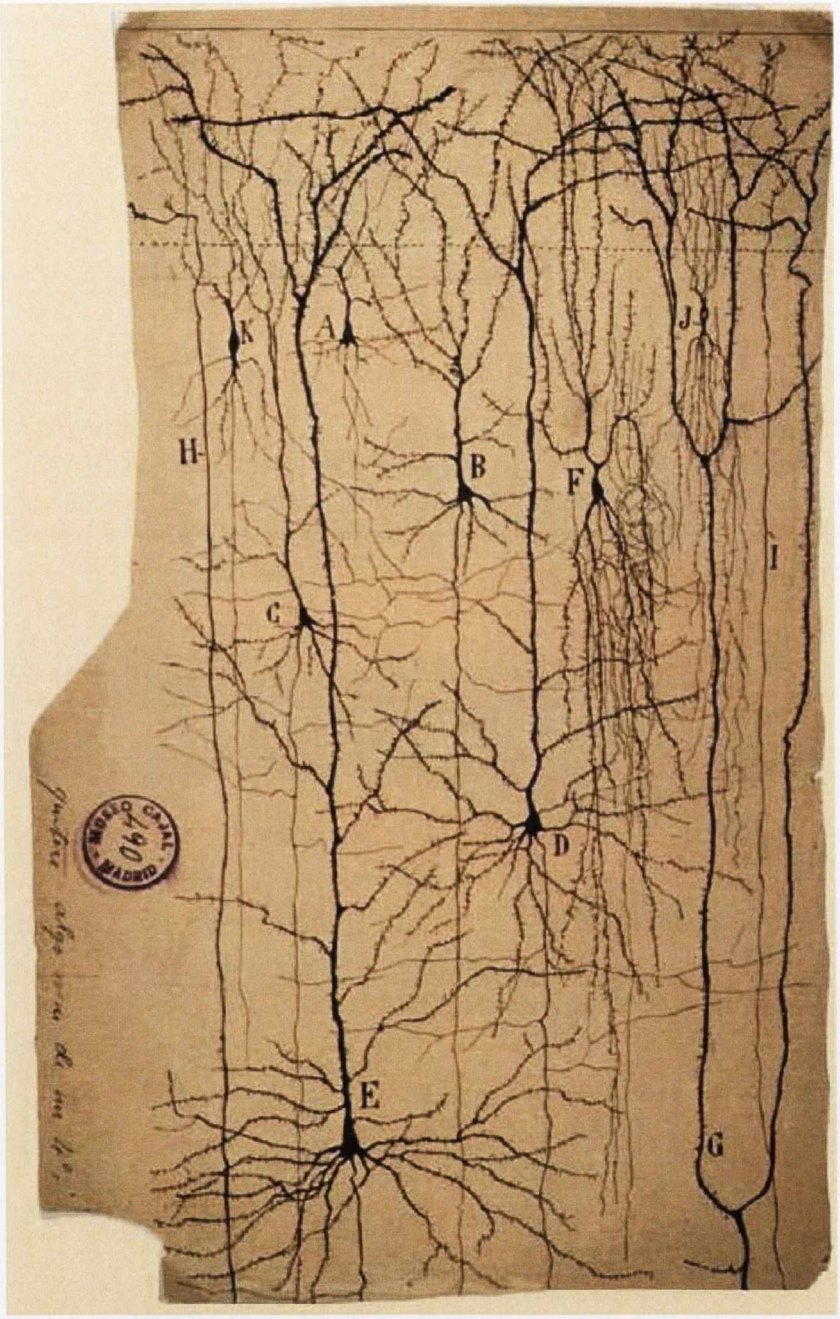

Figure 1.1: An early sketch of pyramidal cells in the human cortex by the neurohistologist Santiago Ramon y Cajal [75]. The brain slice is stained with silver nitrate, making the neural structure visible. The cells labelled A-E are pyramidal cells, so named for the shape of their soma. These are the most common cells in cortex, and the principle carriers of feedforward information. (Picture taken from `kavlifoundation.org`)

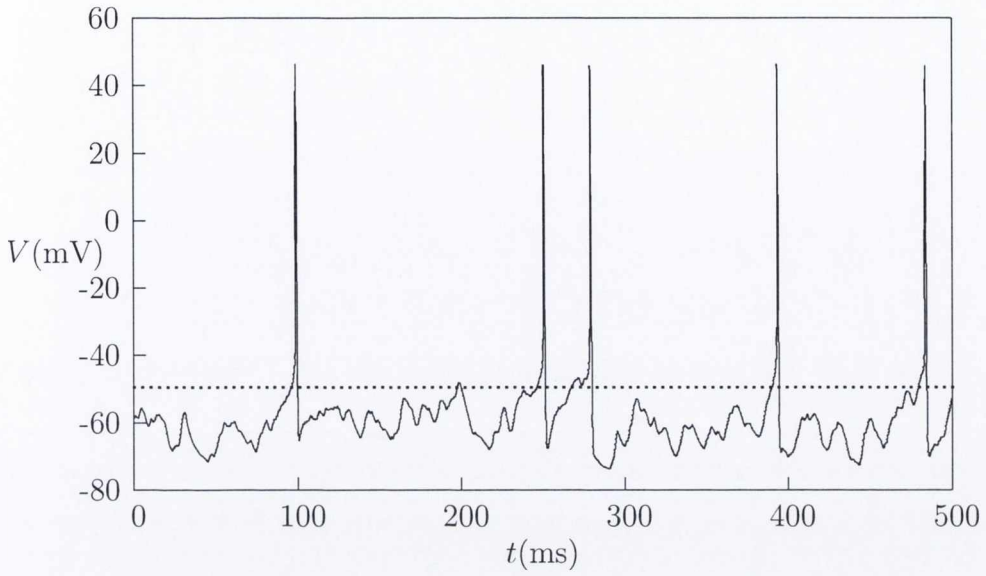

Figure 1.2: A membrane potential plot for a model neuron, produced by numerical simulation of the Hodgkin-Huxley [28] model under stochastic input. The effective threshold value of -50mV is shown.

Anatomically, a typical neuron consists of three principal parts, (see Figure 1.1): the body of the cell, called the soma, which contains the cell nucleus; a large number of small branched projections from the soma called dendrites, which carry input voltage signals; and finally the axon, a single long projection which connects to the soma at a point known as the axon hillock and carries the outgoing voltage spikes. Near its end, the axon typically splits into many branches, which form connections with the dendrites of other cells via junctions known as synapses.

1.1.1 Modelling Neuronal Dynamics

The membrane dynamics of neurons have been widely studied for decades, begining with the seminal work of Hodgkin and Huxley in the 1950s [28]. Their work - in which they modelled the voltage dynamics of a single giant axon found in Atlantic squid - described the neuron in terms of an R-C circuit. Here ion exchange across the cell membrane is modelled by voltage-dependant conductances for different ion species, as well as constant leak conductances.

Though highly succesful in describing membrane dynamics of single cells (see

Figure 1.2), the Hodgkin-Huxley model requires the fitting of a large number of free parameters, and is often considered too complicated to be used in network simulations. Accordingly, many more models of varying complexity have been devised to decribe spiking neurons, such as the Fitzhugh-Nagumo [53], Morris-Lecar [44] and adaptive-exponential models [14]. Though differing in their details, these models are united in defining the membrane potential as a time integral of input and leak currents, and in each case, these models allow the membrane to be described by an equivalent capacitive electrical circuit. In this aspect, they share much in common with a far older neuron model, which was developed long before accurate measurement of ion currents or membrane potentials was possible.

The Leaky Integrate-and-Fire (LIF) model, first developed by Lapicque in 1907 [36], describes a neuron membrane as a capacitor and resistor in parallel. Subject to a time dependent polarizing current, the voltage across the capacitor - equivalent to the membrane potential - increases at a rate determined by the capacitance, C, while the current across the resistor simulates the leak of positive charge through the cell membrane. The voltage dynamics of this circuit are given by the differential equation

$$C\dot{V}(t) = I(t) - \frac{1}{R}(V(t) - V_r), \tag{1.1}$$

where R is the resistance, and V_r is the resting potential. The equivalent circuit diagram is shown in Figure 1.3.

Clearly, neural spiking dynamics cannot be described by such a simple first-order model. However, spikes are short-lived events, and for many practical purposes the voltage trajectory during a spike is of little interest compared to the timing of the spikes . Hence, we can simulate a spiking neuron using this model by the addition of a simple non-linearity: whenever the voltage exceeds a certain threshold value, a delta function spike is produced, and the voltage reset to the resting value. Remarkably, though developed over one hundred years ago, this simple model is quite effective in reproducing the spike timing and firing rate characteristics of sensory neurons, and

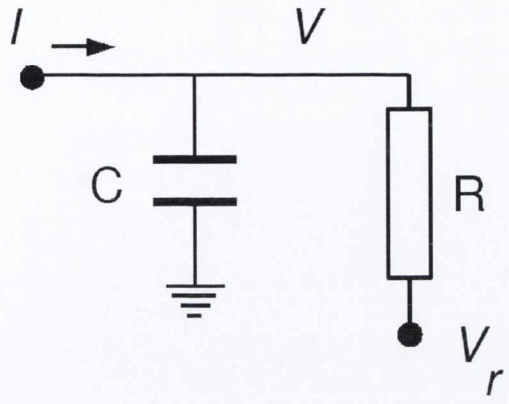

Figure 1.3: Equivalent circuit diagram for the Lapicque's leaky integrating neuron. The membrane potential, V is described by the voltage across the capacitor, C.

is still used to describe neuron dynamics in many recent network models [73, 29, 43].

In chapter 4 we make use of a thresholded integrating neuron model largely based on the LIF. Though our model omits the spiking mechanism entirely in favour of a rate based approach (see Section 1.3.1 on rate coding), the treatment of the neuron as a leaky integrator and the use of a threshold function to simulate firing are directly inspired by their use in the Lapicque's original model.

1.2 Synapses, Neurotransmission and Synaptic Plasticity

Neuronal action potentials are the carriers of both feed-forward and feedback information in the brain. However, such signals are not simply replicated or linearly passed on from neuron to neuron. Rather, information transfer is mediated by chemical signalling across synaptic junctions.

The arrival of an action potential at an axon terminal causes the activation of ion channels in the terminal membrane. This results in the release of vesicles containing neurotransmitting chemicals into the gap between the axon and the post-synaptic dendrite. The neurotransmitter disperses into this gap, known as the synaptic cleft,

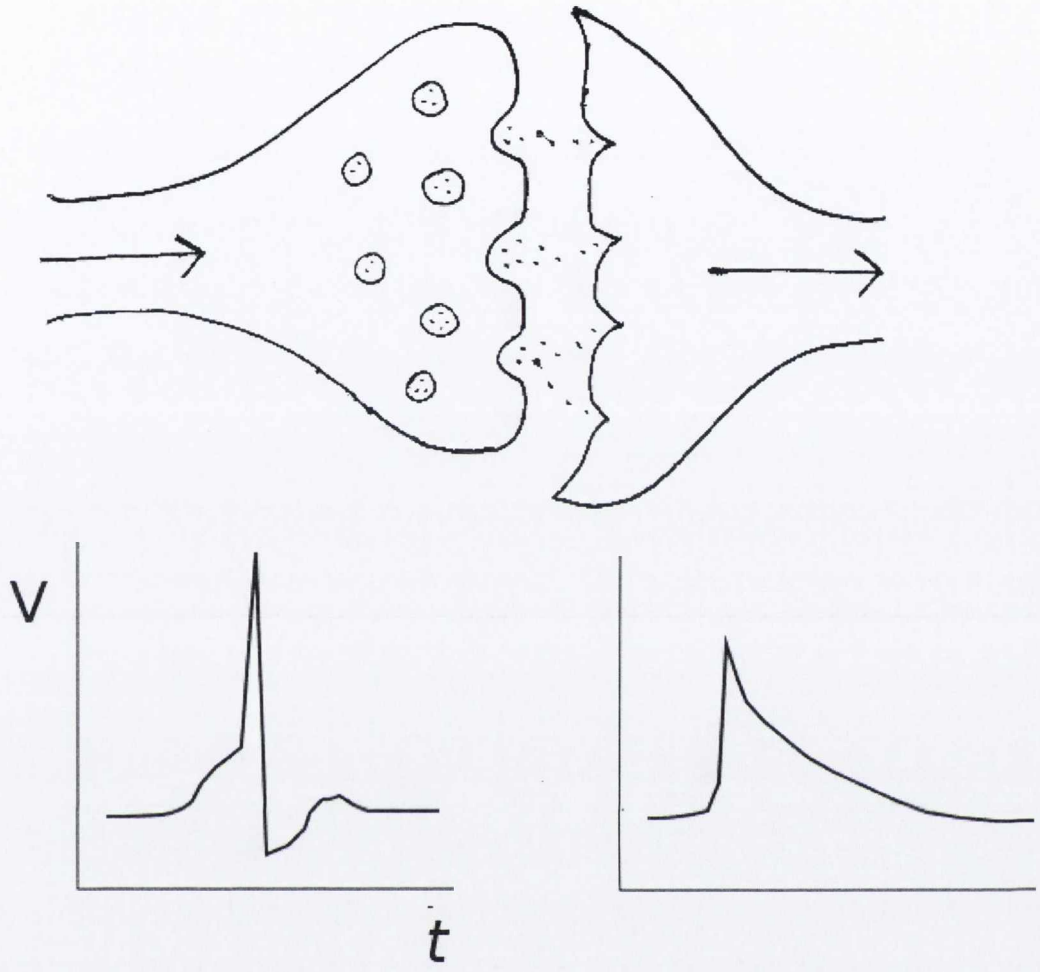

Figure 1.4: Schematic of neuronal synapse. The arrival of an action potential at the axon terminal stimulates the release of neurotransmitters into the synaptic cleft, which then bind to receptors in the post-synaptic dendrite. Temporal profiles of pre-synaptic spikes and post-synaptic potential are shown.

before bonding to receptors in the dendrite membrane. This in turn creates a change in the membrane potential of the post-synaptic cell dendrite, known as the Post-Synaptic Potential (PSP). Incoming PSPs from many dendrites are integrated in the post-synaptic cell soma and, if sufficiently large, may result in the creation of an action potential. This action potential propogates along the axon, and in turn stimulates further synaptic activity. Figure 1.4 shows a schematic of this process of synaptic neurotransmission.

1.2.1 Synaptic Weights

In general, the likelihood of spike generation in a post-synaptic cell is dependent on the number and timing of pre-synaptic spikes, as well as on the strength or effectiveness of the individual synaptic connections. The information conveyed by a neural spike, and indeed, any computation that may be thought to be performed by a neuron, are determined by the relative importance of the cell's synaptic inputs. It is useful, therefore, to define a measure of the connection strength of a particular synapse.

The activity of a synapse is determined by a large number of variables; the temporal profile of pre-synaptic spikes, the concentration of vesicles in the axon terminal button, the availability of receptor sites on the post-synaptic dendrite and so on. In functional terms it may be more useful to reduce the description of a synapse's effectiveness to a single quantity; namely, the conditional probability that a pre-synaptic spike will result in a post-synaptic spike. In practice, however, it may require inputs from several pre-synaptic cells to precipitate a threshold crossing in the soma of the post-synaptic cell, and so this probability can be defined only with reference to all other inputs. Instead, a synapse is perhaps best described by the effect of a pre-synaptic spike on the post-synaptic potential. We quantify this effect by a single scalar parameter, the synaptic weight, w.

The activity of a post-synaptic cell is influenced by that of all the neurons which synapse onto it, and so we might express the evoked post-synaptic potential as a weighted sum over the activities of the pre synaptic cells [17]:

$$\dot{u}(t) = \sum_i w_i a_i(t) \tag{1.2}$$

Where a_i represents the activity of the i^{th} presynaptic cell, and is approximated by

$$a_i(t) = \sum_s \delta(t - t_s) \tag{1.3}$$

where $\{t_s | s \in \mathbb{N}\}$ is the set of spike times. If time is sampled discretely, this becomes

$$(\Delta u)_t = \sum_i w_i a_{it} \tag{1.4}$$

and the synaptic weights are found using the standard least squares solution. In other words, the change in PSP is produced by a linear filtering of a set of inputs, $a_i(t)$, where the filter components are given by the synaptic weights. The use of linear filters to describe neuronal behaviour is discussed at much greater length in Chapters 2 and 3.

In the case where the pre-synaptic inputs, a_i in Equation (1.2) are independent, the weights w_i can be thought of as the causal correlations between the firing of each presynaptic cell i and the PSP.

Importantly, this formulation allows for negative weight values. The firing of a cell j whose synaptic weight w_j is negative will tend to inhibit or supress the firing of the post-synaptic cells. Inhibition is believed to play an important role, both in neural computation and in preventing runaway excitation in neural circuits. It is estimated that up to one third of neurons in certain cortical areas perform an inhibitory function [59].

1.2.2 Synaptic Plasticity

Though it is convenient for many purposes to describe synaptic weights as constants, in fact they vary in time as a function of both pre- and post-synaptic activity. Such changes can result from a range of biochemical effects, and are generally classified according to the timescale on which they occur.

Short term plasticity effects such as Short Term Depression (STD), or synapse fatigue, are observed on the millisecond to second timescale of action potentials and neural spike trains [1]. Such changes are generally attributed to short term changes in the concentration of neurotransmitters in the synapse, and in particular, synaptic

fatigue is often considered to arise from a depletion of vesicles in the synapse during periods of rapid firing. More recent studies have shown that synaptic depression may in fact serve a computational purpose in allowing for the encoding of information in slow-firing cells [1, 15]. Nonetheless, such effects are generally temporary, with the synapse effectiveness rapidly returning to normal after a period of silence. As a result, in many simple models of neurotransmission, such effects are ignored, or treated as minor perturbations around an constant weight value. However, many models exist which take account of such effects, and recent spike-metric techniques explicitly model the effects of short term depression [31].

Perhaps of more interest from the point of view of neural information are long term plasticity effects. These changes may take the form of increases in effectiveness, known as Long Term Potentiation (LTP) or decreases known as Long Term Depression (LTD). These effects are believed to constitiute the principal mechanisms of learning and memory formation in the brain.

The profile of a neural action potential appears unchanged over the lifetime of the cell. Likewise, the threshold potential for spike generation is considered to be a constant for a given cell. Therefore, long term changes in the functional behaviour of a neuron must result from changes in the only significant remaining variables, the synaptic weights.

Changes to a cell's incoming synaptic weights result in changes in the information encoded by that cell's activity. The firing of the cell may indicate the presence a particular pattern of activity in the pre-synaptic neurons. In the case of sensory neurons, this may correspond to a particular image or sound, while the pre-synaptic cells encode the presence of image components. A change in the pattern of synaptic weights results in a different image or sound being encoded by the post-synaptic cell. Likewise, single neurons may perform simple computational tasks. If we imagine a cell with only two non-zero input weights, both pre-synaptic cells may need to fire to precipitate firing of the post-synaptic cell. In this case, the post-synaptic neuron

performs a coincidence detection computation on its inputs. Again, changes in the cell's synaptic weights result in a change in the neuron's computational function.

It is generally believed that changes in synaptic weights are the result of correlations in the firing of pre- and post-synaptic neurons. In this way, the behaviour of a neural system is adapted as a function of the information which it encodes. This process is more familiarly known as learning.

1.2.3 Hebbian Learning

If learning and memory arise from synaptic plasticity, then it is obviously important to derive models for synaptic change. Perhaps the first attempt to describe this behaviour was by Hebb [26]. His hypothesis, often incorrectly summarised as "Cells that fire together, wire together" in fact proposed a more insightful principle which still underlies most studies on plasticity. Hebb's rule can more accurately be stated as

> *If the firing of cell* **a** *results in the firing of cell* **b**, *then the connection from* **a** *to* **b** *should strengthen.*

This rule, which requires a causal link between the firing of the pre-synaptic and post-synaptic cells more accurately reflects modern theories about plasticity mechanisms than the many correlation based rules inspired by Hebb's hypothesis.

There exist many models of synaptic changes, varying greatly in their degree of complexity. In recent times, the most popular synaptic learning mechanisms are those based on Spike Timing Dependant Plasticity (STDP) [9, 10]. These models describe synaptic change as a function of minute differences in the timing of pre-synaptic and post-synaptic spikes. Such rules are in keeping with the spirit of Hebb's conjecture, and account for both positive (LTP) and negative (LTD) weight changes.

A commonly used STDP rule is of the form [63]:

$$\Delta w_{ij} = \sum_f \sum_p W(t_j^f - t_i^p) \tag{1.5}$$

$$W(t) = \begin{cases} A_+ \exp -t/\tau_+ & t > 0 \\ -A_- \exp t/\tau_- & t < 0 \end{cases} \tag{1.6}$$

where w_{IJ} is the synaptic weight from cell i to cell j, $\{t_j^f | f = 1, 2 \dots\}$ is the set of spike times for post-synaptic cell j, and $\{t_j^p | p = 1, 2 \dots\}$ are the pre-synaptic spike times.

Though STDP rules are highly effective in modelling synaptic change where precise spike timings are known, it is often more convenient to model neural systems in terms of neural firing rates (see Section 1.3.1 on rate coding). In these cases, spike timing information is unavailable, and we must fall back on acausal correlation based methods. The simplest Hebb type rule which can be applied in these cases is simply given by

$$\dot{w}_{ij} = \frac{1}{\tau} a_i(t) a_j(t) \tag{1.7}$$

where $a_i(t)$ and $a_j(t)$ are the instantaneous neural activity rates. This rule, though popular in rate-based models, has two significant drawbacks. Firstly, since firing rates are non-negative, it does not model synaptic depression (LTD). Secondly, the update rate is bilinear in the firing rates, which can result in positive feedback and exponential growth in the synaptic weight.

Long-term depression can be more effectively achieved in so-called covariance models, in which the post-synaptic firing rate is replaced by its deviation above or below a threshold, such as in the Bienenstock, Cooper and Munro model (BCM) [11]. However, such models still suffer from the possibility of unbounded growth.

A simpler solution, which models both positive and negative weight changes and places an upper bound on weights, is to use a simple correlation rule as in Equation (1.7) and then normalise the synaptic weight vector at each update. In effect, this

places an upper limit on the magnitude of the excitatory input a cell may receive, and ensures that synapses with proportionally lower correlation functions will be weakened. Though lacking somewhat in realism, such a model does reflect certain biological limitations. In particular, the finite supply of neurotransmitters and limits on ion flow through gated channels do in fact place practical upper limits on the the effectiveness of synaptic connections.

In cases where fast simulation of network interactions are of greater concern than highly accurate modelling of individual synapses, this normalised correlation rule is often highly convenient, In fact, in chapter 4 of this thesis we make use of such a rule when simulating networks of sensory neurons. Output cells in our model have activity rates $r_j(t)$ and receive excitatory input from a layer of input cells with activities $s_i(t)$. Synaptic weights are updated periodically using the Hebbian correlation rule

$$\Delta w_{ij} =< s_i(t) r_j(t) >_t \tag{1.8}$$

where the average is taken over the time since the last update. Weights are then normalised after each update:

$$w_{ij} \rightarrow \frac{w_{ij}}{\sqrt{\sum_i w_{ij}^2}} \tag{1.9}$$

Using this rule, the total potential input to each output cell remains constant, but the weighting of this input is distributed among the input cells in proportion to the correlation of their firing rates with that of the output cell. As we shall see in chapter 4, this very simple learning rule is sufficient to reproduce several features of the behaviour of sensory cortex in a two-layer network model.

