Associative memory properties of multiple cortical modules

ALFONSO RENART, NÉSTOR PARGA
Departamento de Física Teórica
Universidad Autónoma de Madrid
Canto Blanco, 28049 Madrid, Spain
and
EDMUND T. ROLLS
Oxford University
Department of Experimental Psychology
South Parks Road, Oxford OX1 3UD, England

Abstract

The existence of recurrent collateral connections between pyramidal cells within a cortical area and, in addition, reciprocal connections between connected cortical areas, is well established. In this work we analyze the properties of a tri-modular architecture of this type in which two input modules have convergent connections to a third module (which in the brain might be the next module in cortical processing or a bi-modal area receiving connections from two different processing pathways). Memory retrieval is analyzed in this system which has Hebb-like synaptic modifiability in the connections and attractor states. Local activity features are stored in the intra-modular connections while the associations between corresponding features in different modules present during training are stored in the inter-modular connections. The response of the network when tested with corresponding and contradictory stimuli to the two input pathways is studied in detail. The model is solved quantitatively using techniques of statistical physics. In one type of test, a sequence of stimuli was applied, with a delay between them. It is found that if the coupling between the modules is low a regime exists in which they retain the capability to retrieve any of their stored features independently of the features being retrieved by the the other modules. Although independent in this sense, the modules still influence each other in this regime through persistent modulatory currents which are strong enough to initiate recall in the whole network when only a single module is stimulated, and to raise the mean firing rates of the neurons in the attractors if the features in the different modules are corresponding. Some of these mechanisms might be useful for the description of many phenomena observed in single neuron activity recorded during short term memory tasks such as delayed match to sample. It is also shown that with contradictory stimulation of the two input modules the model accounts for many of the phenomena observed in the McGurk effect, in which contradictory auditory and visual inputs can lead to misperception.

[Key words: short term memory; delayed match to sample; McGurk effect; attractor networks; cortical computation; memory]
P.A.C.S. 87.30 Biophysics of neurophysiological processes
Preprint UAM-CXI 98-10
Submitted to NETWORK.

1 Introduction

The association areas of the human neocortex comprise most of it. In a processing stream, the information arrives first at the primary sensory areas, and is then passed on to a series of association cortical areas, often connected hierarchically, with forward and backward connections between the adjacent stages in the hierarchy (see Rolls & Treves (1998), for an introduction to this type of architecture, and some of its possible properties). The existence of these connections implies that these cortical areas cannot be studied in isolation, but they should rather be modelled as multi-modular neural networks. Although some steps in this direction have been made (Amit et al. 1990, O'Kane & Treves 1992, Lauro-Grotto et al. 1997), the study of these type of system is clearly underdeveloped, especially if one takes into account their importance.

As a first example of the type of neural structure we will be dealing with, one could consider two different sensory pathways which converge to a multi-modal sensory area: after a number of unimodal processing stages (in which increasingly complex and abstract features may be extracted), the streams from different modalities (e.g. taste and smell) may converge (e.g. to form flavor, that is a representation that depends on both odor and taste).

Another instance where multi-modular processing has been suggested to play a role is working memory (Fuster et al. 1985, Miller & Desimone 1994, Miller et al. 1996). This is the capacity to actively hold information in memory for subsequent use if needed. This phenomenon (especially visual working memory) is usually studied in delayed match-to-sample tasks, in which a visual image has to be kept in memory for subsequent comparison with different test stimuli on a trial. It has been observed that the activity in the delay period in the inferotemporal (IT) cortex provoked by the presentation of a visual stimulus is disrupted every time a new stimulus is presented (Miller et al. 1993). The fact that the sample has to be kept in memory for the whole duration of the trial while other stimuli are being presented suggests that more than one neural structure is needed for the succesful performance of the task. There is evidence that prefrontal (PF) cortex may play the long delay period memory in this task (Chelazzi, Miller, Lueschow & Desimone 1993, Miller et al. 1996).

Given that the prefrontal cortex (see e.g. (Fuster & Alexander 1971, Fuster et al. 1982, Wilson et al. 1993)) and the inferior temporal cortex (see e.g. (Fuster & Jervey 1981, Miyashita & Chang 1988, Fuster 1990, Amit 1995)) can show delay period related activity which is stimulus-specific, it seems natural to model these areas as recurrent attractor networks and to try to explore if some of the mechanisms underlying the numerous experimental observations indicating that multiple neural structures seem to interact can be captured by a simple model of interacting attractor recurrent networks.

The aim of this paper is therefore, to understand some of the properties of multi-modular associative recurrent networks and to investigate possible functions of the reciprocal connections between separate areas of the cortex.

A possible way in which the architecture we will be concerned with can be described is in terms of two sensory pathways that converge to a third module (see figure 1). Each module contains recurrent collateral connections which allow each module to operate as an attractor network after training. Modules **A** and **B** are not connected to each other, but they both converge to the third module **C**. Because of the frequent conjunction of the particular sensory data arriving as inputs to each of the modules **A** and **B**, which are therefore denoted as input modules, **C** is a convergent module which receives paired inputs.

Its patterns of sustained activity (due to its attractor network properties) are associated by Hebb-like modifiable synapses to the corresponding sustained activity states present in the input modules during processing. In general, one expects these inter-modular associations to involve sets of several features in one module associated with sets of several features in the connected module (Renart et al. 1999a). For instance, in the case mentioned above where modules **A**, **B** and **C** represent cortical areas providing representations of smell, taste and flavor respectively, the need for inter-modular associations between sets of features in each module is clear. (Note that flavor is normally produced by a combination of taste and odor inputs.) Since stimuli with a given taste may have a number of different odors, this particular taste will be associated with all of these odors, giving different flavors for each association (e.g., both strawberries and raspberries are sweet). However, for the sake of simplicity, we will restrict the discussion of the tri-modular architecture to the case where each feature in a given module is associated with a single feature in each of its neighboring modules (even in this very simple situation, we will see that the phenomenology of the multi-modular network is very rich). This will allow us to differentiate the connections between neurons in different modules from the connections between neurons in the same module, by means of a single continuous parameter (to be denoted g) measuring the strength of the inter-modular connections relative to the strength of the intra-modular connections.

The modules are composed of a very large number of graded-response neurons with hyperbolic transduction functions. The form of the intra- and inter-modular connections reflects the result of Hebbian learning of sparsely coded binary patterns by the multi-modular network, and is essentially a generalization to multi-modular architectures of the synaptic matrix for a single module proposed in (Tsodyks & Feigelman 1988, Buhmann et al. 1989).

A central problem regarding the operation of an associative multi-modular network like the one just described is whether, due to the associations between the features stored in each module, the multi-modular network behaves as a single network, in which the features act just as parts of a global activity pattern or, on the contrary, each module can behave independently. What we mean by this is basically that different modules can be simultaneously in attractors corresponding to features not associated together in the inter-modular synaptic connections. To find a general answer to this question is not easy, since the behavior of the network depends simultaneously on many different factors such as the free parameters of the model, especially the relative strength of the inter- to intra-modular connections g, and the type of stimulation procedure, e.g. clamped versus transient stimuli, consistent versus contradictory stimulation of the two input modules, etc... However, if one restricts one's attention to the stable delay activity states after consistent stimulation, and focuses only on the effect of g for fixed values of the remainder of the parameters, one can hypothesize a few possible regimes. First, in the limiting case of a very low value of g, one expects the modules to be effectively disconnected, in the sense that activity will not spread to the internal module **C** from the input ones and therefore the modules will be unable to influence each other. In a second and opposite regime in which the relative strength g is close to one, it is natural to think that the small difference between the magnitude of the inter- versus intra-modular connections will not be enough to give the modules an *identity* of their own. They will probably behave just as parts of a global unimodular network comprising the three modules, in which the attractors will necessarily involve the triplets of associated features stored in each module. This would produce global attractor states. A third set of more complex regimes may exist for intermediate values of g. For example, in one the networks may interact, but not so much that they become just fractions of a

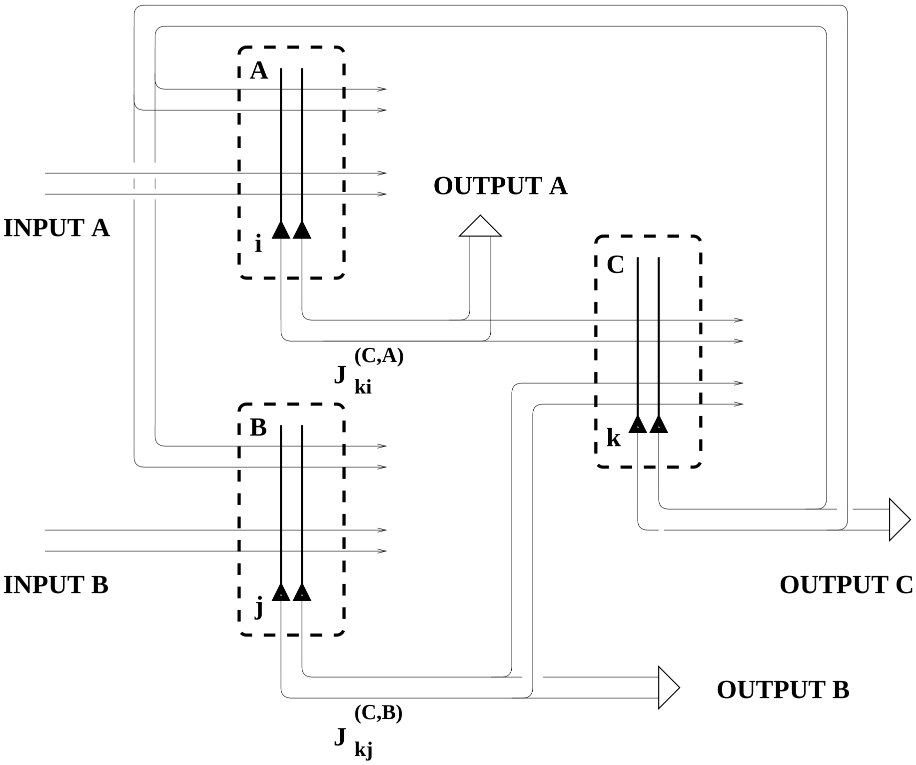

Figure 1: An architecture with one bi-modal module. Triangles represent cell bodies, and thick lines their dendritic trees. $J_{ki}^{(\mathbf{C},\mathbf{A})}$ denotes the connection between the pre-synaptic unit i (in module $\mathbf{A}$) and the post-synaptic unit k (in module $\mathbf{C}$). Similarly $J_{kj}^{(\mathbf{C},\mathbf{B})}$ is the synaptic connection between neurons j and k located in modules $\mathbf{B}$ and $\mathbf{C}$ respectively. The relative strength of the inter-modular connections is measured by an external parameter g ($0 \leq g \leq 1$). Since the synaptic matrix is assumed to be symmetric, the feedforward and back-projecting connections are defined equal. The recurrent connections inside each module are not shown in this figure.

larger, single global network. It might be that the connections between neurons in different modules were large enough so that activity could initially propagate to set up global, consistent attractors, but were also small enough so that if a new cue was presented to the input modules, module **C** would be capable of remaining in an attractor corresponding to the previous stimuli. (Although it might be in a slightly different state from the previous one due to the reciprocally inconsistent signal between the modules.) Such a network might be useful in describing some neural mechanisms related to working memory.

The case of stimulation of the input modules with cues close to features which are not associated in the connections (inconsistent or contradictory stimulation) also offers the possibility to explore the nature of the interactions between the modules. In fact, the interaction between different sensory pathways (for example auditory and visual) has been studied experimentally (McGurk & MacDonald 1976, Howells 1944). In this system the sight of the mouth movements made during vocalization and speech is correlated with the phonemes heard. The associations between the auditory and visual inputs are learned (in a part of the higher order multi-modal processing cortical areas) in such a way that the sight of the mouth moving can influence which phoneme is actually heard (McGurk

& MacDonald 1976). This effect will be studied in detail in Section 5. Consistent with the ideas presented here, activation of auditory cortical areas (e.g. **A** in figure 1) has been observed when the sight of a mouth vocalizing is seen by a subject (producing visual activation in for example module **B** of figure 1) (Calvert et al. 1997). In this case, the influence could be produced by forward connections from module **B** to module **C**, and then backward connections from module **C** to module **A** (see Rolls & Treves (1998), section 4.5). Consistent with the hypothesis that there is such a multi-modal convergence area (**C** in figure 1), single neurons in the primate cortex in the superior temporal sulcus are activated by the sight of the lips moving, and neuronal responses to auditory stimuli have been found in the same region (Baylis et al. 1987, Hasselmo et al. 1989).

The architecture in figure 1 may also be relevant to several different situations apart from the ones already mentioned, and the model should be considered in a more abstract way. For example, if only one of the modules is used for input, in principle the model can describe networks arranged as a (sequential) set of layers with internal recurrent connections and forward- and back-projections between the stages. It could represent, for instance, some of the stages in the visual system (V1, V4, IT) which interact in important ways. A tri-modular architecture could also be interpreted as a model for the hippocampus-neocortex system; in this case the central module **C** is taken as the hippocampus while the other two are interpreted as two different cortical areas (Treves & Rolls 1994, Rolls & Treves 1998). In all these situations, one expects Hebbian associations between patterns of sustained activity to take place. They would be set up during learning as a result of, for example, regular associations between sensory stimuli in different sensory modalities, more generally because of regular associations between inputs to a network, or in the case of back-projections, because of regularly associated activity in one stage of processing with that in a higher stage of processing. The ways in which this learning stage of the process could occur have been discussed by Rolls (1989) and Rolls & Treves (1998).

We begin by describing the model and then proceed to present its solution in terms of a series of self-consistency equations for the quantities that describe the sustained activity states of the network (Section 3). Results are presented first, in Section 4, on the question of the the dependence of the activity of the modules on the parameter g, and then, in Section 5, the model is used to investigate the experimental results on the simultaneous processing of contradictory information found by McGurk & MacDonald (1976). A discussion of the results is presented in the last section.