1.3 Sensory Coding

The goal of mathematical neuroscience is to understand and model the means by which information is encoded, transmitted and stored in the brain. An obvious approach to this problem is to study how external sensory information is translated into neural signals. Electrophysiological methods allow us to record directly from cells in the sensory pathways during exposure to an experimentaly controlled stimulus. Hence, we can directly compare a visual or auditory signal with its neural representation in the brain. The mapping between a stimulus and its resulting cortical activity is known as a neural code. In this thesis, we will address three related questions which arise in the study of neural coding of sensory inputs:

1. How do we characterise the response properties of a neuron under stimulation?

2. What determines these properties, and how do they relate to the theory of efficient coding?

3. How are efficent codes implemented in a neural system?

These questions are explored in chapters 2,3 and 4 respectively. To approach these problems we make use of two common concepts in neural coding. Firstly, we make use of the simplifying, though somewhat inaccurate assumption that sensory inputs are encoded in neural firing rates (rate coding). Secondly we adopt the hypothesis of efficient coding, that is, the assumption that the brain encodes stimuli in a way that is energetically efficient. These concepts are further explained below.

1.3.1 Rate Coding

Signals in the nervous system are transmitted in the form of changes in neuronal membrane potentials. Synapses, which form the junctions between neurons, are generally activated only by large voltage spikes in the form of action potentials, and so we can conclude that sub-threshold variations - that is, variations in membrane potential which do not result in a spike - feed forward little, if any information

to post-synaptic cells. Hence, we proceed on the basis that neural information is encoded entirely by the action potentials. Furthermore, the action potentials of any one neuron are generally stereotypical, displaying little or no variation in temporal profile or amplitude. Hence we can assume that the signal information is contained solely in the timing and frequency of spikes. However, it is still unclear exactly how spikes encode sensory information.

Neural spiking is an inherently noisy process, and a given neuron will rarely respond identically to several repetitions of the same stimulus (see Figure 1.5). Hence, neural firing is often treated as a stochastic, Poisson-like process, where spikes occur randomly with a probability determined by an underlying firing rate. From this viewpoint, known as rate coding, the timing of individual spikes is considered to be unimportant, and the neural representation of a stimulus is described completely by the firing rates.

The assumption of rate coding is in many ways a vast over-simplification, and in fact, it is well known that certain neural functions, such as plasticity, rely heavily on spike-timing effects [54, 63, 64]. Nonetheless, in many systems, and particularly in those areas where neurons display a high degree of variability, rate coding allows an intuitive first-order characterisation of the stimulus-response properties which can explain much of observed behaviour of these cells [3, 2, 32]. This is particularly true in the auditory and visual systems, which encode stimuli which generally vary on a timescale significantly longer than that of neural spiking.

Moreover, since such stimuli are continuously varying, it is reasonable to consider them as being encoded in continuously varying rate functions, rather than in sequences of discrete spike times.

There also exists a more direct biological justification for the use of rate codes, in the form of synaptic transmission. Although action potentials can be reasonably described within a neuron as discrete, instantaneous events, their effect on a post-synaptic neuron is mediated by the release, dispersion and binding of neuro-

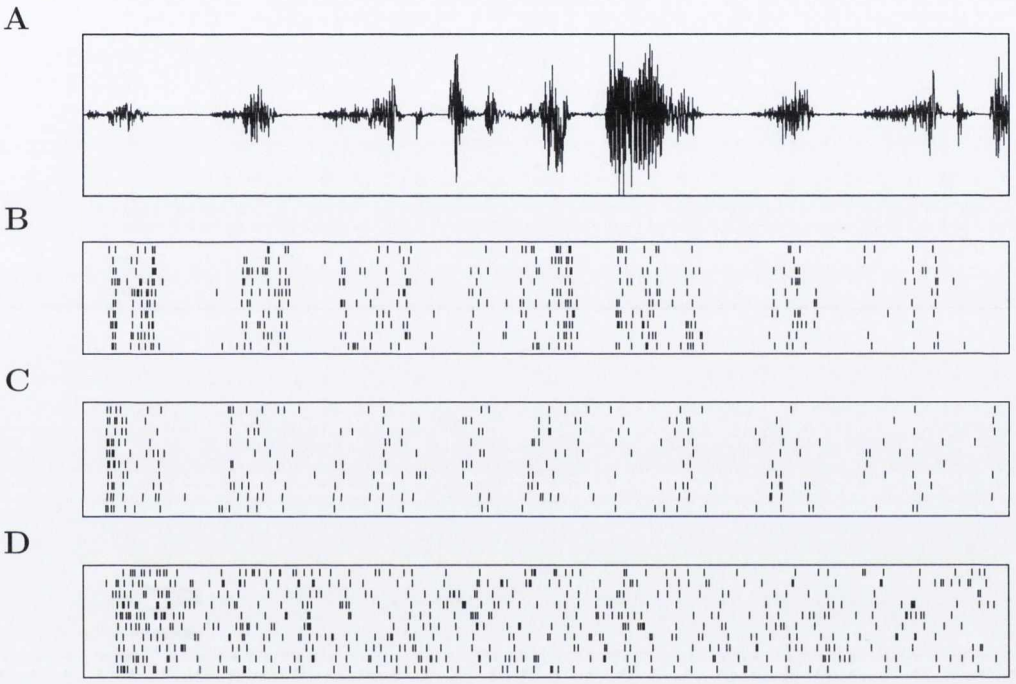

Figure 1.5: Raster plots showing response of auditory neurons in the zebra finch to a known stimulus song recording. **A** shows the one second wave form of the stimulus song. In **B**, **C** and **D** each box corresponds to ten responses at a single recording site to repeated stimulation with the song shown in **A**. Time increases along the horizontal axis and the results of ten different trials are stacked vertically. A spike is marked by a vertical dash. Site **C** shows a typical response, with a wide variation in spike trains between repetitions, site **B** is unusually reliable, though responses still vary significantly, while site **D** is unusually poor, showing very little coherence. Figure adapted from paper by C. Houghton, 2009 [30].

transmitters in the synapse, which are best modelled as continuous time-dependent processes. Hence, although information in a neuron may be carried by spikes, this information is transmitted across a synapse in the form the continuous rate function describing the uptake of neurotransmitter in the post-synaptic dendrite. This notion is commonly applied in Van Rossum type spike metrics, where a spike train is described by a synaptic rate function [69, 31].

In this thesis, we discuss sensory processing largely in terms of rate coding. This has the advantage of allowing us to formulate neural codes in terms of time dependent linear filters (see Chapter 2: *Spectro-Temporal Receptive Fields*) and also allows us to construct efficient codes for natural stimuli by decorrelating neural responses, as discussed in Chapter 2. Though by no means a complete description of a neural signal, the firing rate is believed to carry much of the stimulus information in primary sensory areas. Importantly, the use of a firing rate code does not preclude the possibility that further information is contained in the precise spike times, and so it may be useful to think of rate coding as a first approximation to a true neural code.

1.3.2 Optimal Coding

It is commonly - indeed, almost universally - believed that sensory information is encoded in the cortex in a way that is metabolically efficient [6, 47, 50, 49, 5]. That is, the neural code used in the brain is that which requires the lowest energy consumption, consistent with a sufficiently high processing speed and capacity. This is hardly surprising, and indeed the brain does appear to be remarkably efficient, running at an average power as low as $20W$ according to some estimates. By comparison, a typical desktop computer may consume around $100W$ while idle.

It is natural to wonder how this efficiency might be reflected in the firing patterns of sensory neurons. A neuron's metabolic consumption is highest during, or just after, the creation of a spike, though neurons do consume energy even when idle in

order to maintain their membrane potentials. Hence, energy efficiency is achieved by reducing the total number of action potentials produced, and one might therefore conclude that:

> *The most efficient encoding is that in which the smallest number of spikes are required to encode the stimulus.*

However, an optimal neural code must also reflect the functional goals of the sensory system, and so we can refine this notion further in light of some computational considerations: For a given layer of sensory cortex, the encoding which requires the smallest number of spikes may involve relatively large number of neurons each firing a small number of spikes, or a small number of neurons firing more frequently. An encoding in which a larger number of cells are weakly active - often called a dense representation - can be computationally difficult for higher layers to decode, and may in fact result in higher rates of firing in other layers. Furthermore, such an encoding is more susceptible to contamination by noise, since the firing rate for some cells may be close to the cell's background firing rate. So, for a given level of activity across a cortical layer, an encoding in which a small number of neurons are strongly active may preserve more of the stimulus information than a more distributed representation, and may also result in lower levels of activity in subsequent layers, thus increasing overall efficiency. In light of this, we may augment our statement from above to read:

> *The most efficient encoding is that in which the smallest number of spikes are required to encode the stimulus and, for a given total activity level, the smallest number of cells are simultaneously active.*

This is, in essence a re-statement of an idea known as *Barlow's Efficient Coding Hypothesis* [6] which is at the heart of most common conceptions of primary sensory coding, and is believed to be one of the principle determinants of auditory and visual codes [47, 38, 71, 24].

Metabolic efficiency, however, is not the only factor in determining optimal sensory codes. As mentioned above, a neural encoding can often be described in terms of its density, that is, the proportion of cells which are simultaneously active. This has a considerable influence on the representational and storage capacity of the system. Low-density or local codes can be rapidly and easily decoded, and can allow the concurrent representation of several distinct stimuli. However, such codes place a severe limit on total number of possible stimuli which can be represented by the system. In the low density limit where only one cell is required to encode each stimulus, the system can represent only as many stimuli as there are cells. In addition, the use of completely distinct representations for even quite similar stimuli may hinder the formation of associative connections in the cortex, which are essential for many cognitive tasks.

By contrast, a highly dense code provides a much larger representational capacity, but also brings with it certain deficiencies. In particular, very dense codes are computationally costly to decode, and cannot represent distinct simultaneously occuring events. It is therefore necessary to find a code which interpolates between these extremes.

In chapters 3 and 4 of this thesis we will discuss at some length the notion of Sparse Coding [23], which can be cosidered as a trade off between these regimes, while also generating metabolically efficient representations of highly structured data. Sparse coding neurons are associated with important, commonly occuring stimulus components which are both highly statistically independent and span the relevant stimulus space. This organisation produces efficient, low-density representations of important natural stimuli, while also allowing dense representations of rarer events.

1.4 Songbirds and the Coding of Natural Sounds

Though we often discuss sensory coding in terms of efficiency, it is important to remember that the goal of primary sensory areas is not simply to find a low-energy representation of a signal, but also to ensure that the signal so encoded contains a high degree of relelvant information. The natural auditory environment is to a large extent an ocean of noise from which the auditory system extracts and amplifies relevant or important sounds. The definition of a set of *important* sounds raises many questions in itself, but is most often thought of in evolutionary terms as the set of sounds whose detection is advantageous to the animal in question: snapping or cracking sounds which may indicate the presence of a predator, or animal noises and vocalisations, which may suggest opportunities for procreation, and so on. The auditory system must extract these features from an often cacophanous acoustic environment, while less relevant features are filtered out. It follows, therefore, that the response properties of auditory neurons are adapted to the particular structures and statistics of these important natural sounds [18, 41, 42]. Hence, if we wish to study the behaviour of an auditory system, it is important to ensure that the system is responding to a stimulus which is considered relevant to the animal in question, so as to ensure that the stimulus is, in fact, being encoded in the auditory cortex, and not discarded as irrelevant noise.

Auditory Coding is commonly studied in songbirds. Although songbirds and humans differ greatly in their neurobiology, the songbird auditory system is to a large extent functionally analagous to that of humans, and so we can draw many useful conclusions about auditory processing which are relevant to humans, and indeed to most hearing animals. Importantly, in the case of songbirds, there exists an obvious and easily isolated ensemble of important stimuli, in the form of the birds' songs themselves. In chapters 2 and 3 we discuss the characterisation of auditory neuron responses and the calculation of optimal auditory codes using data gathered from experiments performed on male zebra finches. The zebra finch is a

songbird native to central Australia, and is a popular choice of subject in auditory studies. The zebra finch auditory system is highly adapted to encode conspecific song, and in fact large parts of its auditory pathway are believed to be adapted to the identification and classification of songs.

1.4.1 Avian Auditory Pathway

In chapters 2 and 3 of this thesis we discuss the encoding of natural sounds in the zebra finch auditory forebrain. Zebra finch is widely used as a model organism in studies of avian learning and song production, and so the structure of the auditory forebrain is often discussed in the context of its connections to the rather complicated song system and associated motor pathways. Here however we are concerned with the encoding of incoming auditory signals. Hence, though we make use of birdsong stimuli - indeed, it is our hypothesis that the zebra finch auditory forebrain is specifically adapted to encode song - we do not concern ourself here with any aspects of the vocalisation system, and instead focus solely on the ascending auditory pathway.

As in mammals, the neural processing of auditory signals begins with the so-called hair cells of the cochlea, or inner ear. Hair cells are mechanically stimulated by longitudinal vibrations on the fluid filling the cochlea. Though hair cells do not themselves produce spikes, mechanical perturbation of the inner hair cells results in the depolarization of auditory nerve fibres, resulting in the creation of action potentials. The arrangement and connectivity of hair cells produces a number of different effects, among them the amplification of weak sounds by the outer hair cells. However, two characteristics or cochlear response are of particular interest. Firstly, the response of hair cells to auditory stimuli is thought to be logarithmic in stimulus amplitude, and secondly, different hair cells are known to be selective for particular sound frequencies, and so the cochlea is believed to perform a spectral decomposition on incoming stimuli. Here, as in many studies of auditory coding,

we mimic these effects by using a spectrographic representation of sound stimuli, in which the stimulus is represented by the log amplitude of the sound waveform in each of several frequency bands.

From the cochlea, signals then feedforward, via the auditory nerve, to the brainstem nuclei NM (nucleus magnocellularis) and NL (nucleaus laminaris). NL is the site of convergence of binaural signals, and it is believed that sound localisation computations occur here, through the use of coincidence detection circuits [35]. Intepretation and recognition of signal content, however, is performed at higher level, in the analogue of auditory cortex.

From NL, auditory signals feed forward to nucleus ovoidalis, the avian equivalent to the thalamus in the human brain. As with the thalamus, ovoidalis is believed to function primarily as a relay between peripheral sensory areas, and their corresponding forebrain structures. However, also in common with the thalamus, the large number of feedback connections to ovoidalis suggest that it performs other subsidiary functions in sensory processing, perhaps forming part of a system for estimating the error in cortical representations of stimuli.

Ovoidalis projects principally into the Field L region of the avian forebrain [58, 35, 74]. Field L is the functional analogue of primary auditory cortex, and as such shares many features in common with auditory and visual cortex in mammals. In particular, it is overcomplete, meaning that it contains many times more neurons than the areas which project into it, and it displays tonotopy [76, 45], meaning that neighbouring cells respond to spectrally and temporally similar stimuli.

Field L is often studied as one of the areas which projects forward into the vocalisation system of songbirds [20, 62, 46]. However, it is important to note that the song production is not the primary role of Field L, and that Field L is found even in bird species which do not sing [35]. In chapters 2 and 3 of this thesis we analyse and model the response of Field L neurons to natural stimuli. We do this with refernce to electrophysiological recording taken from zebra finch Field L by

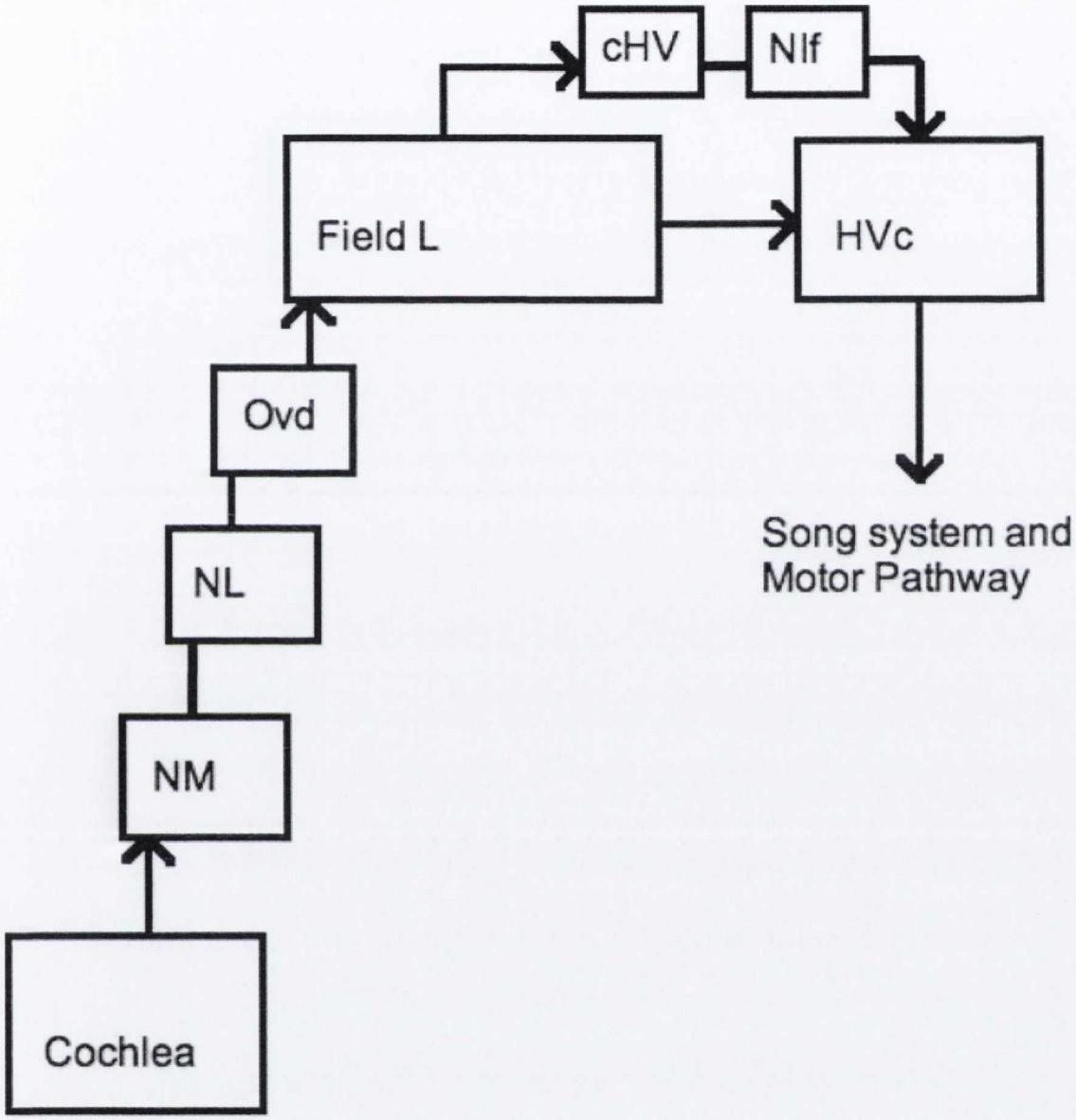

Figure 1.6: Schematic of the ascending auditory pathway in the zebra finch brain. Sound waveforms are converted into neural signals at the cochlea, and projected through the brainstem nuclei NL and NM to nucleus ovoidalis. From there, the signal is projected into Field L, the avian analogue of the primary auditory cortex. Field L representations are projected, via two seperate pathways, into the higher vocal centre (HVC). HVC forms the nexus between the ascending auditory input, and the descending motor pathway and song system.

our experimental collaborators at the Natural Sounds and Coding Lab in Boston University.

1.4.2 Electrophysiological Recording

To gain a significant insight into the function of neurons as computational components of the brain, we must study their behaviour, not just in isolation, but also while functioning as part of a much larger neural system. This can most usefully be achieved by recording the voltage dynamics of active neurons within a living brain. Furthermore, in order to determine what function is performed by a particular neuron or brain structure, it is important to know what input the system is receiving, and what information it conveys. The study of primary auditory and visual systems satisfy both these requirements: In a laboratory environment it is easy to control and record the sensory stimuli to which the system is exposed, while the poitioning of auditory and visual areas in the vertebrate cortex makes them convenient sites for electrophysiological recordings, and in particular, recordings of intra-cellular and extra-cellular potentials.

Neuronal spiking is caused by ionic currents through voltage dependent gates in the cell membrane, causing changes in charge concentrations, both inside and outside the cell [28]. These changes result not only in changes in the membrane potential, but also in fluctuating potential differences relative to an external source. Under laboratory conditions, it is possible to insert electrodes into the brain, which can be used to measure these fluctuations in potential, both inside and outside the cell.

Ideally, neural recordings are made with the electrode inserted directly into the cell axon. This allows accurate measurement of the internal cell potential, and the detection of spikes without interference from external sources. Unfortunately, due to the size of neuronal axons (generally of order 10^{-5}m) such experiments are extremely difficult and time consuming, and are still considered impractical for most purposes.

Instead, it is most common to use measurements of potential fluctuations in the inter-cellular medium, taken ouside the cell membrane. All neuronal data used in this thesis waas acquired using such extra-cellular recordings.

The principal disadvantage of this method is that fluctuations in extra-cellular potential may result from the activity of several neighbouring neurons, as well as a high degree of environmental noise. Using band pass filters, it is generally possible to remove low-frequency noise from such recordings, allowing the identification of potential spikes corresponding to action potentials. It may then be necessary to classify such spikes according to the neuron which produced them.

Generally, experimentalists aim to position the electrode as closely as possible to a single cell, in which case the action potentials of that cell should be clearly distinguishable within the signal due to their large amplitude. This is not always possible, however, and it is necessary to apply a spike sorting algorithm to extract the component signals. A variety of spike sorting algorithms exist which differ widely in their complexity. The majority of these methods rely on finding distinct spike profiles for different cells, and employ a range of techniques including template matching, clustering algorithms, and wavelet analysis.

Details of experimental conditions and spike sorting methods applied to recordings used in this thesis are given in Appendix A.

1.5 Background and State of the Art

1.5.1 Characterising Neuronal Responses.

The most commonly used model for characterising the responses of auditory cortical neurons is the Spectro-Temporal Receptive Field (STRF) model. The STRF is defined as a linear spectro-temporal kernel, which is convolved with the stimulus spectrogram to give a prediction of the neuronal firing rate. The 'true' STRF for a given neuron is that filter which gives the most accurate prediction. Though

generally employed as a purely linear model, the STRF can also be thought of as the first term in a higher order Volterra series.

Theunissen et al.[65] describe the standard method for calculating the STRF. This algorithm, known as STRFPak, proceeds by normalised reverse correlation: The calculation is first transformed into the space of temporal modulation frequencies (or k-space) by means of the convolution theorem, which transforms the convolution of stimulus and STRF into a simple product of fourier transforms. The STRF is then found by reverse correlation using the least squares method. As is common with reverse correlation calculations, the STRF estimate obtained by this method may be highly contaminated by noise, and so the calculation must be regularised. This regularisation is acheived by singular value decomposition in k-space.

This method of STRF calculation is extremely popular, and has been used succesfully to characterise auditory reponses in a number of studies [65, 58, 40, 39, 74]. However, it also possesses a number of drawbacks: the use of the convolution theorem in this case is ill-justified mathematically, and also requires the smoothing of neuronal firing rates so that the fourier transform may be calculated. Furthermore, the regularisation method used is computationally expensive compared to a regularisation in the temporal domain, and is somewhat lacking in transparency.

David et al. [16] describe an alternative method for STRF calculation. This technique, known as boosting, avoids several of the problems associated with the reverse correlation method. This iterative algorithm describes the STRF as a vector in spectro-temporal space. At each iteration, a single STRF component is increased or decreased by an increment ϵ. The component to be incremented, and the sign of the change are chosen from all possible combinations to give the best improvement in the predictive power of the STRF. To prevent over-fitting to noise in the training data, the calculation as stopped after a number of iterations determined by a stopping parameter. This early stopping condition serves a similar role to the regularisation performed in reverse correlation calculations.