2 The multi-modular network

The model we use to describe a multi-modular network is an extension of the bi-modular model proposed in (Renart et al. 1999a). It employs the formalism first studied in (Treves 1990, Shiino & Fukay 1990, Kühn 1990) in which neurons are described as analog variables. This idea was then developed by Amit & Tsodyks (1991a), who showed that it is possible to approximate the dynamics of a network of integrate and fire (IF) units by a dynamics in terms of currents and mean firing rates in certain conditions which are plausible in the cortex. Although a better description of associative recurrent networks which, e.g. allows the inclusion of attractors corresponding to spontaneous activity states and of realistic values for the mean firing rates of the neurons, is now available (Amit & Brunel 1997), it is technically more complicated and we do not expect the main qualitative conclusions of this study to depend on the details of the formalism used, so we have restricted the model to the simpler description in terms of currents and mean rates. The main elements of the model are as follows:

The neurons: Neurons are described as dynamical elements which emit spikes at a given, definite, firing rate. The firing rate of a neuron at a given time is a function of the incoming current it is receiving at that moment. The function which transforms currents into rates will be denoted the transduction function of the neurons and the explicit form we have chosen will be given below.

The network dynamics are those of the input currents received by the neurons, defined according to the set of equations:

$$\mathcal{T}\frac{dI_{ai}(t)}{dt} = -I_{ai}(t) + \sum_{bj} J_{ij}^{(a,b)}\nu_{bj} + h_{ai}^{(ext)} \quad . \tag{1}$$

Here, I_{ai} is the afferent current into the neuron i of the module a, and ν_{bj} is the firing rate of the neuron j of the module b. The current is driven by the output spike rates of the other neurons in the network (located either in the same or in different modules), weighted with the corresponding synaptic efficacies $J_{ij}^{(a,b)}$, and (in the case of the input modules) by the stimulus (or external field) $h_{ai}^{(ext)}$. At the same time, it decays with a characteristic time constant $\mathcal{T}$. The conversion from currents to rates, necessary to complete the definition of the dynamics, will be indicated by $\nu = \phi(I)$, where $\phi(x)$ is the transduction function. We have chosen a hyperbolic transduction function given by:

$$\phi(I) = \begin{cases} 0 & \text{if } I < \theta \\ \tanh[G\ (I-\theta)] & \text{if } I \geq \theta \end{cases} \quad , \tag{2}$$

where G is the gain and θ is the threshold. The main reason for our choice of this function is that it is at the same time simple and supports stable attractors with non-saturating firing rates without any specific modelling of the inhibition in the network. On the other hand we expect no qualitative changes to occur if a different transduction function is used. Notice that in this description firing rates are normalized. The value $\nu = 1$ corresponds to the maximum firing rate achievable by the neurons in spikes/sec.

The stored patterns: In each module a, the number of stored patterns is denoted by P. The patterns are defined in terms of binary variables η_{ai}^{μ} ($\mu = 1, \ldots, P$; $i = 1, \ldots, N$). The η's are independent random variables which are chosen equal to one with probability f (the mean coding level of the stimuli) and equal to zero with probability $(1-f)$. Their variance is therefore $\chi \equiv f(1-f)$.

The synaptic connections: The synaptic matrix will be denoted by $J_{ij}^{(a,b)}$, where again a and b are module indices and i and j are neurons in a and b respectively. The only constraint that we will impose on this matrix is symmetry under the interchange of the neuron indices. This will allow us to solve the model analytically. The intra- and inter-modular connections are given (respectively) by:

$$J_{ij}^{(a,a)} = \frac{J_0}{\chi N_t}\sum_{\mu=1}^{P}(\eta_{ai}^{\mu} - f)\ (\eta_{aj}^{\mu} - f) \qquad i \neq j \ ; \ a = \mathbf{A}, \mathbf{B}, \mathbf{C} \tag{3}$$

$$J_{ij}^{(a,b)} = \frac{g}{\chi N_t}\sum_{\mu=1}^{P}(\eta_{ai}^{\mu} - f)\ (\eta_{bj}^{\mu} - f) \qquad \forall\ i, j \quad , \tag{4}$$

and $J_{ii}^{(a,a)} = 0$. For the tri-modular architecture of figure 1, (a, b) in eq. (4) can take the values $(\mathbf{A}, \mathbf{C})$ and $(\mathbf{B}, \mathbf{C})$ or, equivalently, any other combination is assumed to have a connection strength of zero. The symmetry condition has to be imposed by setting $J_{ij}^{(\mathbf{C},\mathbf{A})} = J_{ji}^{(\mathbf{A},\mathbf{C})}$ and $J_{ij}^{(\mathbf{C},\mathbf{B})} = J_{ji}^{(\mathbf{B},\mathbf{C})}$. The parameters g and J_0 measure the strength of the inter- and intrar-modular connections respectively. One may take into account the fact that if the number of stored features is finite and the network is very large, as will be the case here, randomly diluting the connectivity of the network is equivalent to a renormalization of the strength of the connections (see Appendix). Therefore the parameters J_0 and g can be thought as effective connection strengths taking into account not only the strength of the connections, but also the fraction of them which have been set equal to zero. (Randomly disconnecting half of the synapses of the network is equivalent to dividing the intensity of the connections of the fully connected network by two).

Having noticed this, the weight normalization $N_t \equiv N\Lambda$ with $\Lambda = (J_0 + 2g)$ can be thought of as the average number of synapses to a neuron in the convergent module taking into account their intensity. The inclusion of the parameter g in the normalization constant ensures that the total effective intensity of the connections afferent to any neuron in the convergent module remains fixed as the strength of the associations varies [1]. It allows us to investigate the changes that take place as the fraction of current received from the recurrent collaterals inside a given module is varied and replaced by current from afferents from outside the module. The relative importance of each contribution can therefore be measured.

The external field: The external fields used as stimuli will be chosen as proportional to one of the stored features (e.g. feature μ_0). Their strength is controlled by an intensity parameter h, in terms of which the stimulus at a given neuron i in module a can be expressed as:

$$h_{ai}^{(ext)} = h\, \eta_{ai}^{\mu_0} \quad . \tag{5}$$

Finally, the total number stored features per module, P, is a finite number, chosen not to increase with the size of the network.

3 Solution of the model

If the number of stored features is finite, the solution for the fixed points of equations 1 can be found very simply.