For suitable choices of the step size, ϵ, and a stopping condition, this algorithm matches the performance of the STRFPak method when applied to natural stimuli, and produces a significant improvement in computational efficiency. However, despite avoiding some of the problems associated with the STRFPak method, the boosting algorithm has a number of drawbacks of its own. Most importantly, the performance of the algorithm is highly dependent on the choice of the stopping condition, as well as on the step size, and so this method requires the optimisation of two model parameters, as opposed to a single regularization parameter for the reverse correlation method. As a result, this method is significantly more likely to converge on a highly inaccurate STRF estimate. In addition, this method lacks the intuitive simplicity of the reverse correlation calculation, and, as is the case with the STRFPak algorithm, the precise nature of the regularisation is unclear.

In Chapter 2 of this thesis we will describe an improved method of STRF estimation through reverse correlation which avoids the problems associated with the STRFPak method, but without the introduction of additional model parameters.

1.5.2 Finding Optimal Codes for Natural Stimuli

Having derived a STRF model to characterise the response of auditory neurons, it is natural to wonder what determines the structure and organisation of such STRFs. In chapter 3 we investigate whether zebra finch auditory STRFs are optimised to perform a sparse encoding of con-specific song by calculating a sparse basis for zebra finch songs.

Olshausen and Field [47] have demonstrated that the receptive fields of the primary visual cortex form a sparse code for natural images by calculating such a sparse basis. Using a large ensemble of images samples, they calculated an overcomplete set of optimal basis functions, or image components which allowed an accurate but sparse reconstruction of natural scenes by minimising redundancy in the neural representation. Images were first pre-processed with a regularising low-pass whitening

filter, which normalised the image variance across spatial frequencies, and removed noisy high-frequency contributions. A sparse basis for these whitened images was then calculated by minimising an energy function which combined the error in the image reconstruction with a sparseness cost function, which penalised redundant representations. This method of redundancy reduction can be considered in many cases to be a form of Independent Component Analysis (ICA) [7]

The resulting set of basis functions closely matched the observed receptive fields of neurons in the primary visual cortex, suggesting that the visual cortex utilises a sparse code for natural images.

Algorithms of this type have since been used widely in studies of efficient coding. Vincent et al. [71] made use of a similar algorithm in modelling sparse coding in the early visual system. They constructed a three-layer model of the retina and primary visual cortex and found that the tuning and organisation of both retinal ganglion cells and cortical simple cells could be accurately predicted. In particular, the centre-suround receptive fields of retinal ganglion cells, and the gabor-like tuning of simple cells were found by optimising a sparseness cost function of the type used by Olshausen and Field.

Though such methods have not yet been applied to the auditory cortex, there is evidence to suggest that sparse coding occurs in the early visual system. Lewicki [38] used an ICA algorithm to predict the tuning properties of hair cells in the cochlea of the inner ear. As discussed earlier, the cochlea is believed to perform a spectral decomposition on the incoming sound waveform. This is acheived by temporal filtering in the hair cells.

Lewicki calculated sets of optimal hair cell filters for different sound ensembles, which maximised statistical independence between cells. For a stimulus ensemble comprising a mix of animal vocalisations and environmental sounds, he found a set of predicted optimal filters which closely matched those measured in the auditory nerve fibres. This is suggestive of a sparse code for natural sounds in the early

auditory pathway.

In Chapter 3 we will investigate whether the principle of sparseness extends to the auditory cortex by using a modified Olshausen-Field type algorithm.

1.5.3 Optimal Coding through Network Interactions

Though many studies have found evidence of sparse coding in sensory systems, it remains to be conclusively demonstrated how such codes are implemented on a neural level. Since the neural code in primary visual and auditory areas is known to be highly overcomplete, then for many stimulus there must exist a large number of neurons whose receptive fields closely match the input. There must therefore exist some netowrk non-linearity which prevents all these cells from becoming simulataneously active. Furthermore, many theoretical models of sparse coding, including the ones discussed in Chapter 3 produce representations which are highly unstable with respect to small changes in input, and predict rapid large fluctuations in neural activity in response to smoothly changing inputs. A realistic neural coding model should be expected to produce smoothly varying representations in response to continuously changing stimuli.

Rozell et al. [55] propose a network model in which sparseness arises through lateral inhibitory connections between cells in the visual cortex. In this model each cell receives excitatory input proportional to the similarity between its RF and the input image, in addition to an inhibitory input from all other active cells. The strength of this inhibitory input is given by the product of the activity of the presynaptic cell and the similarity of the two cells' receptive fields. The activity of each cell is determined by a sigmoidal threshold function. These cells then learn the sparse structure of the stimulus through a Hebbian learning rule.

This model successfully yields smoothly varying sparse reconstructions of natural images and, through variation of the threshold function, can be shown to implement a wide family of optimisation algorithms, including Matching Pursuit and Basis

Pursuit De-Noising. The model rapidly sparsifies the network response through inhibitory connections whose strength is determined by the Gram matrix of the cells' receptive fields. However, this connectivity is imposed globally in a biologically unrealistic manner. In order to convincingly demonstrate a network mechanism for sparse coding, such connectivity must develop in an unsupervised manner through synaptic plasticity. Furthermore, this model makes use of direct, instantaneous inhibitory connections between excitatory cells. In reality, such connections must be mediated by inhibitory interneurons, with their own internal dynamics.

Rather than impose inhibitory connections through a global learning rule, it is perhaps more realistic to make use of a model in which such connections might develop in a self-organised manner. Self-organisation is a well observed property of networks which obey Hebbian learning rules, and was famously demonstrated by Song and Abbott [63] in the case of STDP. They simulated a network of leaky integrate-and-fire neurons with local excitatory and long-range inhibitory connections. This network rapidly formed highly organised topographic maps of the input space.In Chapter 4 we propose a network model which learns sparse coding through the development of topographic organisation in a two-layer network of excitatory and inhibitory cells.

Chapter 2

Spectro-Temporal Receptive Fields

The receptive field (RF) [25, 32] has long been used to characterize the response of sensory neurons, particularly neurons in the visual system, using a linear model of the neural firing rate. The RF is defined as the linear filter h which, when applied to a stimulus image, gives the best prediction of the firing rate response. For a two dimensional image, I_{xy}, the response prediction is given by

$$\tilde{r}(t) = \sum_{x,y} h_{xy} I_{xy}(t) \tag{2.1}$$

The RF filter, h_{xy} can be shown to be equal, up to a multiplicative constant, to the stimulus which elicits the maximal response from the cell [17].

While neuronal signalling is most likely too complicated to be typified by a rate response and a linear model can only ever be an approximation, RF models are rather successful and seem to apply in one form or another across many sensory systems.

Linear models of neuronal response have been in use for over 80 years, beginning with the work of Adrian and Zotterman in 1926 [2]. They studied the response of somatosensory nerve fibres in muscle tissue from which weights were suspended.

They found that the firing rate of somatosensory neurons in the muscle tissue varied linearly with the mass of the applied weights.

In the intervening years, linear rate models have been successfully applied to a range of sensory systems, perhaps most famously in the mammalian visual cortex. Hubel and Wiesel [32] developed a receptive field model to describe the stimulus response properties of neurons the the primary visual cortex of cats. A large number of subsequent studies have quantified visual cortex responses in terms of receptive fields [33, 19, 51, 47, 56, 60].

In addition, RF models have been generalised to quantify responses to time-varying stimuli, particularly in the auditory system [3, 21, 67, 65]. Here we describe such a model, known as the Spectro-Temporal Receptive Field [65, 58, 40, 39, 74], which is commonly applied to neurons in primary auditory areas. The STRF describes a neuron's sensitivity to both the spectral and temporal structure of the stimulus. Hence, in order to apply such a model, we must first derive a spectrographic representation of sounds.

2.1 Spectrogram

In the visual RF model, an image is labelled by the spatial indices x and y, corresponding to the two spatial dimensions. An unprocessed auditory signal, however, has no equivalent spatial dimensions. Rather, auditory signals are canonically described by the sound pressure waveform, a single valued function in time. However, it is known that the frequency tuning of cochlear hair cells results in an effective spectral decomposition of sound in the auditory system. It is usual therefore, in studies of auditory coding, to mimic this effect by using a spectrographic representation of stimuli, where the sound waveform is replaced by a vector valued function of time, whose components correspond to the signal amplitudes at different frequencies.

Here, auditory waveforms $s(t)$ are represented by a set of narrowband signals

$\{s_f(t)\}$, where $s_f(t)$ is the deviation from mean of the log-amplitude of the stimulus in the frequency band f. The use of the amplitude logarithm is thought to model the logarithmic response of hair cells.

The spectral decomposition can be achieved either by use of a time windowed Fourier transform, or equivalently, by band pass filtering in the frequency domain. Spectrograms used here are produced by filtering in the frequency domain using a bank of overlapping gaussian filters seperated by one standard deviation. This filtering method has been shown to allow extremely accurate signal reconstruction [66, 67].

The number of frequency bands into which a stimulus is decomposed is determined by the filter width, and hence by the frequency resolution of the spectrogram. The spectral and temporal resolution of spectrographic signals are governed by the signal processing uncertainty principle

$$\Delta f \Delta t \geq \frac{1}{2} \int |s(t)^2| dt \tag{2.2}$$

In short, this means that any increase in spectral resolution of the signal must be traded off against a decrease in temporal resolution. Hence, we must choose an appropriate filter width which best preserves the stimulus structure. Studies of the zebra finch auditory system [66] suggest that recognition of decomposed signals in higher auditory centres is best preserved at a spectral resolution of 250Hz. Unless otherwise stated, all sound spectrograms used here are composed of 32 frequency bands of width 250Hz, spanning the range from 250Hz - 8000Hz.

2.2 Spectro-Temporal Receptive Field Model

Using the spectrographic representation of sound just described, an RF model can be applied to neurons in primary regions of the auditory pathway. In this model the stimulus spectrogram is convolved with a kernel to give a prediction of the neuronal

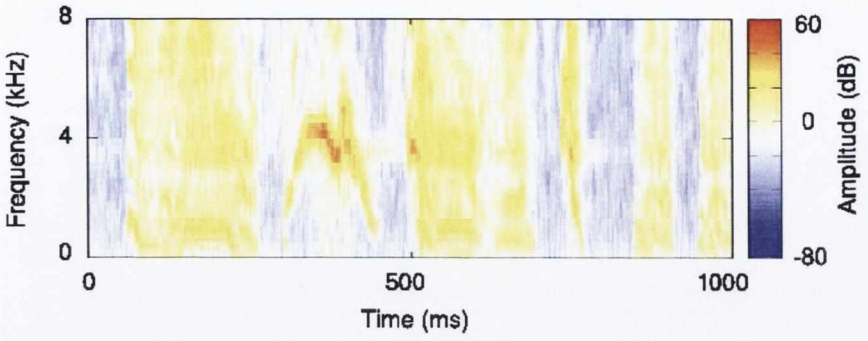

Figure 2.1: A sample spectrogram of one of our zebra finch song recordings. Amplitude is shown on a colour scale from blue (lowest) to red (highest) The temporal resolution of the spectrogram is 1ms, and the spectral resolution is 250Hz

firing rate. The predicted firing rate $r'(t)$ is given by

$$r'(t) = \sum_{f=1}^{n_f} \int_0^T h_f(s) s_f(t-s) ds \tag{2.3}$$

where $s_f(t)$ is the stimulus spectrogram at time t and frequency band f, n_f is the number of frequency bands in the spectrogram and $h_f(\tau)$ is the kernel which is refered to as a spectro-temporal receptive field (STRF).

As is the case with visual receptive fields, the STRF is proportional to the stimulus optimally exciting the cell [17]. The STRF model has been successful in describing the response properties of some auditory neurons [58]. Furthermore, since this model associates a particular stimulus element with each cell, the STRFs of a set of neurons can be considered as the basis of a neural code for natural sounds [24].

The STRF for a given auditory cell is typically found by minimising the squared error of this prediction

$$\mathcal{E} = \sqrt{\int (r - r')^2 dt} \tag{2.4}$$

where r is the actual neuronal firing rate determined from electrophysiological recordings. In this chapter, we describe a novel method of performing this cal-

culation.

The standard package for calculating auditory STRFs, STRFPak [65] makes use of the convolution theorem to transform the convolution in the STRF equation (2.3) into a product of Fourier transforms. The STRF is then found by least-squares minimisation of the k-space error between this product and the Fourier transformed firing rate. Although the use of the Fourier transform to deconvolve the formula for $\tilde{r}$ appears to be somewhat successful, it is not transparent: the deconvolution theorem for Fourier transforms applies to periodic functions or to the acausal convolution whereas the kernel here is compact and the convolution is causal. Furthermore, this method requires that neural spike time data be pre-processed using a smoothing window before the Fourier transform can be applied. Our intention here is to derive an alternative formulation for the STRF which avoids the use of the Fourier transform.

Here, we describe a simplified calculation in which the Fourier transform is avoided by re-formulating the problem as one of regularized matrix inversion in the stimulus space. This formulation is more mathematically robust than the standard STRFPak method, and does not require any smoothing of firing rates. In addition, these improvements allow us to calculate STRFs which more accurately predict the responses to novel stimuli

2.3 A Straightforward Calculation of the STRF

In the standard method for STRF calculation, the linear convolution equation, (2.3), is replaced by a Fourier transform equation:

$$\hat{r'}(k) = \sum_{f=1}^{n_f} \hat{s}_f(k)\hat{h}_f(k) \tag{2.5}$$

where $\hat{s}_f(k)$ and $\hat{h}_f(k)$ are the Fourier transforms of $s_f(t)$ and $h_f(t)$. The aim now is then to minimise the k-space error

$$\mathcal{E}_k^2 = \int [\hat{r}(k) - \hat{r'}(k)]^2 dk \tag{2.6}$$

with respect to $h_f(k)$. If $r'(k)$ was the inverse Fourier transform of $\hat{r'}(t)$ then the Plancherel formula would say that $\mathcal{E}_k = \mathcal{E}$. The minimization problem has a least squares solution and for each value of k, $h_f(k)$ is found by regularized inversion of the $n_f \times n_f$ stimulus autocorrelation matrix. The STRF $h_f(s)$ is then set equal to the causal part of the inverse Fourier transform.

Our straightforward STRF calculation takes advantage of the fact that, since both the firing rate and stimulus are, in reality, discretely sampled, the convolution can be written as a simple matrix multiplication. Thus, the STRF can be calculated by a single inversion. In other words, although the formula for $\mathcal{E}$ has the appearance of an integral, time is, in practice, discretized into bins of width δt and the convolution is rewritten in matrix form with time indices τ and σ corresponding to $t = \tau\delta t$ for time and $s = \sigma\delta t$, corresponding to the temporal support of the STRF. Temporal arguments are replaced by indices

$$R_\tau = r(\tau\delta t), \quad R'_\tau = r'(\tau\delta t), \quad S_{\tau f} = s_f(\tau\delta t), \quad H_{f\sigma} = h_f(\sigma\delta t) \tag{2.7}$$

and the linear model, (2.3), is now

$$\tilde{R}_\tau = \sum_{f=1}^{n_f} \sum_{\sigma=1}^{n_s} S_{\tau-\sigma,f} H_{f\sigma} \tag{2.8}$$

where $n_s\delta t = T$, the temporal width of the STRF. The error is now given by

$$\mathcal{E}^2 = \sum_{\tau=1}^{n_t} (R_t - \tilde{R}_t)^2 = \sum_{\tau=1}^{n_t} \left(R_\tau - \sum_{\sigma,f} S_{\tau-\sigma,f} H_{f\sigma} \right)^2 \tag{2.9}$$

where $L = n_t \delta t$ is the temporal length of the stimulus.

Differentiating with respect to $H_{f\sigma}$ gives the least squares fit solution

$$\sum_{\tau,\rho,g} S_{\tau-\sigma,f} S_{\tau-\rho,g} H_{g\rho} = \sum_{\tau} R_\tau S_{\tau-\sigma,f} \tag{2.10}$$

Thus, the problem of calculating the STRF reduces to a simple matrix inversion problem. To make this clearer, the STRF indices σ and f are vectorized so that $I = (f-1)n_f + \sigma$ and $J = (g-1)n_f + \rho$ and hence $S_{\tau I} = S_{\tau-\sigma,f}$ and, for example, $H_J = H_{g\rho}$. Now, the equation becomes

$$\sum_{J=1}^{n_I} (S^T S)_{IJ} H_J = C_I \tag{2.11}$$

where $C_I = \sum_\tau R_\tau S_{\tau-\sigma,f}$, the dimension $n_I = n_f n_s$ and, for clarity, the shorthand

$$(S^T S)_{IJ} = \sum_{\tau} S_{\tau I} S_{\tau J} \tag{2.12}$$

has been used for the square matrix. Now, H_J, and therefore the STRF, is recovered by inverting $S^T S$.

The matrix $S^T S$ is known as the stimulus autocorrelation matrix, and represents the temporal, spectral, and temporal-spectral correlations in the stimulus ensemble. In the past, experiments of this kind were often performed using Gaussian white noise stimuli. In this case, the matrix $S^T S$ approximates to a scalar multiple of the identity, I_{n_I}, and so the STRF reduces to

$$H_{f\sigma} = \sum_{\tau} R_\tau S_{\tau-\sigma,f} \tag{2.13}$$

which, up to a multiplicative constant, is simply the spike triggered average.

However, it is known that receptive fields, including STRFs, are highly dependent on the choice of stimulus ensemble. Furthermore, since neural spiking is an

inherently noisy process, stimuli must be used which elicit a strong response in the cells being studied, in order to minimise noise in the estimation of neuronal firing rates. In the case of songbirds, a corpus of con-specific songs is considered a good proxy for natural sounds. However, the difficulty is that the correlation structure of such natural sounds makes the matrix inversion ill-conditioned. This must be dealt with through regularization and dimensional reduction.

2.4 Stimulus Correlation and Regularization

Natural sounds such as birdsong are highly structured, and so can contain strong correlations in either time and frequency. The effect of these correlations is that the signal is predominately carried in a relatively small number of high variance dimensions, and as a result, the stimulus autocorrelation matrix has a number of very small eigenvalues, corresponding to the low variance dimensions in the time-frequency space. Given the spectral decomposition:

$$(S^T S)_{IJ} = \sum_{\alpha} \lambda_\alpha E_{\alpha I} E_{\alpha J} \tag{2.14}$$

the inverse is given by

$$(S^T S)^{-1}_{IJ} = \sum_{\alpha} \frac{1}{\lambda_\alpha} E_{\alpha I} E_{\alpha J}. \tag{2.15}$$

and so the small eigenvalues, λ_α become very large on inversion, and dominate the inverse.

In mathematical terms, the autocorrelation matrix is ill-conditioned, and so its inverse is susceptible to noise in poorly sampled dimensions. This presents a problem: in order for the autocorrelation matrix to be well-conditioned - that is, to have a stable inverse - the stimulus ensemble is required to well sample the entire time-frequency space. However, in order to evoke a strong response from auditory neurons naturalistic stimuli must be used and, by their nature, such stimuli contain

a high degree of correlation, and are therefore ill conditioned.

This problem can be circumvented by a number of methods. A common approach to dealing with such ill-conditioned matrices is Tikhonov regularization, also known as ridge regression. This technique progressively diminishes the contributions of smaller eigenvalues to the inverse. Alternatively, the problem may be solved by dimensional reduction. Simply put, the stimuli are projected down into the subspace of significant dimensions, that is, dimensions which are well sampled by the stimulus ensemble; in other words, principle component analysis is performed on the stimulus.

The relative merits of these techniques are discussed below. As a further complication, we must also consider whether the correlations in the stimulus are in time, frequency, or both.

2.4.1 Tikhonov regularization in STRF-space

One possible approach to regularization is to impose a prior condition on the solution which prevents overfitting, and eliminates unstable solutions. Tikhonov regularization [68] allows us to find a reasonable solution to an ill posed inverse problem by incorporating additional qualitative information about the problem. For example, we might intuitively expect a good, noiseless solution to be reasonably smooth, or to have a reasonable upper bound on its norm. We can then impose a smoothness or boundedness condition on the solution, H, which has the effect of suppressing noisier solutions.

The least squares STRF estimate is found by solving:

$$H_I = \arg\min \sum_{\tau=1}^{n_t} \left(R_\tau - \sum_I S_{\tau,I} H_I \right)^2 \tag{2.16}$$

The regularized solution is found by adding an extra term to this expression

which penalises unsuitable solutions.

$$H_I = \arg\min \left(\sum_{\tau=1}^{n_t} \left(R_\tau - \sum_I S_{\tau,I} H_I \right)^2 + \eta \left(D_I H_I \right)^2 \right) \tag{2.18}$$

where $D \in \mathbb{R}^{n_I}$ is a linear operator characterising some property of the solution which is to be minimised. Common choices are $D = I_{n_I}$ where we require the norm of H to be small, or a differential operator in the case where smoothness is requred. This gives the stable regularized solution:

$$\sum_{J=1}^{n_I} (S^T S + \eta D^T D)_{IJ} H_J = C_I \tag{2.19}$$

for some regularization parameter $\eta > 0$, and the STRF is now recovered by finding the inverse of $(S^T S + \eta D^T D)$

In the standard case where $D = I_{n_I}$, this is equivalent to simply adding a small multiple of the identity matrix to $S^T S$. This gives the inverse:

$$(S^T S + \eta I_{n_I})^{-1}_{IJ} = \sum_\alpha \frac{1}{\lambda_\alpha + \eta} E_{\alpha I} E_{\alpha J}, \tag{2.20}$$

and so the contribution to the inverse of small eigenvalues is bounded, converging to $1/\eta$ as λ approaches zero.

2.4.2 Regularization by dimensional reduction in STRF-space

The birdsong data used in our STRF calculations consists mainly of a number of highly structured motifs, which are repeated with little variation. As a result, it is likely that that the information in the stimulus is contained in a small number of dimensions, and that the low-variance components of the stimulus contain little or no relevant information. In such cases, it may be preferable to remove the contributions due to these dimensions entirely. This can be achieved through Principal Component Analysis (PCA) and dimensional reduction.

In the case of STRF calculations, if the stimulus is highly correlated in both time and frequency, the dimensional reduction should be performed on the entire $n_I = n_f n_s$-dimensional time-frequency space; the space of STRF vectors. This is achieved by replacing the inverse of $S^T S$ with the Moore-Penrose pseudoinverse, which removes contributions due to low-variance dimensions:

$$(S^T S)^+_{IJ}(\epsilon) = \sum_{\alpha:\lambda_\alpha > \epsilon} \frac{1}{\lambda_\alpha} E_{\alpha I} E_{\alpha J}, \tag{2.21}$$

where ϵ is a tolerance value, separating significant eigenvalues from noise. By virtue of the orthogonality of eigenvectors,

$$\sum_l E_{\alpha l} E_{\beta l} = \delta_{\alpha\beta} \tag{2.22}$$

this yields a STRF, $H_J(\epsilon)$, which is restricted to the subspace of significant eigenvectors. The tolerance value, ϵ is determined by cross validation.

This gives a dimensional reduction in the time-frequency space of STRFs. As a result both spectral and temporal features have been removed from the STRF estimate. However, neural firing rates contain fine temporal structure, and so by projecting the problem in STRF space, significant temporal information may be lost from the STRF estimate. In other words, though the matrix $S^T S$ exists in time-frequency space, the under-sampling which causes it to be ill-conditioned may be in frequency space only, and so we must also consider a solution which removes noise due to spectral correlations in the stimulus, but retains temporal information.

In practice, the time-frequency approach leads to STRFs which are poor at predicting firing rates for novel stimuli. A better approach, in fact, is to consider only the redundancy in frequency.