The relevant quantities, in terms of which a macroscopic description of the fixed points of the network can be produced, are the overlaps between the state of each module (characterized by the firing rate of its neurons) and each of its stored features averaged over all possible realizations of these features. The reason for performing this average is that, since the features are made up of random variables, one is interested in the average retrieval properties of the network over all possible realizations of its memories, rather than on any specific one in particular. The overlap with an arbitrary feature μ of module a is therefore defined as:

$$m_a^\mu = \frac{1}{\chi N} \ll \sum_i (\eta_{ai}^\mu - f)\nu_{ai} \gg_\eta \quad , \tag{6}$$

[1] Since the number and strength of the connections to a given neuron on the input modules is different, this will not be the case for neurons in these two modules.

where the symbol $\ll \ldots \gg_\eta$ means average over the random variables appearing in the features. This expression is normalized so that the overlap between the feature η^μ_{ai} and a state $\{\nu_{ai}\}$ in which all neurons active in that feature have a (normalized) firing rate of one, be equal to one. At the same time it ensures that the overlap between the feature and a state independent from it is zero.

Next we write the local field afferent to a given neuron i of module a, which is defined as

$$h_{ai} = \sum_j J^{(a,a)}_{ij} \nu_{aj} \; + \sum_{b \neq a} \sum_j J^{(a,b)}_{ij} \nu_{bj} \quad , \tag{7}$$

and notice that it can be expressed in terms of the overlaps defined above

$$h_{ai} = \frac{J_0}{\Lambda} \sum_\mu (\eta^\mu_a - f) m^\mu_a \; + \; \frac{g}{\Lambda} \sum_{b \neq a} \sum_\mu (\eta^\mu_a - f) m^\mu_a \quad . \tag{8}$$

The contributions from outside and inside the module can be put together using a matrix K which contains all the information about the architecture of the network

$$h_{ai} = \sum_\mu (\eta^\mu_a - f) \left(\sum_b K^{ab} m^\mu_b \right) \quad , \tag{9}$$

with

$$K = \begin{pmatrix} \frac{J_0 d_0}{\Lambda} & 0 & \frac{gd}{\Lambda} \\ 0 & \frac{J_0 d_0}{\Lambda} & \frac{gd}{\Lambda} \\ \frac{gd}{\Lambda} & \frac{gd}{\Lambda} & \frac{J_0 d_0}{\Lambda} \end{pmatrix} \quad . \tag{10}$$

In the presence of an external stimulus $h^{(ext)}_{ai}$, the local field is the sum of the contribution from inside the network (9) and this external field.

Finally, one imposes the condition that the state of the system be a fixed point of the dynamics by setting the left hand side of equation (1) to zero, which equates the total effective afferent current to our neuron with its local field. Transforming the current into a mean firing rate through the use of the transduction function ϕ gives the firing rate of each neuron in the network as a function of the overlaps m^μ_a. The self-consistency equation for the overlaps is obtained by replacing this expression for the rates in equation (6)

$$m^\mu_a = \frac{1}{\chi} \ll (\eta^\mu_a - f) \tilde{\nu_a} \gg_\eta \quad , \tag{11}$$

where the quantity $\tilde{\nu_a}$ is defined as:

$$\tilde{\nu_a}(\eta) = \phi \left\{ \sum_\mu (\eta^\mu_a - f) \left(\sum_b K^{ab} m^\mu_b \right) \right\} \quad , \tag{12}$$

and represents the rate of the neurons of module a in the attractor. It is a function of the variables η^μ_a since the rate a neuron of module a will depend, in principle, on whether it is active or inactive in each of the stored features of this module. Since the number of patterns does not increase with the size of the network, one can use the property of *self-averaging* (see e.g. (Hertz et al. 1991)) to replace the sum over the size of the network in (6) by the average over the random variables which make up the features that appears in (11).

As noted by Amit & Tsodyks (1991b), the overlap with a given feature μ of module a can be expressed as the difference between the average activity of neurons active (the *foreground population*) and silent (the *foreground population*) in that feature, in the following way:

$$m_a^\mu \equiv \nu_{a+}^\mu - \nu_{a0}^\mu \quad , \tag{13}$$

where

$$\nu_{a+}^\mu = \frac{1}{fN} \ll \sum_i \eta_{ai}^\mu \nu_{ai} \gg_\eta \tag{14}$$

$$\nu_{a0}^\mu = \frac{1}{(1-f)N} \ll \sum_i (1-\eta_{ai}^\mu)\nu_{ai} \gg_\eta \quad . \tag{15}$$

Notice that if the module is in an attractor such that all neurons active in a given feature are firing at a given rate ν', the overlap of the module with that feature in the attractor will be ν'. In general, a non-zero value for the overlap inplies that the average activity of the two populations in the attractor is different, and therefore that the sustained activity state of the module is correlated with the corresponding feature. The numerical values of the overlaps of the states of the three modules with each of their stored features characterize macroscopically the attractors of the multi-modular system.

We have solved these equations by an iterative procedure in which all the overlaps are set initially to zero, and then the input modules are stimulated with an effective current equal to the external field (5) during a certain number of iterations. The external stimuli could correspond to any pair of features in each of the input modules, thus allowing the study of retrieval of corresponding or contradictory features (those from different modules not associated with each other). In the case of persistent (i.e. clamped) stimuli, the external fields were present during all iterations until a fixed point was reached in which the values of the the parameters and of the firing rate function reproduced themselves when further iterated.

4 Regimes of the multi-modular network

In this section, the problem of the autonomy of the modules as a function of the relative contribution of the extra-modular afferents is studied. In order to do this and to simplify the analysis, we have chosen to fix some of the external parameters of the system. First, since the sum of the intensities of the connections afferent to the central module is chosen to be constant, changing the value of the parameter J_0 is equivalent to a renormalization of g. We have therefore chosen to fix $J_0 = 1$ everywhere, so that g becomes the true relative value (the ratio) between the inter- and intra-modular connection strengths:

$$J_{0_{eff}} = \frac{J_0}{J_0 + 2g} \equiv \frac{1}{1+2g} \quad ; \quad g_{eff} = \frac{g}{J_0 + 2g} \equiv \frac{g}{1+2g} \quad , \tag{16}$$

then,

$$g = \frac{g_{eff}}{J_{0_{eff}}} \quad ; \quad J_{0_{eff}} + 2g_{eff} = 1 \quad , \tag{17}$$

where $J_{0_{eff}}$ and g_{eff} are the strengths of the intra- and inter-modular connections respectively with the normalization constant Λ already taken into account. The coding rate (or sparseness) of the stored features has been set equal to $f = 0.2$, which is similar in magnitude to what is found by integrating the tail of typical spike rate distributions in inferotemporal cortex (Rolls & Tovée 1995, Treves et al. 1999). The values of the gain G and of the threshold θ characterizing the transduction function, have also been fixed. Since these quantities *effectively* reflect features of the real neurons, their values cannot be chosen according to neurophysiological plausibility. They are rather usually chosen according to a desired plausible behavior of the model network being studied. Using the same argument, we have chosen them so that, first, the network showed a wide range of possible behaviors as g was varied, and second, that the firing rates of the neurons in the attractors were far from saturation. According to these criteria, in the next two sections $G = 1.3$ and $\theta = 0.001$. Also, unless stated otherwise, the external stimuli are transient and, if both input modules are stimulated, consistent.

We want to point out that our aim has not been a complete description of the possible behaviors of the system in every point of the (quite large) parameter space. We have rather been interested in demonstrating that some specific interesting regimes were present and that they could be realized without drastically changing the parameters of the model. In particular, we have only varied the relative inter-modular connection strength g and the type of stimulation procedure, fixing everything else as stated above.

We have therefore solved the equations (11) for the values of the parameters stated above and for unclamped conditions as a function of g. A summary of the results is given in Table 1, where the ranges of g corresponding to the different qualitative behaviors (also denoted as *phases*) of the network have been listed.

When g is very close to zero, the activity does not spread through the modules, and only the input modules which have been stimulated show activity in the delay (consistent with the cue). Neurons in non-stimulated modules are in the quiescent state in this situation. Since the modules are effectively not interacting with each other and therefore seem to be disconnected in this regime, it will be referred to as the *Isolated* phase.

For slightly larger g the activity propagates to the central module and from it, if only one module is stimulated, to the non-stimulated input module as well [2]. This happens, for instance at $g = 0.005$ if a single input module is stimulated. In this region, the 'delay period' activity states are global retrieval states in which the three modules are in attractors correlated with the triplet of features selected by the cue.

To test if the modules were capable of some degree of independent processing, even when the values of g were such that their interaction caused global retrieval delay activity states, we studied the behavior of the network when a sequence of stimuli separated by a delay was used (cf. the neurophysiological investigations of (Baylis & Rolls 1987, Miller et al. 1993, Miller & Desimone 1994)). For every fixed value of g, this was done in the following way: First, with the network silent, one input module was stimulated with a cue equal to one of their stored features. Let us denote this stimulus as $\mathbf{a}$. Then, the stimulus was removed, and the network was let to evolve (according to the iterative dynamics described above) to a self-sustained attractor. A new stimulus equal to a second feature $\mathbf{b}$ was then imposed for a number of iterations on top of the delay activity left by the previous one and, after it had been removed, the new attractor was examined.

Of course, the difference between the first time the network is stimulated and the second is that, in the latter case, there are conflicting components in the effective current driving the neurons: on one side is the current coming from the recurrent collaterals and from

[2] Although in a marginally small range of g, a stable state also exists in which, when only one module is stimulated, only the central and the stimulated modules are active in the delay.

the projections from the other modules which initially tend to stabilize the already existing attractor and, on the other, is the current provoked by the external stimulus, which tends to activate neurons which represent a different feature (except for a fraction f of the neurons active in **a**, which are also active in **b**). It is the interaction between these two components that may result in a different delay activity state after the second stimulus.

The result is that, for small g, there exists a range of values of this parameter in which the final attractors of the whole network are such that the central module and the non-stimulated input module remain in a state correlated with the feature selected by the *first* stimuli **a**, while the stimulated input module moves to a new state correlated with the feature selected by **b**. Although the state of the network does not correspond to a triplet of corresponding features anymore, two of the modules do remain in associated features. To test if this behavior was robust, this time, with the network initially in its current state, the other input module was stimulated with a cue equal to a third feature **c** different from the other two. As expected, only the stimulated input module changed to a new attractor correlated with feature **c**. The other two remained in attractors correlated with the same features as before, although with slightly lower rates, due the fact that the signals that the modules are now communicating to each other are inconsistent. Thus, In this final state, the input modules are in states correlated with features **b** and **c** while the central module stays in a state correlated with feature **a** the whole time. Since the modules appear to be able to retrieve independently, we have denoted the region of the space of parameters where this behavior occurs the *independent* (I) phase.

Returning to the case of stimulation of a single input module, the sequence can be as long as one wants, and no noticeable difference occurs if, for instance, a stimulus is repeated in the middle of the sequence. If the first stimulus is presented again, the whole network moves to the initial configuration in which the attractors in three modules are correlated with the three associated features selected by the first stimulus **a**. The difference between this state of consistent activity and the rest is that here the values of the mean firing rates of the neurons in the attractors are slightly higher, due to the reciprocal cooperative influence that the modules exert on each other. This difference in firing rates might be used as a means to detect the appearance of a previously seen stimulus.

The fact that this type of behavior exists is one of the most important results of this paper. It implies that it is possible for a set of connected recurrent networks, at least for some choices of the model parameters, to interact in such a way as to exploit the associative capabilities of the uni-modular attractor networks and at the same time to retain some degree of independence between the individual components (the modules), which allows them to reverberate in a stable attractor under the persistent noisy influence of the attractors in the other modules. Indeed, the fact that each module is in an attractor implies that they can provide each other with a persistent external modulatory input which might be very helpful for things such as making comparisons between stimuli processed sequentially (as in delayed match-to-sample tasks, see e.g. (Miller & Desimone 1994)) or directing attention in visual search experiments (Chelazzi, Miller, Duncan & Desimone 1993, Chelazzi et al. 1998).

It is natural to expect that the results obtained using this procedure will depend on properties of the external stimuli such as their intensity or duration. Yet, it would be desirable that the stimulus necessary for this effect to take place were not very large, so that a typical stimulus was sufficient to produce it. The definition of a typical stimulus is complicated, so we have assessed the effectiveness of the stimuli comparing their properties with similar properties of the recurrent internal current present in a typical attractor. The

g	**Phase**
$g < 0.005$	Isolated
$0.005 \leq g < 0.012$	Independent
$0.012 \leq g < 0.043$	Locked
$0.043 \leq g$	Null

Table 1: The different regimes of the multi-modular network as a function of g when only one of the input modules is transiently stimulated. The rest of the model parameters are given at the beginning of section 4. The definitions of the different phases are also given in the text.

results show that this behavior is, at least for the choice of external parameters used, very robust against changes in the stimulus intensity and duration. A lower bound exists on the intensity and persistence of the stimulus, but it is reasonably low. From that point, the effect persists if the stimulus is made arbitrarily longer and stronger. To prove this we have checked that the final attractors had the desired property for a stimulus strong enough so that the rate of the neurons of the stimulated input modules reached saturation, and long enough so that the input modules converged to a stationary state in the presence of the stimulus. Any further increase in strength or duration does not affect the delay activities. As for the lower bound, the minimal strength is much less (by a factor 5) than the current coming from the recurrent collaterals in the delay activity states, and the duration is less than an order of magnitude smaller than the time that it takes the network to converge to an attractor, so it can reasonably be considered a quite brief and weak stimulus.

When g grows beyond 0.012, the picture changes and the independence between the modules is lost. The delay activity states found in this region, *always* involve the three modules in attractors correlated with consistent features associated in the synaptic connections. Also, since g is now larger, changes in the properties of the external stimuli have more impact on the delay activity states. The general trend seen in this phase under the change of stimulus after a previous consistent attractor has been reached is that, first, if the second stimulus **b** is not effective enough (it is weak or brief), it is unable to move any of the modules from their current delay activity states. If the stimulus is made more effective, then as soon as it is able to change the state of the stimulated input module, the internal and non-stimulated input modules follow, and the whole network moves into the new consistent attractor selected by the second stimulus. In this case, the interaction between the modules is so large that it does not allow contradictory local delay activity states to coexist. Therefore the triplets of features stored in each module are not only associated, but appear to be locked in a single global pattern. We have therefore denoted this region as *locked* (L) phase. As g grows larger, less effective stimuli are needed to move the three modules into a new attractor, since the relative value of the recurrent collaterals that tend to stabilize the previous attractor decreases with g. The L phase seems therefore more limited than the I phase, in the sense that it is functionally closer to a uni-modular network than to a multi-modular one.