2.4.3 Regularization in frequency space

At each time step, τ, $S_{f\tau}$ gives an n_f dimensional vector in frequency space; if there is redundancy, this stimulus vector can be projected down to a lower dimensional space without significant loss of signal. To determine this subspace, the $n_f \times n_f$ matrix of spectral correlations:

$$(C^S)_{fg} = \sum_{\tau} S_f(\tau) S_g(\tau) \tag{2.23}$$

is calculated. If the stimulus contains a high degree of spectral correlation, then this matrix will have a small number of large eigenvalues corresponding to the basis vectors of the sub-space of significant dimensions. The stimulus is then projected down onto this subspace to obtain the projected stimulus

$$S_m(\tau) = \sum_{\gamma=1}^{n_r} \left(\sum_{g=1}^{n_f} S_g(\tau) V_{g\gamma} \right) V_{m\gamma} \tag{2.24}$$

where V_r is an eigenvector of C^s, and m runs from 1 to n_r, the dimension of the restricted subspace formed by the first n_r eigenvectors of C^s. The STRF is then calculated using the projected stimulus. It will be seen that this approach is successful in regularizing the STRF calculation.

This projection in the frequency space is similar to the regularization used in conjunction with the Fourier transform in the STRFPak calculation. In that method, an $n_f \times n_f$ correlation matrix is inverted for each value of k in the k-space of amplitude modulation frequencies. These matrices are then regularized by using the Moore-Penrose inverse. This approach has the advantage that a different subspace can be selected for each value of k, allowing much greater freedom in the choice of information to be retained in the STRF estimate.

Conversely, the approach suggested here has the advantage of being computationally less demanding than STRFPak, since the stimulus projection must be

performed only once. Furthermore, it has the conceptual advantage that the nature of the regularization is intuitively clearer.

2.4.4 Cross-Validation

In all of the regularization methods described above, we must optimise the calculation with respect to a regularization parameter. The tolerance value, ϵ, or equivalently, the number of retained dimensions, n_r is chosen so as to give the most accurate prediction of the response to novel stimuli. In both cases this is acheived by means of cross-validation.

To achieve a reliable STRF estimate, we must ensure that the STRF is not over-fitted to the noise in the training data. To avoid this, we must validate our STRF by applying it to data not used in training. Prior to calculation a subset of the data, known as the validation set, is put aside and the STRF calculated using the remainder of the data, known as the training set. The resulting STRF is then used to predict the response to the validation data. The value of the regularization parameter is chosen as that which miniminses the prediction error on the validation set.

Here, we have used a rigorous validation regime known as 10-fold cross-validation. In this case, the training data is divided into 10 subsets. The cross calidation is then repeated 10 times, with a different subset chosen as the validation set on each repetition. Prediction accuracy for each value of the regulsrisation parameter is then averaged across all repetitions of the validation, and the regularisation value chosen as that which gives the highest mean correlation across all 10 validation sets. This method of cross-validation has the advantage that the result is validated against the entire data set, while still ensuring that the result is not overfitted to noise in the training set. Here our data consisted of 20 zebra finch song recordings, and so we performed a validation in which one of 10 song pairs was excluded from the training set on each repetition.

2.5 Results

We apply our novel STRF calculation methods to a dataset consisting of zebra finch song recordings and corresponding spike trains recorded from 10 cells in the Field L region of the zebra finch auditory forebrain. Song spectrograms were calculated for an ensemble of 20 conspecific songs as described in Section 2.1 above, with $n_f = 32$ frequency bands of with 250Hz, and centre frequencies ranging from 250-8000Hz. Neural data was recorded in spike arrival time format, and binned to form a Peri Stimulus Time Histogram (PSTH). The STRF is calculated directly from an unsmoothed PSTH with 1ms resolution. Results shown are for STRFs with a temporal width of 100ms. However, STRF width may be reduced to as little as 50ms with little deterioration in predictive performace.

Figure below shows the mean prediction accuracy for our calculation method using the three regularization techniques described in the previous section, as well as that obtained from the STRFPak calculation. As can be seen, the frequency-space regularization produces the most accurate STRF estimate, with performance comparable to that of STRFPak. Detailed comparison between this method and the STRFPak algorithm is shown below.

Table 2.1 displays the correlation coefficient for each of our 10 cells for optimal values of the respective regularization parameters. As can be seen, our method produces predictions with similar, and on average, slightly improved accuracy over STRFPak. Correlation values for other regularization methods are also shown for comparison. The final column displays the test statistic for matched sample t-test between Frequency Space regularization and STRFPak predictions. Statistically significant results are shown in bold. In 3 out of 10 cells our frequency based method produces significant improvements over STRFPak, while STRFPak is significantly better in 2 out of 10 cases. For the remaining 5 cells differences in correlation were not significant at the 5% level.

In addition, both of these methods were significantly more accurate than either

the Tikhonov or STRF-space methods in the majority of cases. In no instance did either of Tikhonov or STRF-space methods produce significantly better predictions.

Cell no.	Tikhonov	STRF-space proj.	Freq.-space proj.	Strfpak	t-value
1	0.167(0.028)	0.191(0.031)	0.235(0.036)	0.261(0.04)	**2.06**
2	0.147(0.023)	0.161(0.024)	0.185(0.041)	0.194(0.037)	0.51
3	0.202(0.023)	0.196(0.018)	0.243(0.021)	0.215(0.029)	**2**
4	0.138(0.019)	0.158(0.02)	0.182(0.025)	0.183(0.024)	0.09
5	0.273(0.024)	0.303(0.029)	0.429(0.032)	0.319(0.033)	**7.58**
6	0.027(0.022)	0.042(0.035)	0.052(0.042)	0.034(0.036)	1.03
7	0.192(0.017)	0.202(0.031)	0.240(0.035)	0.214(0.027)	**1.86**
8	0.203(0.034)	0.215(0.038)	0.243(0.066)	0.239(0.051)	0.15
9	0.131(0.026)	0.163(0.032)	0.155(0.03)	0.183(0.034)	1.62
10	0.134(0.029)	0.159(0.033)	0.162(0.041)	0.209(0.047)	**2.37**
Mean (STD):	0.161(0.062)	0.179(0.061)	0.213 (0.092)	0.205 (0.072)	

Table 2.1: The mean correlation coefficient between predicted and actual firing rates for each cell is shown for 4 estimation methods Standard deviations are given in parentheses. Also shown are matched sample t-test values for our frequency space method versus the standard Strfpak calculation (5% significance level: 1.86). Bold type indicates valuea above the significance level.

The STRF calculation is dependent on the number n_r of principal components retained in our representation of the stimulus. This number is determined for each cell by means of cross validation; n_r being chosen as that value which returns the most accurate prediction of responses to the validation stimuli. In the STRFPak calculation, this number is expressed in terms of a tolerance factor, μ. Only those dimensions with eigenvalues larger than $\mu\lambda_1$ are retained in the solution, where λ_1 is the first, or largest eigenvalue.

Figure 2.2 shows the prediction accuracy for both our frequency space method and STRFPak as functions of n_r and μ, respectively. Accuracy is measured using the correlation coefficent of unsmoothed predicted and actual firing rate histograms.

To achieve the best possible prediction, n_r must be determined for each cell. Optimal values were in the range $-5 \leq n_r \leq 8$. The first 8 principle components are shown in Figure 2.3.

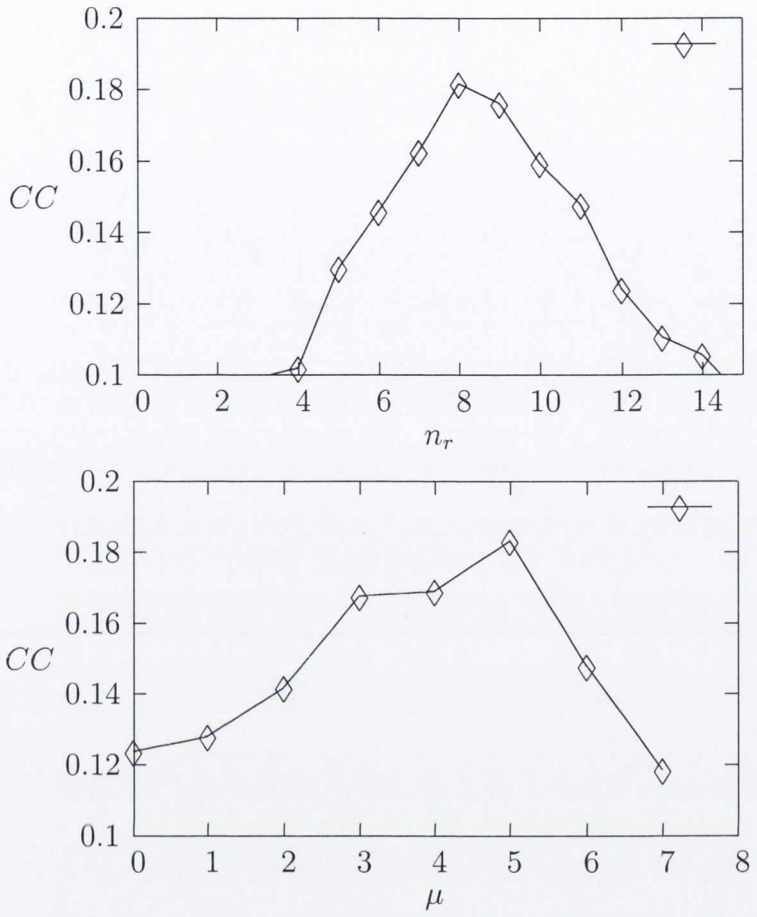

Figure 2.2: Dependence on regularization parameter: Correlation coefficient of predicted and actual responses for cell 4 are shown for our method (**A**) and Strfpak (**B**) as a function of the respective regularization parameters, n_r and μ

2.6 Discussion

The results above confirm that it is possible to calculate reasonably accurate, low-noise STRF estimates without the use of the many approximations implicit in deconvolution based methods.

In addition, by avoiding the transformation of the problem into k-space, we demonstrate how the STRF calculation can be conceptually simplified, and the regularization process more easily understood as a dimensional reduction in the space of stimulus spectrograms. However, although it gives similar predictive power to the STRFPak method, the STRF as calculated here still produces relatively poor predictions of neural firing rates. This is due in large part, to the inherent non-

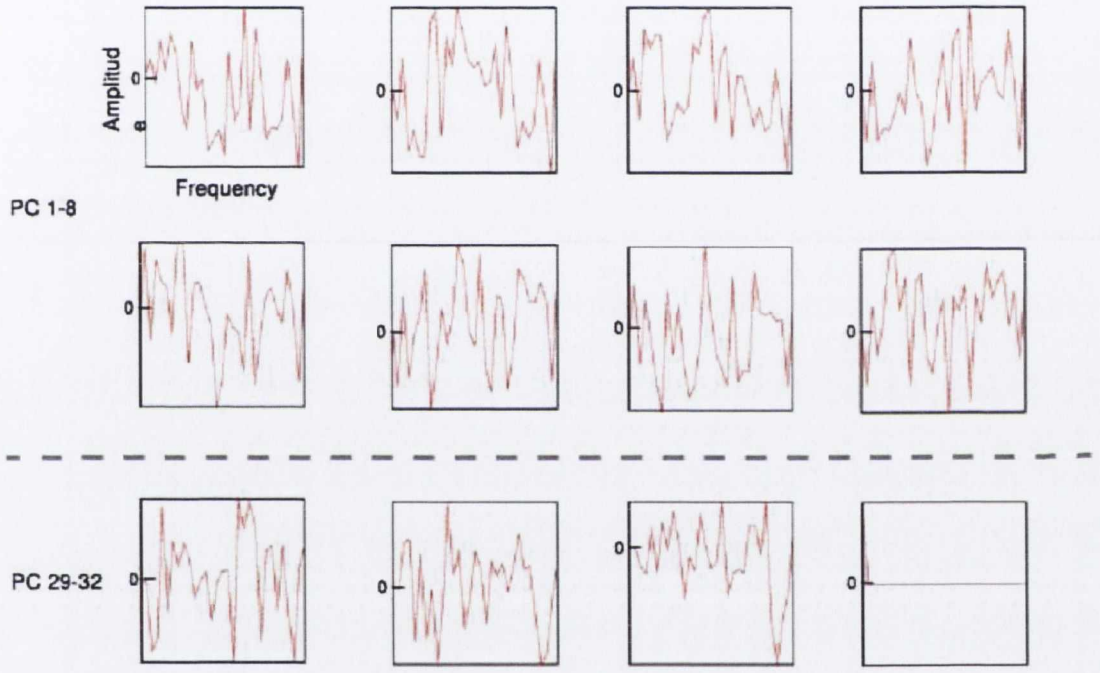

Figure 2.3: Plots of a selection of stimulus principle components in frequency space. The stimulus spectrogram at each timestep can be respresented as a sum over the principal components. **Top:** Principle components 1-8, which were sufficient to produce optimal STRF estimates for all cells. **Bottom:** Principle components 29-31. These were the noisiest, or least significant dimensions in the frequency space. Component amplitudes are independently rescaled for plotting.

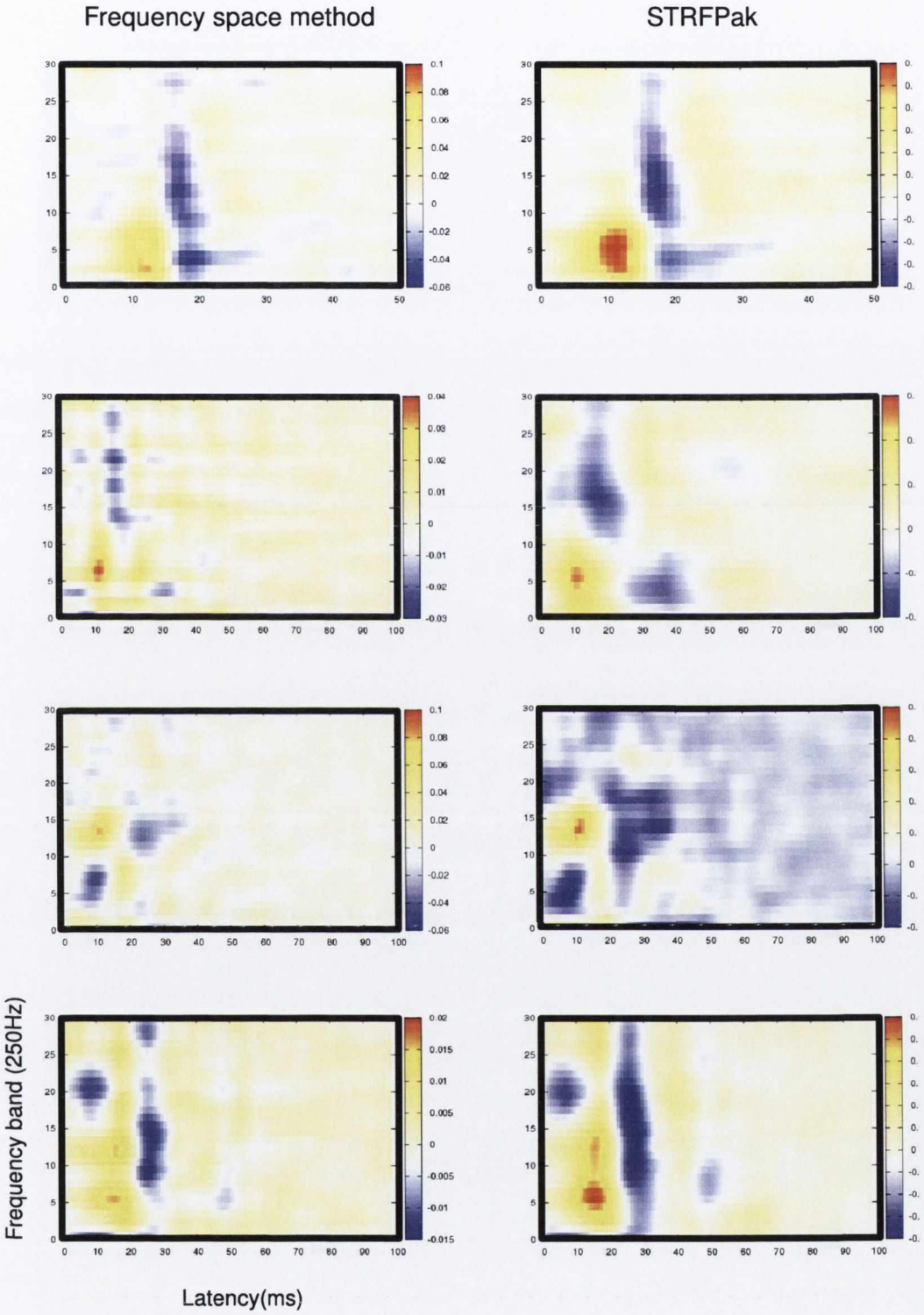

Figure 2.4: STRFs caclulated using our frequency space method (left column) compared to those obtained from STRFPak (right column) for cells 1-4. Colour scales are normalised for comparison. Predictive performance as measured by the correlation coefficient (CC) is invariant with respect to STRF amplitude.

linearity of neural responses, which can at best be roughly approximated by such linear estimates. However, there may still be significant scope for improvements in the linear STRF model.

As discussed previously, it is essential when calculating STRF estimates to use a training set which well samples the relevant stimulus space. However, in order to effectively do this for such a complicates stimulus set as birdsong, it may be necessary to use a much larger dataset than that used here in order to obtain a good estimate. It is instructive to consider the performance of the STRF when predicting responses to the validation data compared to the peformance on data used in training.

Cell no.	Mean training set correlation	Mean validation set correlation
1	0.371	0.235
2	0.325	0.185
3	0.367	0.243
4	0.332	0.182
5	0.494	0.429
6	0.214	0.052
7	0.322	0.240
8	0.354	0.243
9	0.323	0.155
10	0.303	0.162
Mean (STD):	0.341 (0.066)	0.213 (0.092)

Table 2.2: The correlation coefficient between predicted and actual firing rates, averaged over all songs, is shown for both our direct method, and the standard Strfpak calculation.

Table 2.2 shows the mean prediction correlations for both the validation and training data using our frequency space method. The prediction accuracy is seen to be significantly worse for the validation set than for the songs used in training (Matched sample t-test: t=3.6; 5% significance level = 1.73). Of course, it is always possible to produce a highly accurate prediction for the training data if the model is 'over-fitted' to the noise in the training set. However, the regularization procedure used here is specifically designed to prevent such overfitting, and so if the training data well-samples the space of significant stimuli, we might expect the performance on both training and validataion sets to be comparable. This is not the case here,

indicating that there are significant dimensions within the validation stimulus which are not well sampled by the training stimuli. This suggests that the dataset of 20 zebra finch songs used here is not sufficiently large to provide a good STRF estimate, and that significant improvements in performance may be obtained by training with a larger sample.

It may also be possible to obtain improvements in performace by further refinement of the regularization process. We have described three possible methods for regularizing the STRF calculation through ridge regression and dimensional reduction; two in which this reduction is performed on the entire time-frequency space of STRFs, and another in which the projection is performed in frequency only.

As we have seen, the frequency only method produces the more accurate STRF estimate, and so we can conclude that the under-sampling in our stimulus is primarily in the spectral dimensions. However, it is possible that there remain poorly sampled temporal dimensions in the projected stimulus, and so it may be possible to improve our STRF calculation by performing a second, less drastic regularization in the temporal domain. Unfortunately this method has the drawbacks of requiring the optimization of a second regularization parameter, and of significantly increasing the computation time of the calculation.

Another potential improvement to the STRF estimate may be achieved through the introduction of a filtering kernel for the neuronal firing rate data. As discussed above, many STRF estimation methods, including STRFPak, require the smoothing of input firing rates prior to estimation. This is typically achieved using a Hanning window [12], or similar filter function. The inclusion of this pre-processing step is often poorly justified, and primarily motivated by the need to smooth the firing rate function prior to the calculation of its fourier transform. Nonetheless, the use of such a smoothing window has been shown to slightly improve the prediction accuracy of STRFs. It is interesting to consider why this may be so.

Neural action potentials have an extremely narrow temporal profile, and for most

coding purposes, may be treated as instantaneous events. As a result, they are often approximated mathematically as δ-functions. However, such spikes are perceived by other neurons through their effects on those cells' post synaptic potentials, which often persist over a much longer timespan. Typically, the effect of a pre-synaptic spike on a post synaptic dendrite is to produce a rapid increase in the dendrite potential, which then decays exponentially on a timescale significantly larger than that of the pre-synaptic spike. Hence, a neural spiketrain is transmitted across a synapse, not as a series of δ-function spikes, but rather as the sum of a series of decaying exponential functions.

This idea is used extensively in the design of kernel based spike metrics, where a spike train of delta function spikes

$$f(t) = \sum_s \delta(t - t_s) \tag{2.25}$$

is convolved with a causal filter, θ to give a rate function of the type

$$\tilde{f}(t) = \sum_{i=1}^{n} \theta(t - t_i) \tag{2.26}$$

where θ is of the form

$$\theta(t) = \begin{cases} 0 & t < 0 \\ e^{-t/\tau} & t \geq 0 \end{cases} . \tag{2.27}$$

This filtering is functionally similar to that carries out by smoothing firing rates with a Hanning window in STRF calculations, and has the same effect of broadening the temporal profile of spikes. Unlike the Hanning window, however the shape of the filter function used here is directly inspired by neurophysiological measurements. Hence, we may expect that the use of such a filter function in STRF estimation might improve the predictive power of the STRF model.

Chapter 3

Sparse Coding of Natural Stimuli.

It has long been established that the firing rate behaviour of many cells in the primary visual and auditory areas can be predicted by a linear filter model. Any discussion of this prediction must be undertaken with several caveats: the accuracy of the prediction is modest [40, 21, 67, 58] and there are numerous non-linear effects which make the calculation of the kernel dependent on the corpus of stimuli [41, 65, 66, 18]. Furthermore, the model predicts only the spike rate and provides no information about spike timing. Nonetheless, these linear models do associate a particular kernel with a given cell and it is obviously interesting to ask what determines the selection of these kernels.

This question is perhaps unusually well-specified in the case of song birds. Since song birds are adept at distinguishing between different con-specific songs, these songs can be considered an important class of natural sounds. Ideally, sensory processing is studied using stimuli whose statistics reflect those of the natural environment [18]. A guiding principle in neural coding is that sensory systems should efficiently encode such stimuli, and in fact, there is already evidence from the study of the visual system, that the linear kernels of visual neurons are related to a sparse code for natural images [71, 47, 72, 5]. Furthermore, modelling of auditory systems [38] has shown that the tuning properties of cochlear hair cells are well predicted by

a sparse code for natural sound waveforms.

Our aim in this chapter is to extend these ideas to the avian auditory system. The methods used are similar to those employed in these previous studies, however, additional difficulties arise because birdsong does not well-sample the entire frequency-time domain.

The male zebra finch sings; along with a variety of simple calls, such as warning cries, the male bird has a single, identifying song, which develops under the tutelage of an adult male. The female finch does not sing, however, both the male and female birds are able to distinguish songs. Songs usually begin with a series of introductory notes, followed by two or three repetitions of the motif: a series of complex frequency stacks known as syllables, separated by pauses. Syllables are typically about 50ms long, with songs lasting about two seconds. Although perhaps discordant to the human ear, zebra finch songs have a very rich and complex structure. Importantly, the zebra finch auditory system is believed to be highly tuned to detect and recognise this song structure [41, 18, 67].