Finally, if g becomes larger than ~ 0.04 the recurrent collaterals inside the modules become too weak and the network looses the capacity to sustain any activity. This is denoted as the *Null* phase, in which the only stable attractor is the global quiescent state [3].

[3]This particular result is, of course, an artifact of the normalization criterion used. We have checked that

5 The McGurk Effect

The strong interactions between different sensory pathways in the simultaneous processing of information have been demonstrated by McGurk & MacDonald (1976). They considered the effect of the simultaneous processing of contradictory information by the visual and auditory systems. In one of the experimental paradigms the internal representations of the two stimuli in the two input modules are not correlated either to each other or to the representations of any other single stimulus. More precisely, the subject receives one stimulus through the auditory pathway (e.g. the syllables *ga-ga*) and a *different* stimulus through the visual pathway (e.g. the lips of a person performing the movements corresponding to the syllables *ba-ba* on a TV screen). These stimuli are such that their acoustic waveforms as well as the lip motions needed to pronounce them are rather different. One can then assume that although they share the same vowel “a”, the internal representation of the syllables is dominated by the consonant, so that the representations of the syllables *ga-ga* and *ba-ba* are not correlated either in the primary visual cortical areas or in the primary auditory ones. At the end of the experiment the subject is asked to repeat what he heard. When this procedure is repeated with many subjects, it is found that roughly 50% of them claim to have heard either the auditory stimulus (*ga-ga*), which we will call auditory response, or the visual one (*ba-ba*), to be denoted as visual response. The rest of the subjects report to have heard neither the auditory nor the visual stimuli, but actually a combination of the two (e.g. *gabga*) or even something else including phonemes not presented auditorally or visually (e.g. *gagla*).

Given these results, one may wonder if a schematic model like the one under consideration here is able to, at least qualitatively, reproduce some of the features of the experiment. The implementation of the experimental procedure in our model requires the simultaneous processing of non-corresponding features by the two input modules. In this case, contradictory signals would be sent to the convergent module from the other two, each one acting as noise for the other. The question now arises as to which of these two states of sustained activity will be preferred by the bi-modal module and, on the bottom line, as to how this new type of ‘noise’ is going to affect the performance of the network. The result of the experiment just described suggests that, at least in some domain of the space of parameters of the model, the convergent module should converge to one of these two conflicting states.

In describing this experience, we have decided to examine the behavior of the system under persistent stimulation. This is because in the experiment, the total duration of the repetition was about $1s$ so we have assumed that during this time, an earlier area in the processing stream has arrived at an attractor representing the phonemes or the lip motions, and that it is the output of this area which is considered the input to the modules.

We have therefore investigated the behavior of the system under persistent stimulation of the two input modules with a pair of non-corresponding features (μ_A of module **A** and μ_B of module **B**) looking at the sustained activity states as a function of g for the same model parameters given at the beginning of Section 4. Figures 2a-b show the average firing rate of the neurons of the bi-modal module which are active in the two features associated with each of the non-corresponding patterns (a), and the average firing rate of the neurons of modules **A** and **B** which are active in the features μ_A and μ_B respectively (b). The stimulation was performed in a symmetric way (i.e. the intensity of both stimuli was the

the qualitative picture remains if no g-dependent normalization is used. In that case the only difference is that once the locked phase appears it extends all the way up to $g = 1$. However in this case g no longer has the meaning of the relative inter-modular connection strength so we have preferred to use the normalization described in the text.

same and equal to $h = 0.1$).

For g greater than ~ 0.1, module **C** is in a *symmetrical* situation, in a state with the same overlap with the two features associated with the stimuli on **A** and **B**. The mean firing rates of the populations of neurons active in any of those features are therefore equal, and the neurons which are active are those active in either of the features or in both. At the same time, the two input modules are in retrieval states, showing correlation with their stored features selected by the stimuli.

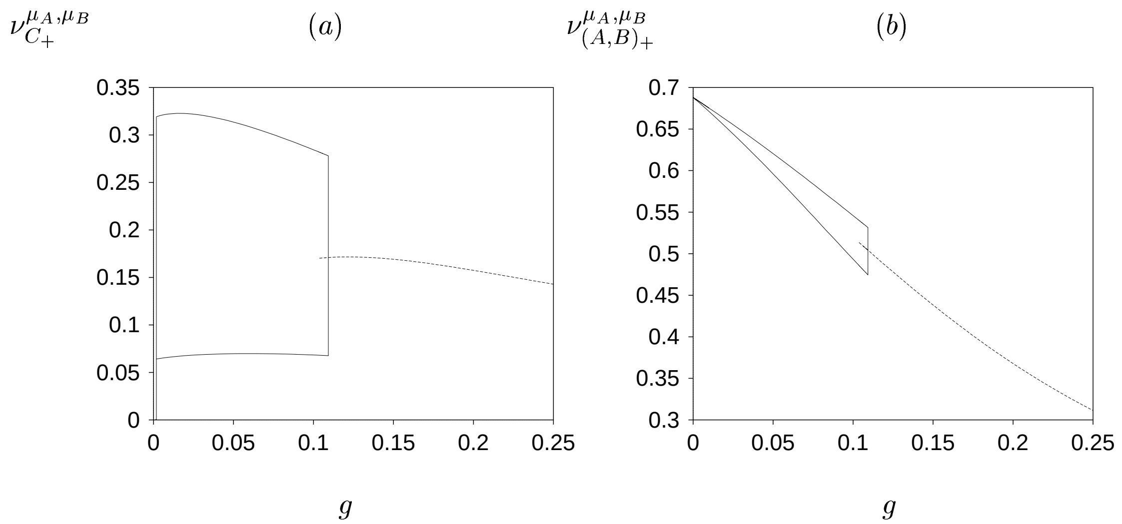

Figure 2: Symmetry breaking and symmetry restoration as a function of g. The model parameters are as described in the text. Both input modules were persistently stimulated with contradictory features of intensity $h = 0.1$. In **(a)** the mean firing rate of the neurons of module **C** active in each of the two features associated with the stimuli are shown: The solid line corresponds to the asymmetric phase in which these mean firing rates are different. In this phase the probability of each feature being 'selected' by the central module is the same. The dashed line corresponds to the symmetric phase in which the mean firing rate in the two neural populations is the same. **(b)** shows the mean firing rate in the neural population of each input module composed by neurons active in the feature being used as stimulus to that module. In the asymmetrical phase, the neurons of one of the input modules which are active in the feature associated with the one 'selected' by module **C** have a larger average firing rate due to back-projected signal from this module. There is a small transition region in which both symmetric or asymmetric states can be realized by the network.