As discussed in the previous chapter, the response characteristics of auditory neurons are described by the spectro-temporal receptive field (STRF), a linear kernel relating the spectrogram of the stimulus to the firing rate response of the neuron. While linear in the spectrogram, the STRF model is non-linear in the stimulus due to a non-linear transformation in the calculation of the spectrogram. Such a linear mapping from spectrogram to response is rather naive and, not surprisingly, gives an incomplete description of neuronal behaviour [40, 21, 67, 58]. Nonetheless, the model does provide a good approximation for some cells, and a description of how information is processed and encoded in the primary auditory areas should account for this linear behaviour.

Numerous previous studies, [65, 40, 58], have produced STRFs for auditory neurons from electrophysiological recordings, using the traditional STRFPak algorithm. In particular, the STRFs of Field L neurons in the zebra finch auditory forebrain

have been calculated and parameterised by a number of quantitative measures, such as the location and width of the time and frequency peaks. These STRFs are also characterised by a number of distinctive spectral and temporal features such as narrowband selection and on-off switching.

We investigate whether these properties of simple neurons can arise naturally from a sparse coding strategy for natural sounds. Specifically, we consider optimal strategies for the encoding of an ensemble of 20 zebra finch songs and generate a set of optimal kernels which sparsely encode this ensemble using an Olshausen-Field type algorithm [47].

Such learning algorithms have been succesfully used to calculate sparse bases for natural images [47, 71]. Here,we adapt it for use with highly correlated, ill-conditioned data, and apply it to the birdsong spectrograms. Using our novel formulation of the STRF from Chapter 2, we can then calculate optimal linear filters for zebra finch auditory neurons which are explicitly equivalent to neuronal STRFs.

3.1 Recap: Spectrotemporal Receptive Fields

As in Section 2.1, we consider a spectrographic representation of our songs, where the spectrogram represents the log amplitude of the stimulus in frequency and time, obtained by narrowband filtering Fourier transform of the song waveform. Spectrograms are represented by a combination of $n_f = 32$ narrowband signals, $\{s_f(t)\}$, with centre frequencies between 250 and 8000Hz.

To recap from Chapter 2, according to the STRF model, an approximate firing rate is calculated by convolving the spectrogram with a kernel $h_f(s)$:

$$\tilde{r}(t) = \sum_{f=1}^{n_f} \int s_f(t-s) h_f(s) ds \tag{3.1}$$

This calculation can then be discretised and the STRF indices combined, to give

the STRF equation in vector form:

$$\tilde{R}_\tau = \sum_{I=1}^{n_f n_s} S_{\tau I} H_I. \tag{3.2}$$

where n_s is the temporal width of the STRF, and the vector index I labels both spectral and temporal dimensions.

The STRF is defined as the best approximate solution of Equation (3.2) which minimises the squared error between the predicted firing rate, $\tilde{R}_\tau$ and the actual neuronal firing rate, R_τ. This is given by the least squares solution

$$\sum_{J=1}^{n_f n_s} (S^T S)_{IJ} H_J = C_I \tag{3.3}$$

where $C_I = \sum_\tau R_\tau S_{\tau-\sigma,f}$ and, for clarity, the shorthand

$$(S^T S)_{IJ} = \sum_\tau S_{\tau I} S_{\tau J} \tag{3.4}$$

has been used for the square matrix. Now, H_J, and therefore the STRF, is recovered by inverting $S^T S$.

In practice, however, this precise solution does not give the best STRF estimate, and the STRF calculated in this way will, in fact, give a poor prediction of the response to novel stimuli not used in the calculation. This is a consequence of overfitting to the training data.

As discussed above, to realistically characterise neural responses, we must use stimuli which provoke a strong response in the neurons of interest [18, 67, 66]. In fact, the existence of an easily specified ensemble of natural stimuli is a key advantage of using song birds in studies of the auditory system. However, there is a disadvantage: natural sounds such as birdsong have a high degree of temporal and spectral auto-correlation, and so the majority of the information in the stimulus is contained in a relatively small number of significant dimensions. As a result, there exist dimensions

within the stimulus space along which the variance is extremely low and in which noise becomes significant. Since the least squares solution gives equal weighting to all dimensions in the stimulus, this results in the STRF being fitted to the noise in these dimensions. In other words, the stimulus autocorrelation matrix $S^T S$ is generally ill-conditioned.

As described in Section 2.4, this problem can be overcome by regularized inversion in the space of STRFs: If

$$(S^T S)_{IJ} = \sum_{\alpha} \lambda_\alpha E_{\alpha I} E_{\alpha J} \tag{3.5}$$

is the spectral decomposition of $S^T S$ over its eigenvalues, λ_α, and eigenvectors, $\mathbf{E}_\alpha$, then the inverse is given by

$$(S^T S)_{IJ} = \sum_{\alpha} \frac{1}{\lambda_\alpha} E_{\alpha I} E_{\alpha J} \tag{3.6}$$

Regularization is achieved by taking the Moore-Penrose pseudoinverse [52]

$$(S^T S)^+_{IJ}(\epsilon) = \sum_{\alpha:\lambda_\alpha > \epsilon} \frac{1}{\lambda_\alpha} E_{\alpha I} E_{\alpha J}. \tag{3.7}$$

where the contributions due to poorly sampled dimensions, which correspond to small eigenvalues, are removed from the inverse.

The tolerance value, ϵ, is chosen by cross validation so as to give the most accurate prediction of the response to novel stimuli.

3.2 Sparse Coding

According to the sparse coding hypothesis for sensory systems, only a small subset of the neurons in a sensory pathway need be strongly active while accurately encoding a given stimulus [50, 47, 22, 4]. From an information theoretic point of view, an

ideal sparse coding regime is one in which the neuronal firing rates are statistically independent [4, 38, 47] and individual cells favour either low or high activity. Here, neurons are identified with kernels or STRFs, so each neuron corresponds to one direction in a stimulus space. In this section, we calculate an optimal set of linear kernels which sparsely encode zebra finch song.

Using the same vectorized notation as in the previous section, $S_{\tau I}$ can be thought of as a patch of the stimulus spectrogram with the same temporal width as the STRFs, ending at the time $t = \tau\delta_t$. The spectrogram patch is decomposed at fixed time τ over a basis B_{nI} where n is a component index. Hence, let

$$\tilde{S}_{\tau I} = \sum_n A_{\tau n} B_{nI} \tag{3.8}$$

where A is the matrix of components. Assuming that the basis B is invertible, it is possible to choose the component matrix A so that $\tilde{S}_{\tau I} = S_{\tau I}$ by setting

$$A_{\tau n} = \sum_I S_{\tau I} (B^{-1})_{In}. \tag{3.9}$$

Where each row of B is a basis vector associated with the neuron n. Notably, this equation shares the form of the vectorised STRF equation (3.2) In this way, the n^{th} column of the inverse basis B^{-1} is equivalent to the STRF, H_I, of the neuron n and $A_{\tau n}$ is equivalent to the firing rate of that neuron at time $t = \tau\delta_t$. However, in an efficient coding regime, we must also require that the firing rates be sparse. This is achieved by placing a constraint on A which enforces sparseness in the distributions of firing rates. We can then determine an accurate sparse coding of our stimuli by allowing a trade off between the sparseness of the representation and the accuracy of the stimulus reconstruction $\tilde{S}$. Furthermore, as with the STRFs, it will be neccessary to regularize the calculation.

Following the method of Olshausen and Field and Vincent et al. [47, 71] we seek

to minimise an energy function, $E(A, B; \mu)$ for the sample at each time τ:

$$E = \sum_I \left(S_{\tau I} - \tilde{S}_{\tau I} \right)^2 - \mu \sum_n C(A_{\tau n}) \tag{3.10}$$

where the first term represents the reconstruction error in the representation, and where $C(\cdot)$ is some sub-linear cost function which penalises redundancy in the coding for a given sample. A typical choice [47] which is used here, is

$$C(A_{\tau n}) = -\left[\log(1 + A_{\tau n} A_{\tau n})\right], \tag{3.11}$$

which favours representations having fewer non-zero coefficients, since $\sum_i \log_i(1 + x_i^2) \geq \log(1 + \sum_i x_i^2)$ for all x_i. μ is a positive constant which determines the relative importance of sparseness and reconstruction accuracy.

To find the minimum a two-step iterative method is used: $E(A, B; \mu)$ is first minimised with respect to the components $A_{\tau n}$ by conjugate gradient descent, averaged over many samples. The basis functions are then updated by

$$\Delta B_{nI} = \eta \left\langle A_{\tau n} \left(S_{\tau I} - \tilde{S}_{\tau I} \right) \right\rangle_\tau \tag{3.12}$$

where $\eta > 0$ is the learning rate. Beginning with a random basis set, this algorithm converges after several thousand iterations to a matrix B of optimal basis functions which allow an accurate sparse encoding of the stimulus. Figure 3.1 shows the increase in the sparseness of the system after learning. The optimal kernels are now given by B^{-1}. Hence, B must be required to be invertible and well conditioned.

Difficulties arise in this calculation due to the highly correlated nature of the data used. Such difficulties are dealt with in many studies [47, 8] by a process of whitening, or sphering the data. However, in this case, in order to allow for the inversion of our basis, and the direct comparison of our sparse kernels with auditory STRFs, we proceed by means of dimensionality reduction, as used in the calculation

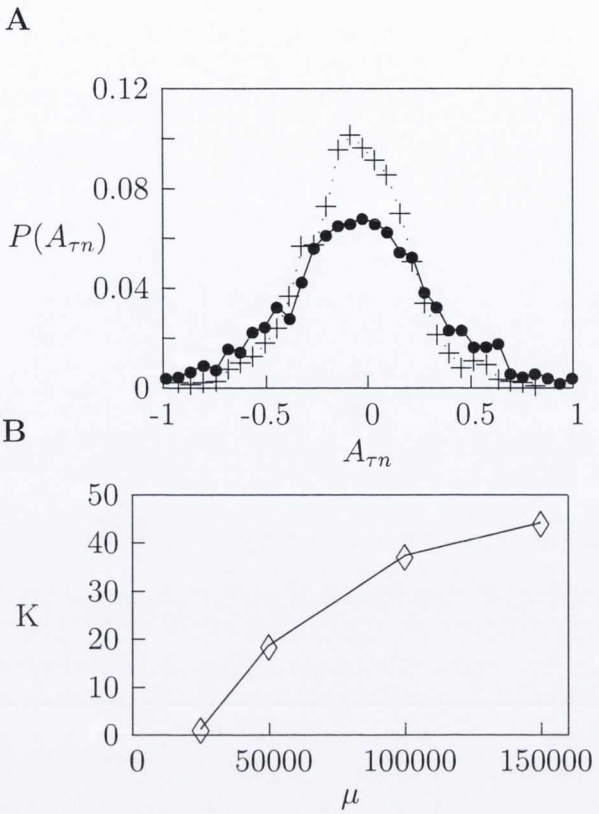

Figure 3.1: **A**: Distribution of weights for learned basis functions (dotted line) compared to those for a random basis (solid line) averaged over all filters. The large peak and heavy tail of the distribution for learned basis functions is characteristic of a sparse response. **B**: The kurtosis, K, of the distribution of weights for learned basis functions increases as a function of the sparseness parameter.

of STRFs. As we have described in Section 3.1 above, the regularized STRF is calculated by removing the contributions due to low variance dimensions in the stimulus, and projecting the songs onto a sub-space of high-variance dimensions. Hence, if, as above

$$(S^T S)_{IJ} = \sum_\alpha \lambda_\alpha E_{\alpha I} E_{\alpha J} \tag{3.13}$$

and anything with an I index can be decomposed over the eigenbasis

$$\begin{aligned} S_{\tau I} &= \sum_\alpha s_{\tau\alpha} E_{\alpha I} \\ \tilde{S}_{\tau I} &= \sum_\alpha \tilde{s}_{\tau\alpha} E_{\alpha I} \\ B_{nI} &= \sum_\alpha b_{n\alpha} E_{\alpha I}. \end{aligned} \tag{3.14}$$

Substituting these into the energy function, and using $\sum_I E_{\alpha I} E_{\beta I} = \delta_{\alpha\beta}$,

$$E = \sum_{\alpha} (s_{\tau\alpha} - \tilde{s}_{\tau\alpha})^2 - \mu \sum_{n} C(A_{\tau n}). \tag{3.15}$$

To project the problem onto the significant stimulus dimensions, we need only restrict the range of α. We write $\mathcal{A}(\epsilon)$ for the set of α such that $\lambda_\alpha > \epsilon$, where ϵ is a cut-off value seperating the high-variance dimensions of the songs - the ones that will be preserved - from the noise. The energy function now becomes

$$E(\epsilon) = \sum_{\alpha \in \mathcal{A}(\epsilon)} (s_{\tau\alpha} - \tilde{s}_{\tau\alpha})^2 - \mu \sum_{n} C(A_{\tau n}) \tag{3.16}$$

and we minimize over $b_{n\alpha}$ rather than B_{nI}. B_{nI} can be reconstructed as

$$B_{nI} = \sum_{\alpha \in \mathcal{A}(\epsilon)} b_{n\alpha} E_{\alpha I}. \tag{3.17}$$

To obtain a set of optimal kernels, B must now be inverted. If a complete representation is chosen, where the number of basis elements N is the same as the number of dimensions in the stimulus; $N = |\mathcal{A}(\epsilon)|$ then the matrix b is square and

$$B_{In}^{-1} = \sum_{\alpha \in \mathcal{A}(\epsilon)} E_{\alpha I} b_{\alpha n}^{-1}. \tag{3.18}$$

Alternatively, an overcomplete representation can be considered where $N > |\mathcal{A}(\epsilon)|$, in which case b^{-1} must be replaced by the Moore-Penrose pseudoinverse, $b^{+}(\epsilon)$.

The eigenvalue cut-off, ϵ, is often expressed in terms of the tolerance factor ϵ/λ_1, where λ_1 is the largest eigenvalue. In STRF calculations the optimal tolerance factor is determined through cross validation. Here, we have used a tolerance value of 0.004. This is within the range of tolerance values used in the calculation of actual Field L STRFs [67, 58] and gives 20 dimensions: $|\mathcal{A}(\epsilon)| = 20$, (see Figure 3.2). This value is sufficient to remove noise while still allowing an accurate reconstruction of

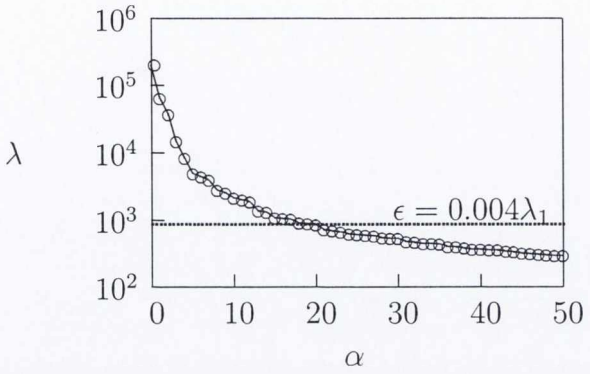

Figure 3.2: $\mathcal{A}(\epsilon)$ is the set of α such that $\lambda_\alpha > \epsilon$. Here $\epsilon = 0.004\lambda_1$. The contributions due to low-eigenvalue dimensions are ignored.

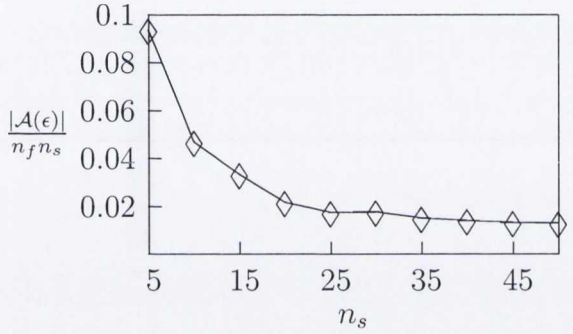

Figure 3.3: Proportion of dimensions above tolerance as a function of sample width ($\epsilon = 0.004$). Longer samples contain a higher degree of temporal correlation, and so have proportionally fewer high variance dimensions. Here $|\mathcal{A}(\epsilon)|$ is the number of significant dimensions, and $n_f n_s$ is the total number of dimensions in the stimulus space.

the stimulus, with more than 90% of stimulus variance explained.

It should also be noted that $|\mathcal{A}(\epsilon)|$ is dependent on the length of the samples chosen, since longer samples will display a higher degree of temporal auto-correlation, and hence will have a higher proportion of noisy, low-eigenvalue dimensions. Figure 3.3 shows the proportion of dimensions above tolerance as a function of sample length.

3.3 Results

We apply these methods to an ensemble of 20 zebra finch song spectrograms, each of one to two seconds duration. For suitable choices of ϵ and μ, we obtain a set of

optimal kernels sharing many of the observed characteristics of STRFs in the zebra finch auditory forebrain.

Figure 3.4 shows the set of optimal kernels of length 50ms calculated for $N = |\mathcal{A}(\epsilon)| = 20$. Though qualitatively differing from Field L STRFs in a number of respects, these kernels nonetheless display certain similarities with neuronal receptive fields: there are excitatory and inhibitory peaks on similar scales to those found in Field L STRFs, with both excitatory and inhibitory regions having similar amplitude, and kernels are localized in space and time, though possibly not as markedly localized as some STRFs of experimentally observed cells. Many of the sparse kernels show sensitivity to complex features such as frequency stacks, which are a common feature of zebra finch song. These kernels display somewhat similar tuning to many found in Field L of the zebra finch forebrain, though it should be noted that the multiple peaks obseved in these kernels are not common to the majority of Field L STRFs. Importantly, though these kernels do differ from Field L STRFs, it appears that those similarities which are observed increase with sparsification of the system, and are not observed in non-sparse filters.

Furthermore, we can quantitatively characterise the sparse kernels using a number of spatial parameters, and compare these values to those obtained from auditory STRFs. Parameters commonly used to characterise STRFs include the width of the largest peak in both time and frequency directions, W_t and W_f; the peak frequency, F_{peak}; the time to the largest peak T_{peak}; the quality factor, Q; the best modulation frequency, BMF and the spectral-time separability, SI.

These values, as calculated from the sparse kernels, agree well with those found in several studies of the avian auditory forebrain [76, 45, 27, 67] (See Table 3.1). The observed range of peak frequencies, F_{peak} closely matches that found in Field L STRFs, as do the seperability index, SI, and quality factor, Q.

The sparse kernels exhibit fine spectral tuning with localized peaks of average width W_f =1.1 kHz, and temporal tuning with W_t typically in the range of 10 -

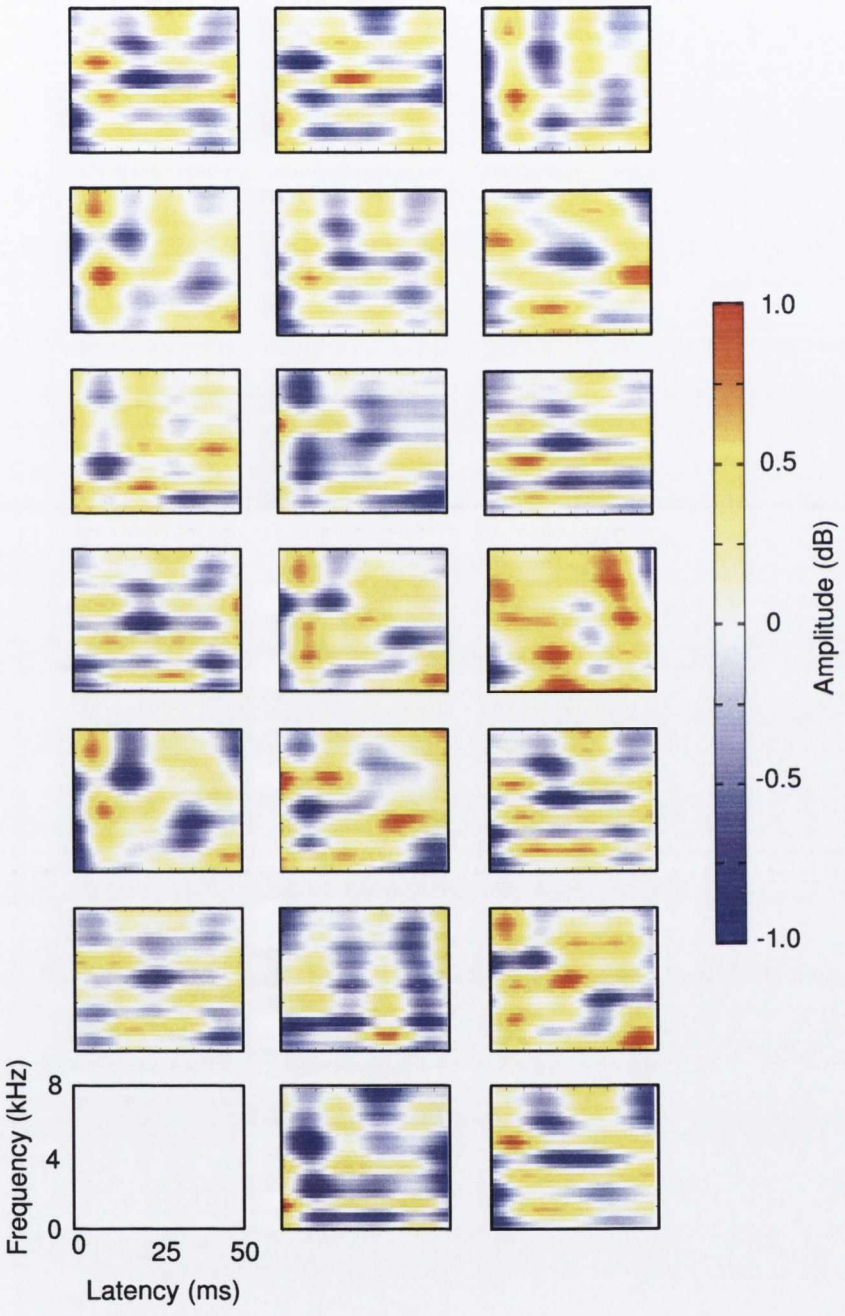

Figure 3.4: Set of 20 learned optimal filters calculated using a complete representation with $\mu = 150,000$ and tolerance value 0.004. Amplitude is shown on a colour scale from blue (lowest) to red (highest).