As g decreases the rates in the input modules grow progressively until the symmetric phase is no longer stable. What is observed is that module **C** 'chooses' to stay in one of the two conflicting retrieval states: the symmetry of the system under the interchange of modules **A** and **B** is broken. There is however a small transition region between $g \sim 0.10$ and $g \sim 0.11$ in which both situations coexist, and the convergent module can either remain in the symmetric state or choose between one of the features, with either situation being stable. For lower g only the asymmetrical situation is realized. It is the random fluctuations produced during the convergence to the attractor that determine the pattern selected by module **C**, both in this small intermediate region and in the fully asymmetrical one. When the bi-modal module becomes correlated with *one* of its stored patterns, the signal back-projected to the input module stimulated with the feature associated with that pattern becomes stronger and the overlap in this module is increased. This effect can be seen in figure 2b, as well as an overall increase in the average firing rates in the retrieval states of the input modules as g is further reduced.

This behavior continues until g is extremely small (0.001), where the intermodular connection strength becomes so low that the convergent module does not get activated by the stimuli, the input modules becoming effectively disconnected.

The existence of this symmetry breaking regime could explain the effect observed by McGurk & MacDonald (1976). The two alternatives for the convergent module in our model would correspond to the visual and auditory responses in the experiment and the symmetric state with similar contributions from the visual and auditory components appearing for large g would correspond to combination-like responses. In reality, however, the final state is probably not selected by random fluctuations in the dynamics of the neural networks involved, but by asymmetries in the values of e.g. the stimulus intensity, the distortion of the stimuli or the strength of the inter-modular connections in the two sensory pathways. In principle a small degree of asymmetry could be enough in order to produce such an explicit breaking of the symmetry. In the asymmetric phase, its effect would just be to determine which of the two components is chosen. If the symmetry were explicitly broken by different external stimuli to the two input modules in the symmetric phase, we do not expect the behavior of the bi-modal module to change qualitatively. The attractor in this module would not be perfectly symmetric, but it will still have similar overlaps with the visual and auditory features.

We have checked numerically that this is indeed the case for our model. As expected, if the intensity or the degree of distortion of the stimuli on the two pathways is slightly different, the bimodal module always prefers the strongest signal or the one with fewer errors. For g less than ~ 0.1 the asymmetry determines which feature is preferred, and for large g it slightly unbalances the attractor towards one of the features, but leaves the overlaps with the visual and auditory components similar in magnitude.

The important point is that although perfect symmetry between the two sensory pathways may never be realized, the real multi-modular structures are probably close enough to this situation so that small, *subject to subject* differences in the structure or performance of the networks allow for *qualitatively* different responses across a given subject population. The precise statistics (distribution of responses) across this subject population in the experiment of McGurk & MacDonald (1976) would then reflect the distribution of the irregularities across subjects.

In this sense we propose that the mechanisms underlying the type of sensory processing displayed by the subjects in the experiment may be present in our model in a range of values of g close to the region where the symmetry breaking transition occurs. In that case, assuming that the stimuli are applied symmetrically to both sensory pathways, the visual or auditory responses would be given by subjects in the broken symmetry phase, and the combination-like responses would be given by subjects in the symmetric phase, which are not able to decide between any of the two alternatives. This was noted by McGurk & MacDonald, who pointed out that '*in the absence of domination*' which would correspond to a symmetric stimulation condition, this symmetrical situation probably exists in which the subject '*oscillates between the two possibilities*'.

Finally, it has to be said that there is a number of results from the experiment that we have not attempted to describe here. There is another experimental paradigm in which there exists a syllable (*da*) whose visual and auditory representations are correlated with those of the visual and auditory stimuli. In this case, most of the subjects (an overwhelming majority for some populations of subjects) report to have heard this, non-stimulated, syllable. The modelling of these situations is beyond the scope of the present analysis and we leave it for future work.

6 Discussion

A model for coupled attractor neural networks with independent, sparsely coded patterns has been studied in detail. Although the model is quite general, we have focused on the study of an architecture in which two of the individual attractor networks or modules receive external projections from outside but are not connected to each other (the *input* modules), and a third module only receives projections from the input modules but not from outside. The connections between the modules are assumed to be reciprocal and symmetrical. The inter-modular connections reflect the learning of associations between the memories of the individual modules. We have modelled them in such a way that each memory in the internal module is associated with one memory in each of the input modules, so the memories, or stored features, of the three modules can be grouped in triplets. Features belonging to the same triplet are said to be consistent or corresponding and features from different triplets are said to be inconsistent or contradictory.

A single parameter g measures the relative intensity of the inter- to intra-modular connections. Their absolute value is fixed by the constraint that the sum of the strengths of the connections onto a neuron in the central module does not change with g. Since the memories of different modules are associated in a similar way to the memories in a given module but with a different strength, we have focused on the study of the changes in performance of the multi-modular network as a function of their relative intensity g.

The model was also studied in conditions similar to those in experimental investigations on the processing of contradictory stimuli by different sensory pathways (McGurk & MacDonald 1976), as a means to see if the interaction between sensory modalities proposed in the model was able to capture some of the experimental evidence. This was indeed the case for the results in the experimental paradigm to which the model could be applied.

The conclusion of the first study was that, for the values of the parameters selected in this work, two qualitatively different regimes exist, corresponding to lower or higher values of the relative inter-modular connection strength. The first regime, that we called independent (I) phase, and which was realized for low values of g, has the property that the modules can interact with each other and at the same time be in any of their memories independently of the features present in the other two modules. There are several instances that exemplify the interaction between the modules in this regime, for example, if the network is in a quiescent state (corresponding to a state of spontaneous activity [4]), an external stimuli to only one of the modules similar to one of the features is enough to put the whole network in a global retrieval state in which the three modules are in attractors corresponding to the triplet of memories selected by the external cue. Also, the firing rates in the delay period in global attractors corresponding to consistent features are higher than those in attractors where the state of the modules corresponds to inconsistent features. This is a result of

[4] The spontaneous activity state and the quiescent state, although similar are not equivalent. That this property of multi-modular networks holds also in a more realistic model able to incorporate stable spontaneous activity states is something that still has to be studied.

the fact that when the network is in a global state corresponding to a triplet, the signals between the modules are consistent and reinforce each other, resulting in higher effective currents and therefore in higher firing rates.

What is important about this, is that each module is communicating to the others a persistent signal (the attractor is self-sustained) which is weak enough to not determine by itself the state of the receiving modules, but strong enough to have observable effects on their activity. It is serving as an external *bias* in the direction of the memory associated with the current state of our module. Such a *bias* has been predicted to play a role in the neural mechanisms of working memory (Fuster et al. 1985, Miller & Desimone 1994, Miller et al. 1996) and visual search (Chelazzi, Miller, Duncan & Desimone 1993, Desimone 1996). In fact, with a model similar to this one we have been able to describe satisfactorily, within the technical limitations imposed by the simplicity of the model, many of these experimental findings. A detailed exposition of these results is outside the scope of the present work and it will be presented elsewhere (Renart et al. 1999b).