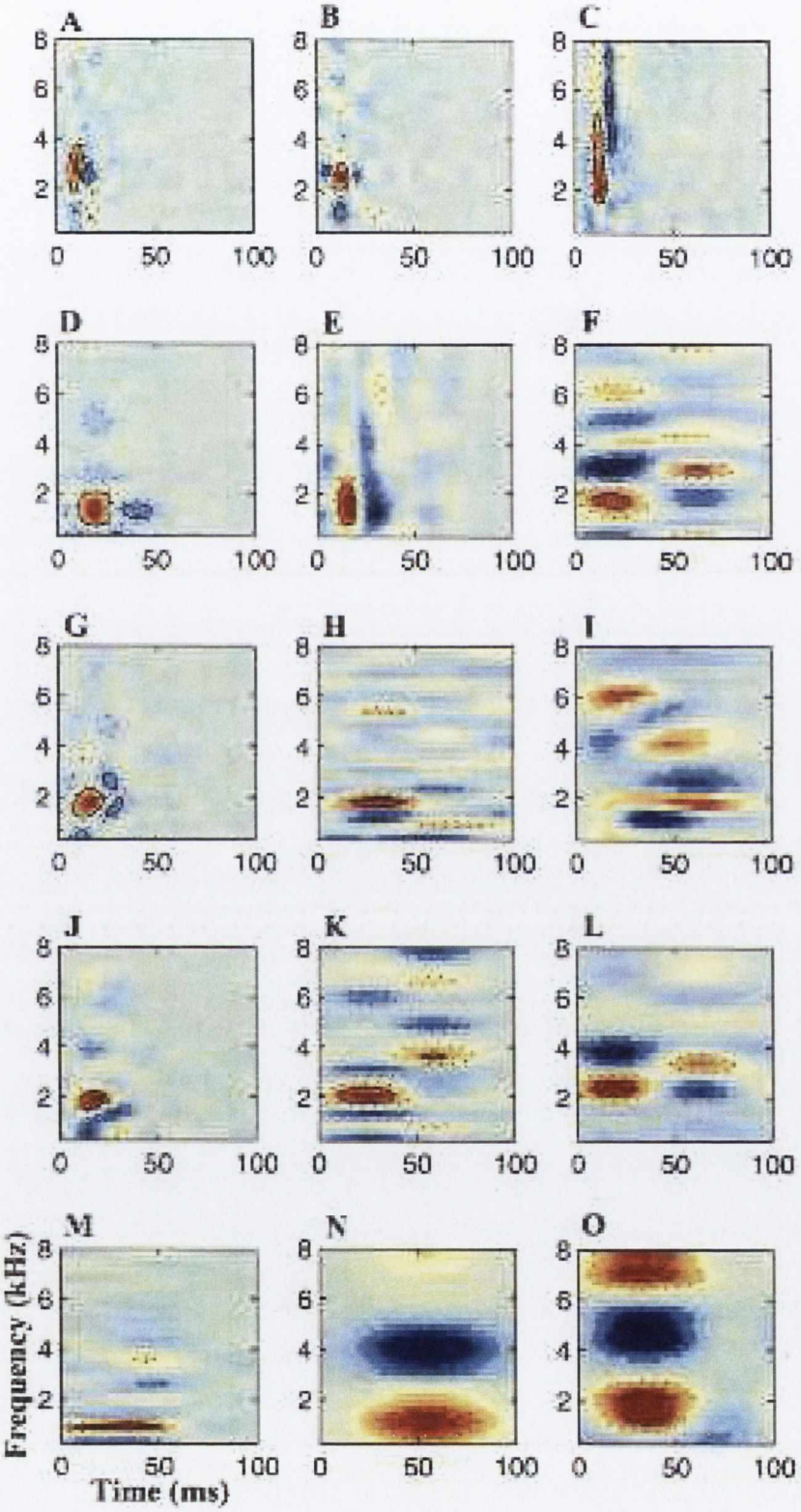

Figure 3.5: Field L neuronal STRFs as calculated using the STRFPak algorithm. Figure adapted from paper by K. Sen et al. [58].

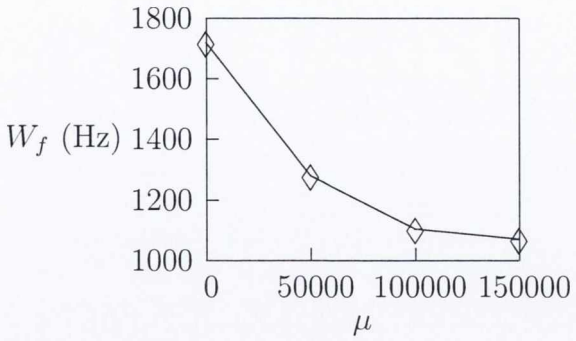

Figure 3.6: Average spectral peak width, W_f as a function of sparseness parameter. Sharp spectral tuning arises from increased sparseness in the system.

20ms (mean value 14.2ms). The kernels show little variation in peak widths, and W_f appears largely independent of peak frequency. Interestingly, W_f is seen to decrease as a function of the sparseness parameter, μ, as shown in Figure 3.6, suggesting that localized kernels arise as a result of sparsification.

The sharpness of the spectral tuning is measured by the quality factor, Q, defined as the ratio of the peak frequency to the width: $Q = F_{\text{peak}}/W_f$. Values of Q are in the range 1 - 5 (mean value 2.9), matching the findings of Theunissen et al. [67].

The best modulation frequency, BMF, is a measure of the AM frequency to which a neuron is best tuned, and is obtained from the Power Spectral Density of the linear kernel. The BMFs of individual sparse kernels were in the range 0 - 40Hz, with 90% of sites having BMF $\leq$ 20Hz (resolution 20Hz), indicating a strong preference for low frequency amplitude modulations, as seen in auditory STRFs [58]. The overall BMF of the set of optimal kernels was obtained by concatenating peak timeslices of all the sparse kernels. This gives an overall BMF value of 8Hz (resolution 1Hz).

Spectral-temporal seperability is measured by the SI value, obtained from the Singular Value Decomposition (SVD) of the STRF [58]. It is

$$SI = \frac{\rho_1^{'}}{\sum_{i=1}^{n-1} \rho_i^{'}} \tag{3.19}$$

where $\rho_i^{'} = \rho_i - \rho_n$, and ρ_i is the i^{th} singular value. As in previous studies [58],

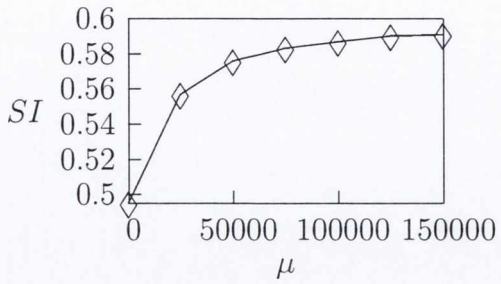

Figure 3.7: The average separability of the 20 sparse kernels increases with the sparseness parameter μ.

we choose $n = 4$ since the majority of features in the sparse kernels are accurately reconstructed from the first three singular values. As is the case with actual Field L neuronal STRFs, kernels are obtained with a wide variation in separability, ranging from relatively complex inseparable kernels to simpler, roughly separable kernels. Values of SI are in the range 0.43 - 0.84, with a mean value of 0.59 for the kernels shown in Figure 3.4. In general, we observe that the average separability of the sparse kernels increases as a function of the sparseness parameter, μ, (Figure 3.7) indicating that separability arises as a consequence of sparse coding.

For comparison, we also calculated a set of non-sparse kernels by setting $\mu = 0$ (Figure 3.8). These kernels exhibit significantly broader tuning than our sparse kernels and more closely resemble PCA kernels than auditory STRFs. Peak frequencies, F_{peak} are not restricted to low frequencies, occuring over the range (250 - 7750Hz), while peaks are broader in both time and frequency directions, with mean values $W_f = 1.7$kHz, $W_t = 20$ms and $Q = 1.4$. These kernels displayed statistically significant differences in both SI (t-test: $t = 3.1$, significance level: 1.73) and Q ($t = 4.1$) from our sparse kernels.

In addition, in order to rule out ensemble effects, the calculation of the sparse basis was repeated using new song recordings not used in the initial calculation. The inclusion of this new song data was found to have no significant effect on the results.

Furthermore, we applied our algorithm to an ensemble of low noise human voice

recordings and calculate the corresponding sparse kernels (Figure 3.9). As with the non-sparse filters mentioned above, these filters differ significantly in their tuning from those calculated for birdsong (t = 5.2 for Q, t=3.8 for SI) and from Field L STRFs. This disimilarity further supports the hypothesis that the tuning of Field L neurons is specifically adapted to encode con-specific song. To better illustrate this, we calculate the standard deviation of the distributions of SI and Q for each of our three filter sets, and use this to quantify the deviation of the Field L filter mean from the mean of each of these sets. As can be seen, the deviation for our sparse filters is significantly smaller than for our two control filter sets. However, in the absence of more detailed Field L data, we lack a suitable statistical model by which to further analyise the significance of our prediction.

Table 3.1 below summarises the tuning properties of each of the filter sets. Table 3.2 shows the deviation of the field L mean values for Q and SI from the mean of each calculated filter set.

Parameter	Field L	Sparse	Non-Sparse	Voice
F_{peak}(Hz)	375-5125	750-5250	250-7750	250-3750
Q	0.4-7.8 (2.5)	1.4-4.8 (2.9)	0.3-3.9 (1.4)	0.3-3.0 (0.93)
SI	0.49-0.83 (0.66)	0.43-0.84 (0.59)	0.38-0.67 (0.5)	0.41-0.75 (0.52)
BMF(Hz)	5-30 (15)	0-40 (8)	0-20 (10)	0-40 (10)
W_f(Hz)	n/a	850-2200 (1100)	700-2750 (1700)	250-3750 (1300)
W_t(ms)	n/a	10-20 (14.2)	11-37 (21.4)	3-9 (6.6)

Table 3.1: The range of STRF parameter values obtained from each of the three sets of calculated kernels, compared to those for Field L STRFs as found by Theunissen et al. [67], (F_{peak}) and for subregion L3 STRFs as given by Sen et al. [58], (Q, SI, BMF). Mean values are shown in parentheses.

Parameter	Sparse	Non-Sparse	Voice
ΔQ	$0.4 = 0.3\sigma$, $\sigma = 1.3$	$1.1 = 1.1\sigma$, $\sigma = 1$	$1.57 = 1.4\sigma$, $\sigma = 1.1$
ΔSI	0.0 7= 0.6σ, $\sigma = 0.11$	$0.16 = 2.3\sigma$, $\sigma = 0.07$	$0.14 = 1.2\sigma$, $\sigma = 0.12$

Table 3.2: The deviations, ΔQ and ΔSI of field L mean parameter values from predicted mean values for our three filter sets. In each case, σ is the standard deviation in the given parameter for the corresponding filter set.

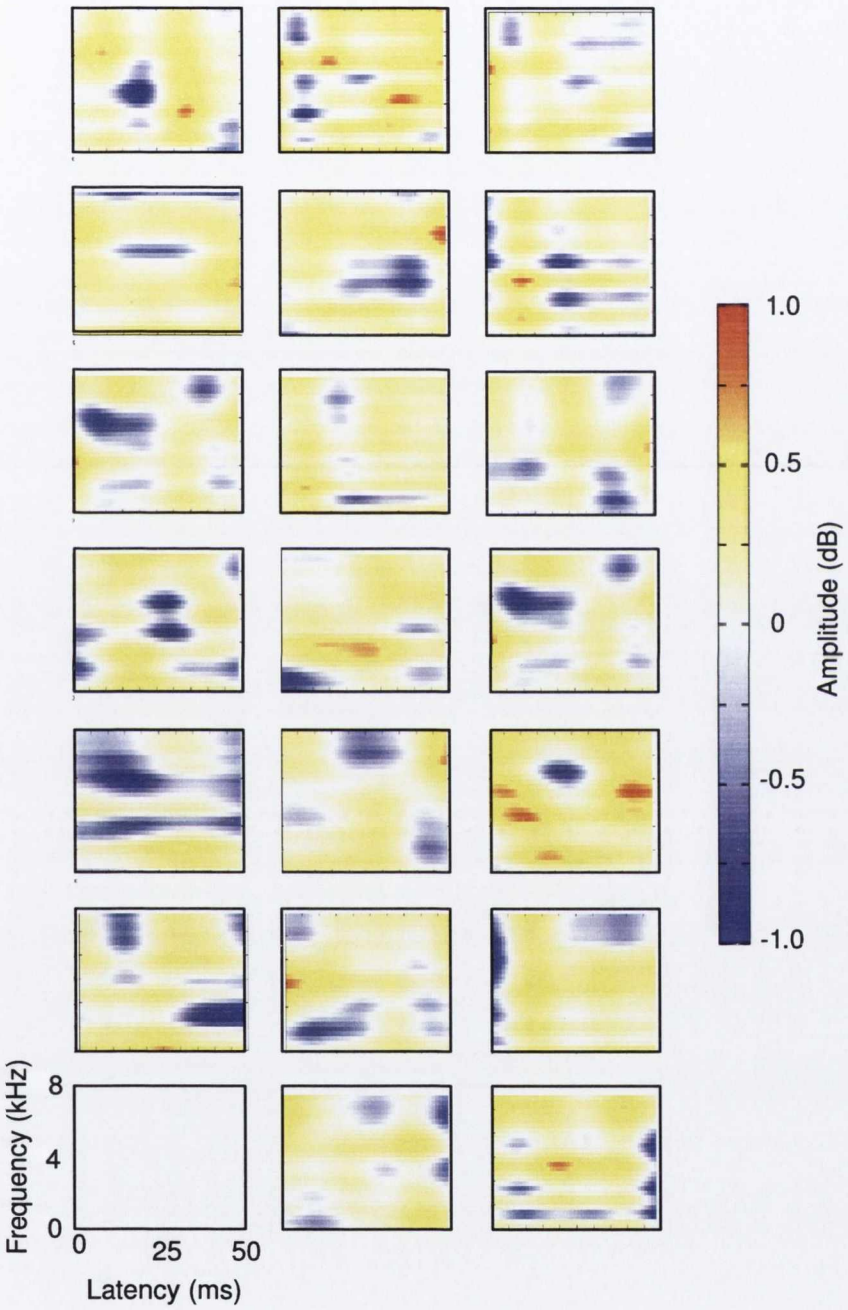

Figure 3.8: A set of non-sparse filters for our ensemble of birdsong.

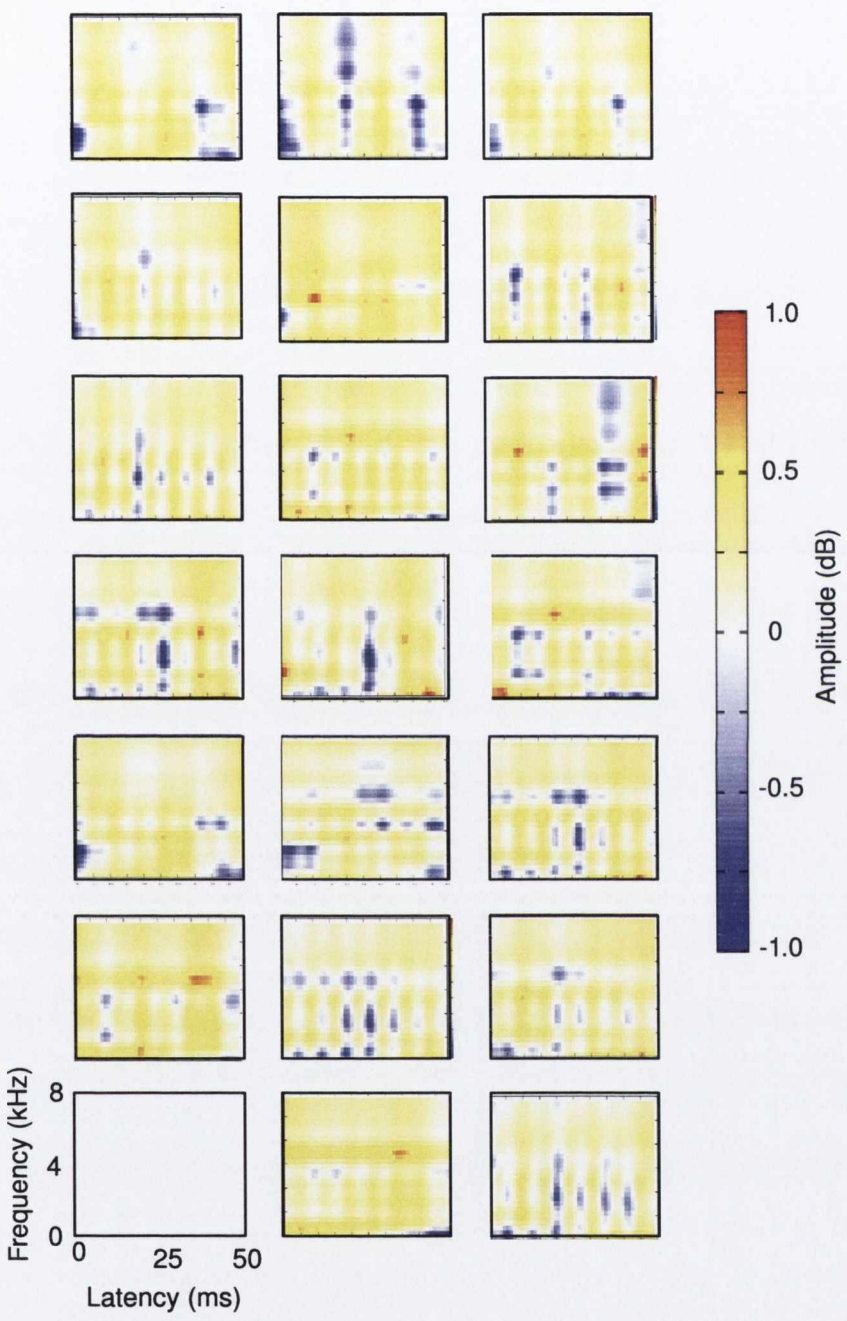

Figure 3.9: Sparse filters calculated for an ensemble of human voice recordings. These recordings were made using a Pearl CC 30 microphone in a semi-anechoic chamber, at a sampling rate of 44,100Hz. The text used was Jonothan Swift's *A Modest Proposal*. This text is in the public domain. Both the text and the original WAV files can be found at http://www.maths.tcd.ie/mnl

3.4 Discussion

The modified Olshausen-Field type algorithm described above identifies a sparse structure of dimension $|\mathcal{A}(\epsilon)|$ within the song spectrograms. We generated a system of STRF-like linear kernels which accurately and efficiently encode this structure. Comparison with neuronal receptive fields from the zebra finch, and with non-sparse filters shows that the sparseness constraint produces kernels which more closely resemble Field L STRFs. In addition, these sparse kernels differ significantly from kernels calculated for human voice recordings, which share little similarity with neuronal STRFs. The relative similarity between the sparse kernels and neuronal STRFs from the zebra finch suggests that the zebra finch auditory pathway is well adapted to encode the sparse structure of birdsong. In particular, the fine spectral tuning and localized peaks characteristic of many Field L STRFs are seen to arise in the sparse kernels as a consequence of sparsification. Similarly, greater separability is seen to arise from increased sparsification of the system. By comparison, both the set of non-sparse kernels and the sparse kernels calculated for human speech differ significantly in their tuning parameters from zebra finch auditory STRFs. This supports our hypothesis that the tuning is specifically optimised to encode conspecific song.

The main result here is the comparison of the sparse kernels with experimentally measured STRFs. In order to make this comparison, it is necessary to regularize the calculation. There are three reasons for this. Firstly, the biologically relevant timescale appears to be quite long, at about 50ms: as shown in Figure 3.3, longer samples possess a lower proportion of significant dimensions. Secondly, the corpus of stimuli we consider is limited to bird songs. It would be tempting to add other natural sounds to sample other stimulus dimensions, however, since sensory neurons are non-linear [41, 18, 67], the sparse kernels would be less relevant to the electrophysiological experiments which were performed using clean songs in an acoustically isolated environment [58]. Finally, the sparse kernels are computed by inverting the

sparse basis, potentially allowing noise to dominate the result.

Efficient or sparse coding certainly seems to be one of the primary goals of early visual processing [8, 47, 5, 71] and there is reason to believe that the same is true for auditory systems. Lewicki, for example, considered a sparse basis for an ensemble of natural sound waveforms composed of animal vocalizations and environmental noises [38] . Interestingly, for a specific mixture of sounds he found that this sparse basis has similar tuning properties to the fibres of the auditory nerve. Since the focus is on an earlier stage of sound processing, far shorter, 8ms, samples are used and the basis is not inverted; for this reason regularization is not required and so this calculation differs from ours, though the conclusion is very much in the same spirit. Furthermore, Smith and Lewicki [61] have shown that such sparse codes yield extremely efficient representations of acoustic signals.

In the specific case of birdsong, the idea that the receptive fields are adapted to song is supported by Woolley et al. [74], who produced is a comparison between the tuning properties of cells and the statistical structure of the songs themselves. Recent modelling of avian auditory areas by Blattler and Hahnloser [13] also suggests that sparse coding in Field L could play a role in higher level avian auditory processes such as song selection.

The results presented in this chapter suggest that there does in fact exist such a sparse coding in Field L, and imply the existence of a sparsifying interaction between Field L cells. However, the nature of this interaction remains obscure. It seems unlikely that a direct gradient descent of the type described here could be implemented in a realistic neural network. Instead, sparsification is assumed to come about as a result of a locally inhibitory interaction between cells. In the next chapter, we describe a simple neuronal network model in which sparsification arises through inhibitory competition between cells.

Chapter 4

Network Models of Sparse Coding

Any theory of primary sensory coding represents a trade off between the competing demands of representational accuracy and computational efficiency. Local codes, in which each neuron responds to one unique stimulus and the average activity of each cell is near zero, constitute the ideal in coding efficiency. However, such codes are unable to represent a large range of stimuli. Densely distributed codes - in which all cells are highly active - allow for much greater representational capacity, but are energetically costly and less obviously subserve computational goals such as recognition [50, 23]. The theory of sparse coding represents a compromise between these two ideals in which commonly occuring stimuli are represented by a small number of cells, while allowing for a denser representation of rare inputs. An effective sparse coding system must identify the particular sparse components within the corpus of natural stimuli to which it is adapted. Hence, sparse coding is often regarded as a form of independent component analysis (ICA). [7, 49]

Sparse coding may be one of the central tasks of primary sensory processing [22, 38, 24]. Studies of both auditory and visual systems [48, 71, 47, 72] have shown that primate V1 and avian Field L areas are well adapted to allow sparse representations of natural stimuli, and indeed, we have shown in the preceeding chapter that Field L may perform a sparse coding of birdsong. However, though these studies

demonstrate that sparse representations are possible, they do little to demonstrate how this sparseness may arise through neuronal interactions. Specifically, it remains unclear how an analogue of the ICA algorithms and gradient descent methods employed in these studies can be implemented in a biologically realistic neuronal setting.

Representations in V1 and Field L are highly overcomplete, with both of these areas having many times more neurons than their respective input layers, the retina and cochlea. Hence, there are more cells in these layers than there are independent dimensions in their input, and so a simple, feed-forward projection of activity from the input layer would result in a highly dense representation, in which many cells with neighbouring receptive fields are simultaneously active. If this were the case, the response would not be sparse. Even when the sensory input contained only a few sparse components, cells coding for distinct but nearby components would also fire, resulting in a redundant representation. Therefore, it is likely that, to achieve an efficient sparse representation, several cells compete to represent each sparse component of the input signal, with those cells which give the most accurate sparse representation winning out.

Sparse coding studies usually rely on gradient descent methods to determine the best representation for a given stimulus, as we have done in the previous chapter [24, 47, 71]. Alternatively, one can artificially enforce direct competition between cells [55]. However, if sparse coding is to be considered a reasonable theory of sensory coding, it should be shown that sparse representations can arise naturally through neuronal interactions. Similarly, it should be demonstrated that the sensory code can be dynamically learnt by training on the relevant corpus of stimuli.

Here we model a neuronal network in which sparse coding arises naturally in a layer of excitatory neurons through interaction with a layer of inhibitory interneurons. This network rapidly and efficiently sparsifies its response to stimuli that have an underlying sparse structure, while simultaneously learning an optimal code for the corpus. The network also displays ordering of its receptive fields, analagous to

retinotopy in the visual system, or tonotopy in the auditory system [76, 45].

4.1 Sparse Coding Models

In standard rate models of sensory coding, the neuronal representation of a stimulus S at time t is a linear combination over a set of component vectors, B_n, corresponding to the available neurons with the coefficients given by the firing rates, $r_n(t)$, of the cells:

$$\tilde{S}(t) = \sum_n r_n(t) B_n \tag{4.1}$$

Rather than seek the most accurate representation, the optimal response in a sparse coding model is the one which minimises an objective function combining the error in the stimulus reconstruction and the density of the representation, [47]

$$E(t) = (S(t) - \tilde{S}(t))^2 - \mu \sum_n C(r_n(t)) \tag{4.2}$$

where the first term penalises inaccuracy in the stimulus reconstruction and the second penalises non-sparse representations, through a sparseness cost function, $C(\cdot)$. $C(\cdot)$ can be chosen as any suitable function satisfying Jensen's Inequality, so that the energy function, $E(t)$ favours representations with a small number of strongly active cells.

In our previous work on sparse coding [24], described in Chapter 3, sparse representations for auditory stimuli were found by minimising this cost function using conjugate gradient descent. Here, instead, we achieve a similar trade-off through network interactions.

Rozell et al. [55] have shown that sparse responses can arise in artificial networks of thresholded neurons through local competition. We achieve a similar sparsification through realistic interactions between excitatory and inhibitory layers. Upon presentation of each new stimulus the network rapidly reaches an equilibrium state

in which a few excitatory cells 'win', accurately encoding the stimulus, while others 'lose' and are inhibited by the interneurons. Moreover, using simple Hebbian rules, this network dynamically learns the sparse structure of a stimulus set.

4.2 Methods

4.2.1 Network Model

Our network consists of three layers. The first layer, **S**, simulates the sensory input. These excitatory cells are not modelled dynamically, rather, the input firing rates are chosen randomly, but in such as way that the input has a sparse structure. There are two network layers, an excitatory layer, **E**, with lateral nearest-neighbour connectivity. **E** should be thought of as determining the output which would then feed-forward along the sensory pathway. Finally there is an inhibitory layer, **I**, of interneurons, which regulate activity in excitatory layer **E**. In the two network layers, **E** and **I**, the neurons are modelled as thresholded leaky integrating neurons.

The signal representation is given by the firing rates of the cells in the excitatory layer, **E**. These cells receive excitatory inputs from the input layer, **S**, and through lateral connections from other layer **E** cells. They also receive inhibitory input from the inhibitory layer, **I**. The layer **I** cells are excited only through connections from layer **E**.

For simplicity, we consider a small model encoding simple vector inputs. The input layer **S** consists of three cells, corresponding to orthogonal unit vectors. The stimulus therefore consists of non-negative combinations of these three components. These input cells feedforward to layer **E**, consisting of eight excitatory cells with a periodic boundary. Output cells have excitatory connections, both to their nearest neighbours in the output layer, and to neighbouring cells in the **I** layer. The **I** layer consists of four cells, each of which has two-way connection to each of the three nearest cells in the excitatory layer.

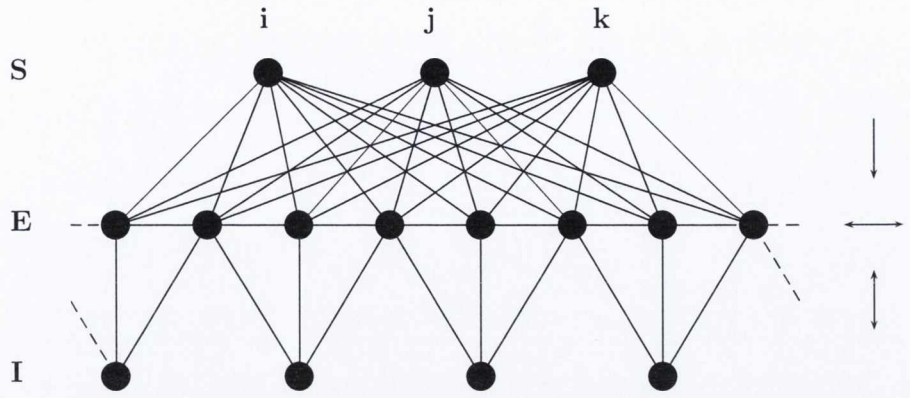

Figure 4.1: Schematic of network connectivity. We have two layers, **S** and **E**, of excitatory cells, and one layer, **I**, consisting of inhibitory interneurons.**S-E** connections are one way, while **E-E** and **E-I** connections are two-way.

This network connectivity, while highly simplified, models several important characteristics of primary sensory cortex, and could be easily scaled up to allow encoding of complex higher-dimensional stimuli. In particular, individual cells in our model may be seen as modelling the behaviour of localised populations, while the short range lateral connections and longer range inhibitory interactions reflect the 'Mexican hat' connectivity commonly observed in visual systems. This connectivity, as we will show, may bring about the desired rapid sparsification of output signals and, on a longer time scale, lead to retinotopic organisation of receptive fields through local entrainment and non-local repulsion, via a standard Hebbian learning rule for synaptic weights. Similarly, the competing effects of inhibitory and excitatory inputs may bring about oscillatory firing patterns, which are widely observed in biological neural systems.

4.2.2 Cell dynamics

In the thresholded leaky integrating neuron model [55] used for the **E** and **I** layers, a cell has an internal state described by its membrane potential, u, which is governed

by the ordinary differential equation

$$\tau \dot{u}_a(t) = \sum_i s_i(t) w^S_{ia} + \alpha \sum_{m \neq n} w^E_{ma} r_m(t) - \beta \sum_p w^I_{pa} r_p(t) - \gamma u_a(t) \tag{4.3}$$

where w^S, w^E and w^I are the synaptic weights corresponding to the **S**, **E** and **I** layers, respectively. The $i \in \{1, 2, 3\}$ index runs over the three **S** neurons, the $m \in \{1, \ldots 8\}$ index over the **E** neurons and the $p \in \{1, \ldots, 4\}$ runs over the **I** neurons. In the equation the index a can refer to a both **E** and **I** neurons. r_i, r_m and r_p are firing rates, the stimulus **S** firing rates r_i are an input determined by the corpus of stimuli. For layers **E** and **I** the cells activity determines its firing rate, r. This is given by a sigmoidal threshold function:

$$r_a(t) = \frac{u_a(t) - T}{1 + e^{-\rho(u_a(t) - T)}} \tag{4.4}$$

In the limit where ρ goes to infinity, this threshold sharpens, and becomes a simple step function, with

$$r_a(t) = \begin{cases} 0 & u(t) < T \\ u(t) - T & u(t) > T \end{cases} \tag{4.5}$$

Learning in the network occurs through changes in the synaptic weights, w, determined by a standard Hebbian update rule;

$$\Delta w_{ab} = \delta \left\langle r_a(t) r_b(t) \right\rangle_t \tag{4.6}$$

where the average is taken over a much longer timescale than the rate at which the stimulus changes, and δ determines the learning rate. The input, excitatory and inhibitory weight vectors for each cell are then normalised after each update.

The firing rate response characteristics of a neuron are generally characterised by a receptive field (RF) [33], a vector which is convolved with the stimulus to give

an estimated firing rate. Here, the RF is given by the spike triggered average:

$$h_a = \langle r_a(t)s(t)\rangle_t \,. \tag{4.7}$$

4.2.3 Training data and learning

The simulation begins with random synaptic weights, w. The network is trained on a dataset consisting of three-dimensional vectors with non-negative components. These components correspond to the firing rates of the three layer **S** neurons. The dataset is constructed so as to contain three significant sparse directions. At each new presentation, one sparse vector, v, is randomly selected and a stimulus vector is generated of the form

$$s_i = v_i + \mathcal{N}(0,\sigma)v_i^{\perp} \tag{4.8}$$

where $v^{\perp}$ is a randomly chosen unit vector perpendicular to v, and σ determines the noise level of the stimulus. Negative components resulting from the addition of noise are reset to zero.

Each presentation lasts for $\theta^P = 10\tau$ timesteps, where τ is the characteristic timescale of the neuron model, given in Equation 4.3. Learning takes place over a longer timescale, with the Hebbian update rule applied every $\theta^U = 500\tau$ timesteps.

4.3 Results

For appropriate choices of the neuron model parameters, α, β and γ (see Appendix B) the network proves to be highly adept at learning the sparse structure of the stimulus. It relaxes into a stable organisation on a timescale of the order of $10^5\tau$ timesteps. The network is seen to arrange itself into a locally redundant clusters of cells with similar RFs, corresponding to the sparse directions in the stimulus space. These clusters consist of between one and three neighbouring excitatory cells grouped around a single inhibitory cell. The closeness of this clustering is seen to

depend on the noise level, σ, of the stimulus.

This arrangement allows a very accurate sparse representation of our stimuli, with only one or two active cells. We also observe a number of mediatory cells whose RFs do not corresponds to sparse vectors. These cells, which are on average very weakly active, allow for the encoding of rare stimuli which do not lie along one of the sparse directions. Figure 4.2 below shows a stereographic projection of the RF vectors of the output excitatory cells, both before and after learning on a typical sparse dataset. As can be seen, these RF vectors quickly converge on the sparse vectors of the training data. Details of the sterographic projection used are given in Appendix C.

The network proves surprisingly robust at isolating widely seperated sparse dimensions of the stimulus, though this ability breaks down as the distance between sparse vectors decreases towards the noise level. In general, the network fails to distinguish sparse vectors v_i and v_j seperated by a distance $|v_i - v_j|$ less than 2σ. This is illustrated in Figure 4.3.

After learning, the network settles into a stable organisation, corresponding to the optimal sparse code for the training dataset. Subsequent stimuli are represented sparsely as a result of competition between cells. Both local and non-local competition is observed. Initially, particularly during the learning phase, several clusters may compete to represent a novel stimulus, with one, or possibly two clusters quickly becoming dominant. We also observe a longer lasting competition between similar cells within the dominant cluster, with the cell whose RF most closely approximates the stimulus winning out.

Figure 4.4 shows the internal potential, u, of output layer cells in the post-learning phase as a function of time after the onset of stimulus presentation. Here we can observe the distinct clustering of layer **E** responses as well as local competition between neighbouring cells in the dominant cluster. Oscillatory behaviour, which is typical of excitatory-inhibitory networks, can also be observed with a period of

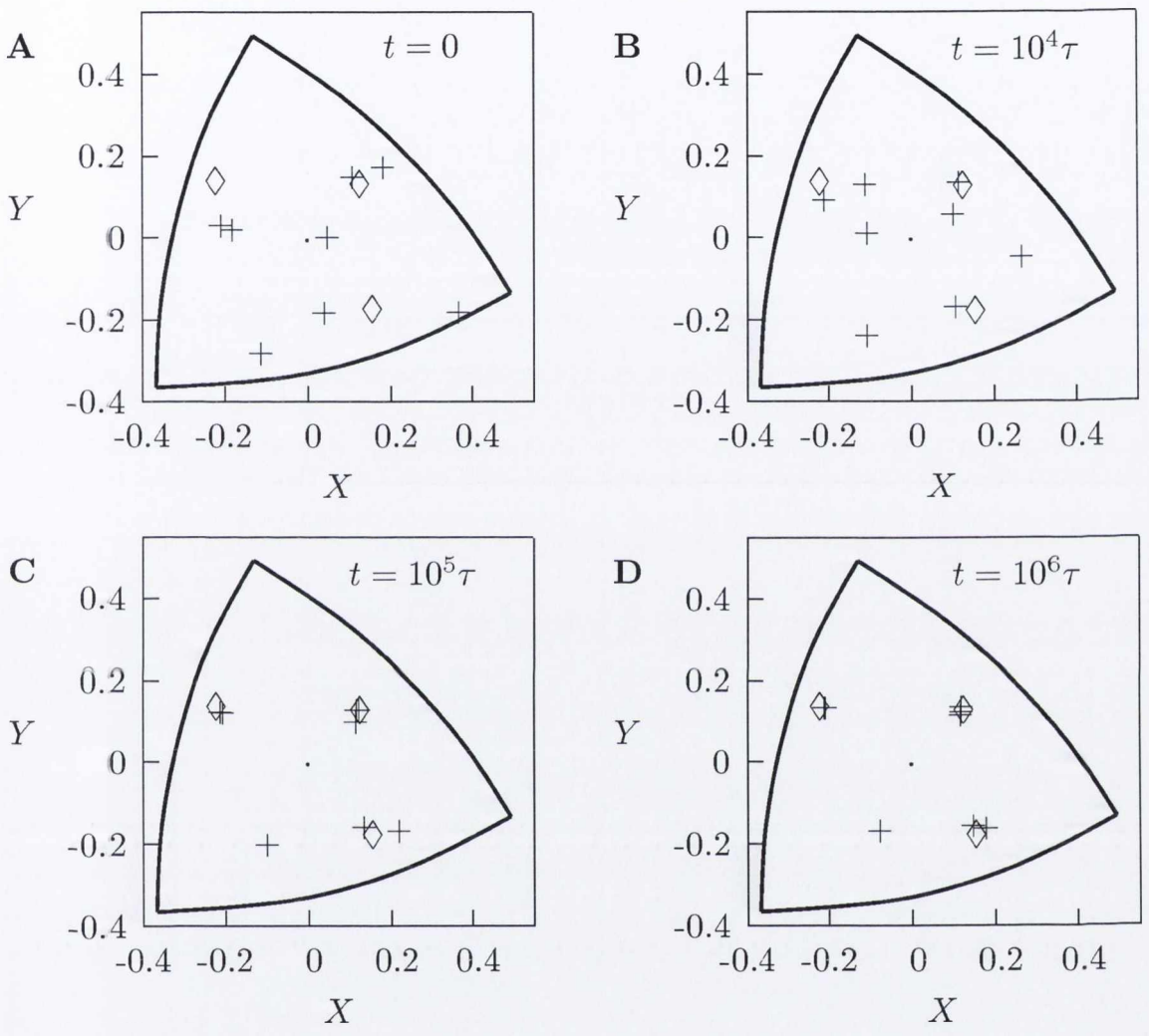

Figure 4.2: RFs of cells in the output layer are shown before (**A**), during (**B**,**C**) and after (**D**) learning. We use a stereographic projection of the first octant of the unit sphere, projecting from the point $p = (-1, -1, -1)$. The point $(X = 0, Y = 0)$ in the projection represents the $(1, 1, 1)$ direction on the sphere. The curved boundaries shown correspond to the intersections of the unit sphere with the $x = 0$, $y = 0$ and $z = 0$ planes. The sparse directions within the training stimulus are represented by diamonds, while RF vectors of layer **E** cells are shown with crosses. The majority of RF vectors converge on the sparse directions after $\sim 10^5\tau$ timesteps.

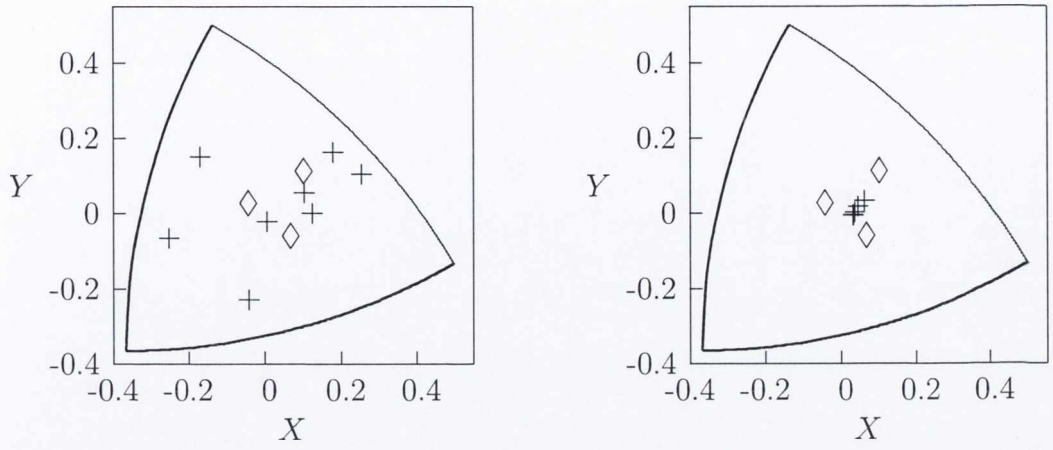

Figure 4.3: The network fails to accurately distinguish sparse vectors with a seperation of less than approximately 2σ. The stereographic plots above show the initial and final network states for a set of sparse vectors with average seperation $|v_i - v_j| \simeq 0.2$, where $\sigma = 0.15$. Output RFs converge on a point roughly equidistant between the two most closely positioned sparse vectors.

approximately 2τ. The corresponding firing rates are given by the value of u above the threshold, T.

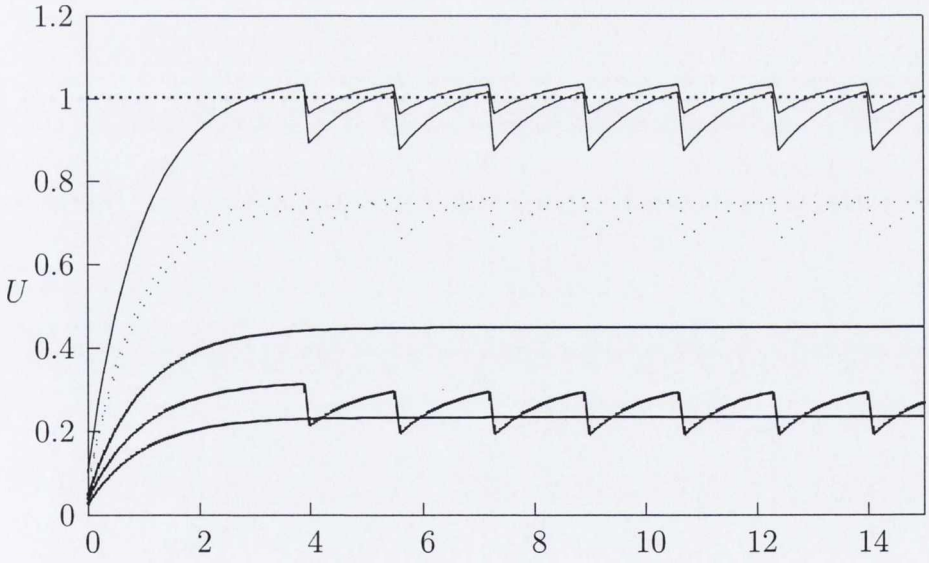

Figure 4.4: Internal potential u of excitatory cells as a function of time after stimulus onset. The threshold value is $T = 1$. In the limit as $\rho \to \infty$, firing rates are given $u - T$. Three distinct clusters can be seen, as well as a single intermediate cell. The two cells whose RFs best match the stimulus compete, with only one becoming significantly active. In addition, we can observe firing rate oscillations with a period of 1.8τ. Line types reflect distinct RF clusters, and the threshold value $u = 1$ is shown.

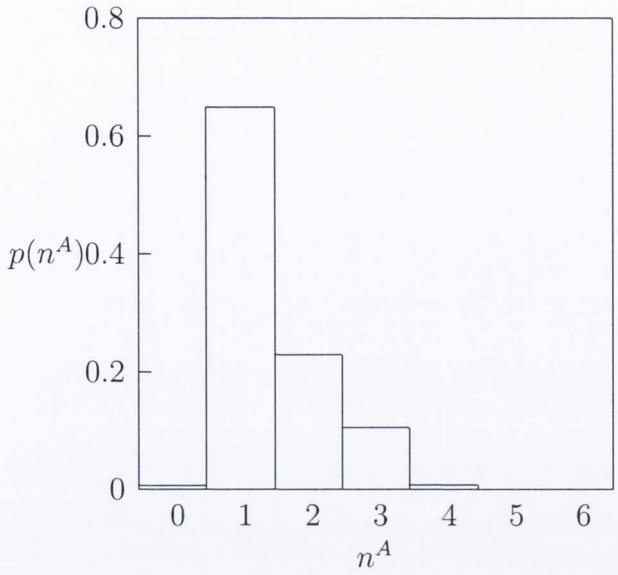

Figure 4.5: Representation density distribution. n^A is the number of active cells in the equilibrium state for each stimulus. Though the network favours sparse representations, denser representations are possible for rare inputs.

Importantly, the network allows both sparse representations of commonly occuring stimuli, and non sparse representations of rarer inputs, with several cells active in the equilibrium state in these cases. This confirms that the network does in fact achieve a sparse code, as opposed to a pure local code for the sparse directions.

Furthermore, each output cell is active on average less than 15% of the time, which demonstrates that the network achieves sparseness both across the population, as demonstrated in Figure 4.5, and in the firing of individual cells.

Interestingly, in addition to sparsifying firing rates, our model reproduces other non-linear behaviour which has been observed in sensory areas. Narayan et al. [46] have shown that the response of zebra finch auditory neurons to natural stimuli is altered in the presence of a noisy masking signal. They measure the response of Field L neurons which are highly tuned to specific syllables of their bird song stimulus and they find that in the presence of an auditory mask the response of these cells to the target syllables is reduced compared to the response to the unmasked signal, even where the intensity of the target syllables is conserved in the masked signal. Like sparsification, this behaviour can not be explained by a simple linear filter model of

sensory processing.

We reproduce this phenomenon in our model by presenting stimuli consisting of the sparse vectors on which the network has been trained, masked with random noise. Given a sparse vector, v, we produce stimulus vectors of the form

$$s_i = v_i + \eta v_i^{\perp} \tag{4.9}$$

where $v^{\perp}$ is a randomly chosen unit vector perpendicular to v, and η is a parameter controlling the noise level. For comparison, we also produce a set of noise free stimuli of the form $s_i = \sqrt{1+\eta^2}v_i$, such that for a given value of η, the total signal strength is the same in both cases. We examine the response to these stimuli of layer **E** cells whose RFs are tuned to the sparse vector v. As can be seen in Figure 4.6 below, whereas the response to the noise free stimulus increases linearly with signal strength, in approximately 90% of cases the average response to the noisy stimuli falls off as a function of η.

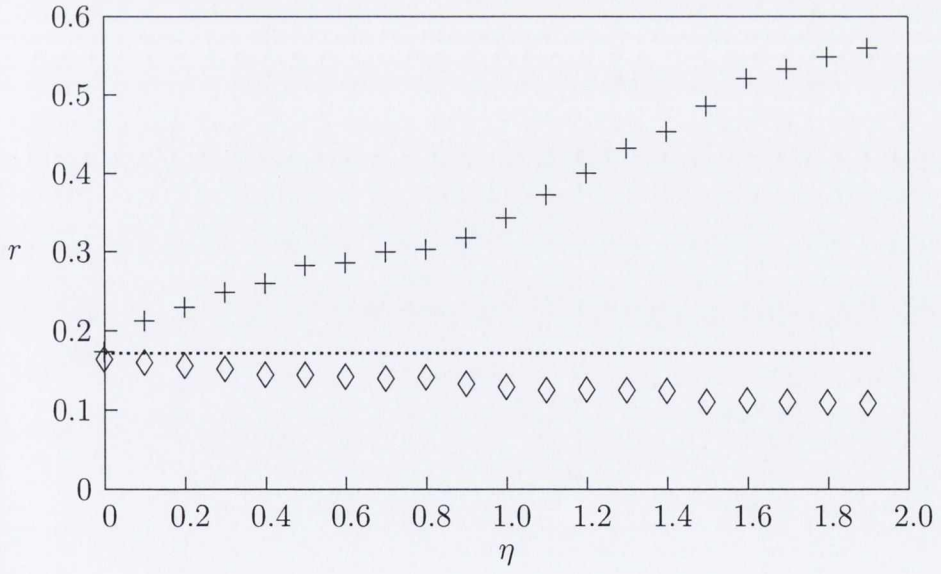

Figure 4.6: Average firing rate response as a function of the noise parameter η for a typical layer **E** cell in the post-learning phase. Response to noisy stimuli is represented by diamonds. For comparison, the reponse to the noise free stimulus set is represented by crosses. For each value of η, the signal strength is the same for both data sets.

4.4 Discussion

Despite numerous simplifications, the network model presented appears to demonstrate a viable mechanism for sparse coding in sensory systems. The limitations of such a model are obvious: its size does not allow the encoding of complex stimuli, the **E-I** connectivity used is somewhat artificial, and the rate based neuron model entirely neglects spike timing effects. Nonetheless, the model is surprisingly effective in reproducing numerous salient features in the behaviour of sensory cortex. In particular, the sparsification of responses and the diminished response to masked stimuli suggest that even apparently complicated soft linearities can be reproduced through simple and biologically realistic rules [46], namely, thesholding of neural reponses, mexican hat connectivity, and Hebb-like plasticity.

Furthermore, though limited in application by its size, such a network may quite easily be scaled up to allow encoding of higher dimensional auditory and visual stimuli. This presents interesting possibilities for investigating the neuronal mechanisms underlying some surprising effects observed in sensory cortex. In the songbird auditory system, for example, Anterior Forebrain cells have been found to display a remarkable degree of selectivity for specific song motifs [62, 20] while in the visual system, V1 cell responses have been shown to depend on fourth order correlations in the stimulus [70]. It may be of great interest to determine whether these effects can arise in such networks through known principles of connectivity and learning or, if not, what conditions are necessary to bring them about.

Chapter 5

Discussion and Conclusions

5.1 Discussion of Linear Models

The fundamental notion underpinning much of this thesis is that of efficient coding, and in particular, the hypothesis that sensory systems should efficiently encode natural stimuli. This idea has been addressed at length in our discussion of sparse neural codes and sparsifying networks (Chapters 3 and 4). However, in order to usefully study sensory codes on a system scale, it is first necessary to derive a method of quantifying the response properties of a single neuron under stimulation. Here, as in much of the literature on sensory systems [58, 67, 65, 47, 32, 33, 3, 21, 74, 56], we have largely relied on classical linear models of neuronal response.

It is to some degree surprising that such linear models remain popular in describing the sensory cortex: it has been established by many studies [32, 57, 67, 37] that cortical neurons rarely respond linearly to sensory stimuli, while common theories about sensory coding - such as the sparse coding hypothesis discussed in previous chapters - predict explicitly non-linear responses. Conversely, however, there is significant evidence to support linear models of neuronal behaviour: neural firing rates in peripheral sensory systems have been shown to be well approximated by linear models [2, 34], while the majority of established neuron simulation models describe

membrane potential as being linear in the synaptic input [36, 14, 28].

The solution to this apparent paradox is perhaps implicit in the formulation of these neuron models, and indeed in the neuron model used in our own network simulation (Chapter 4). In all of these cases the neuron performs a roughly linear integration of the feed-forward input current. However, this linear behaviour is subject to interference due to threshold or near-threshold non-linearities, as well as the effects of excitatory and inhibitory inputs from other cells. In cortical areas, where neurons generally receive a higher number of lateral connections, we may expect these effects to be more prominent, causing the cell's behaviour to deviate further from linearity.

In light of this theory, though linear models may give poor predictions of the behaviour of cortcal cells, they may still be of much use in providing an informative description of the underlying tuning properties of a neuron. In this spirit, we have used such models in developing theories of auditory coding in the zebra finch. More generally, throughout this thesis, we have considered the neuron as a fundamentally linear integrator of input current, which may nonetheless exhibit non-linear behaviour as a result of threshold and network effects.

5.2 Summary

In Chapter 2, we reformulated the commonly used STRF model for auditory neurons in terms of a regularized inversion. In doing so we comapre several possible regularization techniques, and find that the most effective method untilises a dimensional reduction in the space of sound frequencies. This formulation is computationally and conceptually simpler that that used in previous studies, and avoids several approximations used in the standard deconvolution method. This method produced a modest improvement in the predictive power of the STRF when applied to novel stimuli. In addition, and just as importantly, this formulation makes possible the

explicit computation of an optimal sparse code for auditory stimuli.

In Chapter 3 we made use of our simplified STRF calculation to compute a set of predicted sparse coding filters for zebra finch song spectrograms. Using an Olshausen-Field type algorithm, we calculated a sparse basis for an ensemble of zebra finch songs by minimisation of a representational cost function. We subsequently obtained a set of optimal STRF-like filters by regularized inversion of this basis. These predicted filters shared certain similarities with STRFs of cells from zebra finch Field L which were not found in a set of optimal non-sparse kernels. In addition, similar calculations using human voice recordings failed to produce such similarities in tuning. This result suggests that the zebra finch primary auditory system may use a sparse representation of conspecific song.

Finally, in Chapter 4, we suggested a viable neuronal mechanism whereby a sparse coding can arise in biological sensory systems. We described a two-layer neural network consisting of a layer of excitatory feed-forward neurons with lateral excitation and a layer of inhibitory interneurons. Cell dynamics were simulated using a leaky inegrating rate model. This network rapidly sparsified its responses to stimuli through inhibitory competition, and dynamically learned a sparse code for a stimulus ensemble. The network also reproduced several characteristic features of sensory cortical neurons, including receptive field ordering, and response suppression in the presence of noise. Though highly simplified, this network models several important aspects of neural connectivity in sensory cortex, and the results here represent strong evidence that sparse codes may arise in sensory systems through the interplay of excitatory and inhibitory connections.

In summary, though cortical firing rates may be poorly predicted by linear models, we have attempted to demonstrate that these models are still of considerable value. Specifically, they allow an intuitive description of the stimulus to which a cell would preferentially respond in the absence of network effects and, in the case of the STRF model, provide a framework in which to formulate sparse codes. Here,

we have derived an improved version of such a linear model for auditory neurons and, using this model, have produced evidence of sparse coding in the zebra finch auditory pathway (Chapters 2 and 3). This result begs the question of how a non-linear effect such as sparse coding can arise in the framework of linear neurons. To address this question, we describe a realistic mechanism whereby sparse coding and other non-linear effects might arise in networks of linear integrating neurons through network interactions.

5.3 Remarks on Scientific Approach.

The work described here has several strengths associated with it. Specifically, we have tried, with a certain degree of success, to clarify and simplify a number of important but poorly understood concepts in neural coding while, at the same time, making a novel contribution to the theory of optimal coding of natural stimuli. For example, our work on calculation of auditory STRFs produced a new STRF formulation which is inuitively far simpler to understand, while producing comparable performance to the standard method. In addition our comparison of several regularisation methods provides valuable insight into the relative importance of spectral and temporal information in natural sounds.

Furthermore, in our work on network models of sparse coding, we have demonstrated a conceptually simple model of sparse coding which is nonetheless biologically plausible. More generally, though this is a largely theoretical study, we have made a virtue of mathematical simplicity and intuitiveness. We hope that this brings the advantage of making the work presented here more accessible to a general audience.

There are however, a number of limitations in the work presented here which should also be mentioned. Perhaps the most obvious of these is the lack of sufficient comparison to experimental results. The work on auditory STRFs presented in chapter 2 is somewhat compromised by the use of a very small training data set,

which may not be large enough to allow complete characterisation of neuronal responses. A similar problem affects our work on sparse coding in Chapter 3. In this case, a shortage of data on experimentally obtained STRFs severely limits the statistical comparisons possible between our predicted sparse kernels and the receptive fields of auditory neurons.

In a broader sense, we have taken a purely model based approach to the work covered here and, while this has its advantages, the absence of meaningful comparison to experiment may reduce the informativeness of some results. Such studies may benefit greatly from the input of an experimental collaborator.

5.4 Further Work

As previously discussed in the conclusion of Chapter 2, several possibilities exist for further improving the linear STRF model. It may be possible to further increase the prediction accuracy through an additional, seperate dimensional reduction in temporal space. This would, however, have the disadvantage of requiring optimisation of a second model parameter. A perhaps more interesting potential approach involves smoothing the spectrogram with a synapse-like temporal filter. Temporal smoothing has been shown to improve prediction accuracy in other STRF models, but is generally poorly justified, and also requires the choice of an appropriate filter, which may also require optimisation. Our proposed method avoids these problems by using a biologically inspired filter whose ideal width is already from experiment.

We also propose to further expand the network model described in Chapter 4 to enable the encoding of complex visual and auditory stimuli. This would allow verification of our hypothesis that simple inhibitory connectivity is sufficient to bring about sparse coding. It may also be of interest to investigate the predicted model receptive fields of excitatory and inhibitory populations in a sparse coding regime.

Appendix A

Zebra Finch Song Data and Neuronal Recording in Zebra Finch

Zebra finch song samples used in Chapters 2 and 3 were taken from an ensemble of 20 songs, previously used in the work of Theunissen et al. (2000) [67] and Sen et al. (2001) [58]. Songs are of average length 2.1s, and song waveforms recorded at a sampling rate of 44kHz. Spectrograms of these songs were produced at a spectral resolution of 250Hz.

Electrophysiological recordings used in the calculation of Field L STRFs were kindly donated by Kamal Sen and co-workers at the Natural Sounds and Neural Coding Lab in Boston University. A detailed description of experimental procedures is given in their paper, Sen et al., (2001). [58]. To summarise:

All recordings were made in anesthetized adult male zebra finches. Extracellular potential wave- forms were obtained using parylene-coated tungsten electrodes inserted into the neostriatum of the bird. Waveforms are first band-pass filtered, and then transformed into spike trains, using a window discriminator. Successive action potential profiles are compared to determine the number of units present in

the recording. Waveforms containing a single stereotyped spike profile are classified as single-unit recordings. Data used in this thesis is taken entirely from such single unit recordings.

Spiketrains obtained from extracellular waveforms are subsequently converted into spike arrival time format, with a sampling rate of 32kHz. For STRF caluclations these data are downsampled into spike count format at a resolution of 1kHz. In this format, entries correspond to the number of spikes observed in each 1ms time bin.

For the recordings used in Chapter 2, spiketrains are obtained for 10 repetitions of each stimulus song. Firing rates used in STRF calculations are found by averaging over the responses for each repetition.

Appendix B

Typical Values of Neuron Model Parameters

Neuron dynamics within our model are determined by the neuron parameters α, β and γ, which simulate the inhibitory, excitatory and leak conductances of the cell, as well as the threshold parameter, ρ, which determines the speed of transition at threshold. Typical values for excitatory and inhibitory cells are shown in Table 1. In addition, network learning behaviour is dependent on the learning rate paramater, δ, and the noise level, σ. Except where otherwise stated, all data and plots shown here were generated using values of $\delta = 0.005$ and $\sigma = 0.15$.

Cell type	α	β	γ	ρ
Excitatory	1	0.6	0.98	1000
Inhibitory	0	1	0.8	1000

Table B.1: Typical values of cell paramaters for layer **E** and **I** model neurons.

Appendix C

Stereographic Projection

The plots of output receptive fields in Figure 2 are generated using a stereographic projection. RFs in our model are represented by three dimensional unit vectors with non-negative components, and so can be expressed as points on the first octant of the 2-sphere. In order to illustrate the learning behaviour of these RFs, we seek a planar representation of our three-dimensional vectors. We can define a stereographic projection of the 2-sphere at a point p as a mapping

$$\phi : S^2 \backslash \{p\} \mapsto \mathbb{R}^2 \tag{C.1}$$

from the sphere onto the plane, where p is the projection point. This map is defined at all points on the sphere except p, and by convention the antipode of p is mapped to the origin.

Here, we seek a projection of the first octant of S^2, centered on the point $[1/\sqrt(3), 1/\sqrt(3), 1/\sqrt(3)]$, with projection point $[-1/\sqrt(3), -1/\sqrt{3}, -1/\sqrt{3}]$. Equivalently, and more straighforwardly, we can perform a rotation, ψ on our set of vectors, such that $\psi([1/\sqrt(3), 1/\sqrt(3), 1/\sqrt(3)]) = [0, 0, 1]$ and hence perform the projection from the south pole, $p = [0, 0, -1]$. A commonly used formula for this

projection is given by

$$\phi\left([x,y,z]\right) = \left[\frac{x}{1+z}, \frac{y}{1+z}\right]. \tag{C.2}$$

This generates a two-dimensional representation of our data with the centre point, $[1/\sqrt{(3)}, 1/\sqrt{(3)}, 1/\sqrt{(3)}]$ mapped onto the origin in $\mathbb{R}^2$. The boundary curves of the first octant, defined by the intersections of the $x-y$, $x-z$ and $y-z$ planes with the unit sphere can be similarly transformed to give the boundary curves shown in Figure 2.

Bibliography

[1] L. F. Abbott, J. A. Varela, Kamal Sen, and S. B. Nelson. Synaptic depression and cortical gain control. *Science*, 275(5297):221–224, 1997.

[2] E.D. Adrian and Y. Zotterman. The impulses produced by sensory nerve endings: Part ii: The response of a single end organ. *Journal of Physiology*, 61:151–171, 1926.

[3] A.M.H.J. Aertsen and P.I.M. Johannesma. The spectro-temporal receptive field. *Biological Cybernetics*, 42:133–143, 1981.

[4] J.J. Atick. Could information theory provide an ecological theory of sensory processing? *Network*, 3:213–251, 1992.

[5] R. Baddeley. An efficient code in v1? *Nature*, 318:560–561, 1996.

[6] H. Barlow. Possible principles underlying the transformation of sensory messages. In *Sensory Communication*. MIT Press, 1961.

[7] A.J. Bell and T.J. Sejnowski. An information maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[8] A.J. Bell and T.J. Sejnowski. The independent components of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

[9] G. Bi and M. Poo. Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience*, 18:10464–10472, 1998.

[10] G. Bi and M. Poo. Distributed synaptic modification in neural networks induced by patterned stimulation. *Nature*, 401:792–796, 1999.

[11] E.L. Bienenstock, L. Cooper, and P. Munro. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, 2:32–48, 1982.

[12] R.B. Blackman and J.W. Tukey. *The Measurement of Power Spectra, from the Point of View of Communications Engineering.* Dover, 1959.

[13] F. Blatter and R. Hahnloser. A sparseness hierarchy models song selectivity. *Poster at Society for Neuroscience Meeting, Washington D.C.*, 2008.

[14] R. Brette and W. Gerstner. Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *Journal of Neurophysiology*, 94:3637–3642, 2005.

[15] S. Chung, X. Lia, and S.B. Nelson. Short-term depression at thalamocortical synapses contributes to rapid adaptation of cortical sensory responses in vivo. *Neuron*, 34:437–446, 2002.

[16] S.V. David, N. Mesgarani, and S.A. Shamma. Estimating sparse spectro-temporal receptive fields with natural stimuli. *Network*, 18(3):191–212, 2007.

[17] P. Dayan and L. F. Abbott. *Theoretical neuroscience.* MIT Press, 2001.

[18] R.C. deCharms, D.T. Blake, and M.M. Merzenich. Optimizing sound features for cortical neurons. *Science*, 280:1439–1443, 1998.

[19] R. DeValois, E.W. Yund, and N. Hepler. The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research*, 22:531–544, 1982.

[20] A. J. Doupe. Song- and order-selective neurons in the songbird anterior forebrain and their emergence during vocal development. *Journal of Neuroscience*, 17:1147–1167, 1997.

[21] J.J. Eggermont, A.M. Aertsen, and P.I. Johannesma. Prediction of the responses of auditory neurons in the midbrain of the grass frog based on the spectro-temporal receptive field. *Hearing Research*, 10:191–202, 1983.

[22] D.J. Field. What is the goal of sensory coding? *Neural Computation*, 6:559–601, 1994.

[23] P. Foldiak. Sparse coding in the primate cortex. In *The Handbook of Brain Theory and Neural Networks, Second edition.* MIT Press, 2002.

[24] G. Greene, D. Barrett, K. Sen, and C. Houghton. Sparse coding of birdsong and receptive field structure in songbirds. *Network: Computation in Neural Systems*, 20:162–177, 2009.

[25] H.K. Hartline. The response of single optic nerve fibers of the vertebrate eye to illumination of the retina. *American Journal of Physiology*, 121:400–415, 1938.

[26] D.O. Hebb. *The organization of behavior: A neuropsychological approach.* John Wiley Sons, New York, 1949.

[27] P. Heil and H. Scheich. Functional organisation of the avian auditory cortex analogue. i. topographic representation of isointensity bandwidth. *Brain Research*, 539:110–12, 1991.

[28] A. Hodgkin and A. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, 117:500–544, 1952.

[29] J.J. Hopfield and A.V. Herz. Rapid local synchronization of action potentials: Toward computation with coupled integrate-and-fire neurons. *Proceedings of the National Academy of Sciences*, 92:6655–6662, 1995.

[30] C. Houghton. Studying spike trains using a van rossum metric with a synapses-like filter. *Journal of Computational Neuroscience*, 26(1):149–155, 2009.

[31] C. Houghton and K. Sen. A new multi-neuron spike-train metric. *Neural Computation*, 20(6):1495–1511, 2008.

[32] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195:215–243, 1968.

[33] J.P. Jones and L.A. Palmer. The two-dimensional spatial structure of simple cell receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58:1187–211, 1987.

[34] E. C. Kandel and J. H. Schwartz. *Principles of Neural Science.* Elsevier, New York, 1991.

[35] C. Kopple, G.A. Manley, and M. Konishi. Auditory processing in birds. *Current Opinion in Neurobiology*, 10:471–481, 2000.

[36] L. Lapicque. Recherches quantitatives sur l'excitation electrique des nerfs traitee comme une polarization. *J. Physiol. Pathol. Gen.*, 9:620–635, 1907.

[37] S.R. Lehky, T.J. Sejnowski, and R Desimone. Predicting responses of nonlinear neurons in monkey striate cortex to complex patterns. *Journal of Neuroscience*, 12:3568–3581, 1992.

[38] M.S. Lewicki. Efficient coding of natural sounds. *Nature Neuroscience*, 5(4):356–63, 2002.

[39] C. K. Machens, H. Schtze, A. Franz, O. Kolesnikova, M. B. Stemmler, B. Ronacher, and A.V. Herz. Single auditory neurons rapidly discriminate conspecific communication signals. *Nature Neuroscience*, 6(4):341–2, 2003.

[40] C.K. Machens, M.S. Wehr, and A.M. Zador. Linearity of cortical receptive fields measured with natural sounds. *Journal of Neuroscience*, 24:1089–1100, 2004.

[41] D. Margoliash. Acoustic parameters underlying the responses of song-specific neurons in the white-crowned sparrow. *Journal of Neuroscience*, 3:1039–1057, 1983.

[42] D. Margoliash. Preference for autogenous song by auditory neurons in a song system nucleus of the white-crowned sparrow. *Journal of Neuroscience*, 6:1643–1661, 1986.

[43] F. Michler, R. Eckhorn, and T. Wachtler. Using spatiotemporal correlations to learn topographic maps for invariant object recognition. *Journal of Neurophysiology*, 102:953–964, 2009.

[44] C. Morris and H. Lecar. Voltage oscillations in the barnacle giant muscle fiber. *Biophysical Journal*, 35:193–213, 1981.

[45] C.M. Muller and H.J. Leppelsack. Feature extraction and tonotopic organization in the avian auditory forebrain. *Experimental Brain Research*, 59, 1985.

[46] R. Narayan, V. Best, E. Ozmeral, E. McClaine, M. Dent, B. Shinn-Cunningham, and K. Sen. Cortical interference effects in the cocktail party problem. *Nature Neuroscience*, 10(12):1601–1607, 2007.

[47] B.A. Olshausen and D.J. Field. Energence of simple-cell receptive field properies by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[48] B.A. Olshausen and D.J. Field. Natural image statistics and efficient coding. *Network*, 7:333–339, 1996.

[49] B.A. Olshausen and D.J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311–3325, 1997.

[50] B.A. Olshausen and D.J. Field. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4):481–487, 2004.

[51] A. J. Parker and M. J. Hawken. Two-dimensional spatial structure of receptive fields in monkey striate cortex. *Journal of the Optical Society of America A*, 5:598–605, 1988.

[52] R. Penrose. A generalized inverse for matrices. *Proceedings of the Cambridge Philosophical Society*, 51:17–19, 1955.

[53] FitzHugh R. Mathematical models of threshold phenomena in the nerve membrane. *Bull. Math. Biophysics*, 17:257–278, 1955.

[54] F. Rieke, D. Warland, R.R. van Steveninck, and W. Bialek. *Spikes: exploring the neural code.* MIT Computational Neuroscience Series, 1999.

[55] C.J. Rozell, D.H. Johnson, R.G. Baraniuk, and B.A. Olshausen. Sparse coding via thresholding and local competition in neural circuits. *Neural Computation*, 20:2526–2563, 2008.

[56] N.C. Rust, O. Schwartz, J.A. Movshon, and E.P. Simoncelli. Spatiotemporal elements of macaque v1 receptive fields. *Neuron*, 46:945–956, 2005.

[57] E.L. Schwartz, R. Desimone, T.D. Albright, and C.G. Gross. Shape recognition and inferior temporal neurons. *Proceedings of the National Academy of Sciences: Biological Sciences*, 80:5776–5778, 1983.

[58] K. Sen, F. E. Theunissen, and Allison J. Doupe. Feature analysis of natural sounds in the songbird auditory forebrain. *Journal of Neurophysiology*, 86:1445–1458, 2001.

[59] C.C. Sherwood, M.A. Raghanti, C.D. Stimpson, C.J. Bonar, A.A. de Sousa, T.M. Preuss, and Patrick R. Hof. Scaling of inhibitory interneurons in areas v1 and v2 of anthropoid primates as revealed by calcium-binding protein immunohistochemistry. *Brain, Behaviour and Evolution*, 69:176–195, 2007.

[60] E.P. Simoncelli, J. Pillow, L. Paninski, and O. Schwartz. Characterization of neural response with stochastic stimuli. In *The Cognitive Neurosciences*, pages 327–338. MIT Press, Cambridge, MA,, 2004.

[61] E. C. Smith and M. S. Lewicki. Efficient auditory coding. *Nature*, 429:978–982, 2006.

[62] M. M. Solis and A. J. Doupe. Anterior forebrain neurons develop selectivity by an intermediate stage of birdsong learning. *Journal of Neuroscience*, 17:6447–6462, 1997.

[63] S. Song and L.F. Abbott. Cortical development and remapping through spike timing-dependent plasticity. *Neuron*, 32:339–350, 2001.

[64] S. Song, K.D. Miller, and L.F. Abbott. Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *Nature Neuroscience*, 3:919–926, 2000.

[65] F.E. Theunissen, S.V. David, N.C. Singh, A. Hsu, W.E. Vinje, and J.L. Gallant. Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems*, 12:289–316, 2001.

[66] F.E. Theunissen and A.J. Doupe. Temporal and spectral sensitivity of complex auditory neurons in the nucleus hvc of male zebra finches. *Journal of Neuroscience*, 18(10):3786–3802, 1998.

[67] F.E. Theunissen, K. Sen, and A. J. Doupe. Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *Journal of Neuroscience*, 20(6):2315–2331, 2000.

[68] A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math Dokl.*, 1963.

[69] M. van Rossum. A novel spike distance. *Neural Computation*, 13:751–763, 2001.

[70] J. D. Victor, F. Mechler, I. Ohiorhenuan, A. M. Schmid, and K. P. Purpura. Laminar and orientation-dependent characteristics of spatial nonlinearities: Implications for the computational architecture of visual cortex. *Journal of Neurophysiology*, 102:3414–3432, 2009.

[71] B.T. Vincent, R.J. Baddeley, T. Troscianko, and I.D. Gilchrist. Is the early visual system optmised to be energy efficient? *Network: Computation in Neural Systems*, 16:175–190, 2005.

[72] W.E. Vinje and J.L. Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287:1273–1276, 2000.

[73] T.P. Vogels and L.F. Abbott. Signal propagation and logic gating in networks of integrate-and-fire neurons. *Journal of Neuroscience*, 25:10786–10795, 2005.

[74] S.M.N. Woolley, T.E. Fremouw, A. Hsu, and F.E. Theunissen. Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nature Neuroscience*, 8:1371–1379, 2005.

[75] Ramon y Cajal. *Textura del sistema Nervioso del Hombre y los Vertebrados (1894-1904); Histology of the Nervous System of man and vertebrates (English translation N. Swanson & L.W. Swanson).* Oxford University Press, 1994.

[76] M.D. Zaretsky and M. Konishi. Tonotopic organization in the avian telencephalon. *Brain Research*, 111:167–71, 1976.