The model was also studied in the context of the processing of contradictory (non-associated) information by different sensory pathways as studied in the experimental work by McGurk & MacDonald (1976). A qualitative explanation of one of the experimental paradigms can be extracted. When the persistent stimuli to the input modules are non-corresponding features of the same intensity (assumed to be the output of a previous area in a self-sustained state), two regimes appear, separated by a transition region: For large g symmetrical states appear in the convergent module equally correlated with the two stored features in this module associated with the two stimuli. The human perceptual correlate would be to give a combination response, e.g. *ba-ga* when the inputs to each module were *ba-ba* for visual and *ga-ga* for auditory. When the value of g is smaller, the attractor in this module becomes correlated with one of the stored features *or* the other, i.e. the symmetry is broken. In this case either the visual or the auditory stimulus is chosen. In between these two regions a small range of g exists in which the two behaviors coexist, so either type of response could, in principle, be given in this case. The fact that a simple model such as this one, is able to qualitatively reproduce some experimental results on the subject of the simultaneous processing of contradictory information suggests that, at least, some features of the interaction of different sensory modalities in associative areas in the cortex may be described by Hebb-like associations of the different local uni-modal representations existing in previous stages of the processing stream.

We have decided to present here for simplicity results regarding the behavior of a network with a small number of stored features. Having the number of stored features grow with the size of the network should not be essential for the phenomenology described to take place, and yet we believe that the main results obtained would still hold if this were the case. In fact, results obtained previously confirm that this is the case for most of the results presented here.

Finally, there are several aspects of the model which should be improved: The connectivity pattern proposed, although useful, does not take into account a proper separation of the synapses of the network into excitatory plastic, and inhibitory ones. No low-rate global unselective stable attractor exists in the model to take into account spontaneous activity, which is represented in the model by a quiescent state of no activity. The values of the mean firing rates in the selective attractors, although far from saturation, are too high to stand any quantitative examination, which diminishes the dynamic range of response of the neurons in the network, with the associated loss of flexibility. It will therefore be necessary to extend these results to a more detailed and realistic model able to take into account these and other quantitative considerations. A model of this sort would allow quantitative comparison with the numerous experimental data and would further develop the importance of associative recurrent networks in describing diverse cognitive phenomena. Progress is heading in this direction.

Acknowledgements

This research was partly supported by a British Council-Spanish Ministry of Education and Science bilateral program HB 96-46, and by Medical Research Council Programme Grant PG8513790 to E.T.Rolls. A Spanish grant PB96-47 is also acknowledged. Authors AR and NP are most appreciative for the hospitality shown to them while visiting the Department of Experimental Psychology, Oxford University, during the completion of this work. We also want to acknowledge an anonymous referee who helped us improve the manuscript significantly.

Appendix: Treatment of the dilution

To understand why a random dilution of the synapses is equivalent to a renormalization of the intensity of the connections if the number of stored patterns does not increase linearly with the size of the network, we follow the arguments and the procedure given in (Sompolinsky 1986, Sompolinsky 1987). We refer the reader to those references for the details. Let us first write the general expression for the synaptic connections (3),(4) when a fraction d_0 and d of the intra- and inter-modular connections respectively are randomly set equal to zero:

$$J_{ij}^{(a,a)} = \frac{J_0 d_{ij}^{0a}}{\chi N} \sum_{\mu=1}^{P} (\eta_{ai}^{\mu} - f)\,(\eta_{aj}^{\mu} - f) \qquad i \neq j \;\; ; \;\; a = \mathbf{A}, \mathbf{B}, \mathbf{C} \tag{18}$$

$$J_{ij}^{(a,b)} = \frac{g\; d_{ij}^{ab}}{\chi N} \sum_{\mu=1}^{P} (\eta_{ai}^{\mu} - f)\,(\eta_{bj}^{\mu} - f) \qquad \forall \; i, j \quad , \tag{19}$$

where J_0 and g are the initial intra- and inter-modular connection strengths (no normalization criterium has been used yet), and the variables d_{ij}^{0a} and d_{ij}^{ab} describe the connectivity within and between the modules, respectively. They are both $(0,1)$ binary quenched random variables which take the value one with probability d_0 and d (again respectively) independently of the pair of neurons observed and of the module they belong to. For the connections to remain symmetric, only half of these variables are drawn randomly, the other half are set equal to their symmetric counterparts.

This expression becomes simplified once the average over the dilution variables is performed. The result for an arbitrary number of stored patterns is that the effect of the random ablation of some of the synapses is equivalent to the addition of a random gaussian term to the mean value of the diluted synapses, where the average is performed according to the distribution of the dilution variables. The resulting final form of the synaptic efficacies of the multi-modular system is

$$J_{ij}^{(a,b)} = \frac{1}{\chi N} \sum_{\mu,\nu=1}^{P} (\eta_{ai}^{\mu} - f) \tilde{K}^{ab} (\eta_{bj}^{\nu} - f) + \tilde{D}_{ij}^{ab} \qquad ai \neq bj \quad , \tag{20}$$

where

$$\tilde{K} = < K >_{d,d^0} = \begin{pmatrix} J_0 d_0 & 0 & gd \\ 0 & J_0 d_0 & gd \\ gd & gd & J_0 d_0 \end{pmatrix} \quad ; \quad \tilde{D} = \begin{pmatrix} \delta_{ij}^{0\,(\mathbf{A})} & 0 & \delta_{ij}^{(\mathbf{AC})} \\ 0 & \delta_{ij}^{0\,(\mathbf{B})} & \delta_{ij}^{(\mathbf{BC})} \\ \delta_{ij}^{(\mathbf{CA})} & \delta_{ij}^{(\mathbf{CB})} & \delta_{ij}^{0\,(\mathbf{C})} \end{pmatrix} \quad . \tag{21}$$

The symbol $< \ldots >_{d,d^0}$ stands for an average over the random variables d_{ij}^{0a} and d_{ij}^{ab}. The variables $\delta_{ij}^{0\,(a)}$ and $\delta_{ij}^{(ab)}$ represent a random contribution to the synaptic efficacies between neurons i, j of module a and neurons i from module a and j from module b respectively. They are drawn from gaussian distributions of zero mean and variances given by

$$< (\delta_{ij}^{0\,(a)})^2 >_{d^0} \;=\; \frac{1}{N}\left[\frac{d_0(1-d_0)P}{N}\right] \qquad a = \mathbf{A}, \mathbf{B}, \mathbf{C} \tag{22}$$

$$< (\delta_{ij}^{(ab)})^2 >_{d} \;=\; \frac{1}{N}\left[\frac{d(1-d)g^2P}{N}\right] \qquad (ab) = (\mathbf{AC}), (\mathbf{BC}) \quad . \tag{23}$$

Since these variances are to be interpreted as synaptic couplings, the first factor $1/N$ in both equations provides the correct scaling of the connections with the size of the network in the thermodynamic limit. However, since we are assuming that P does not scale linearly with N, the two variances are actually proportional to $1/N^2$ and therefore they vanish as the size of the network becomes arbitrarily large. The only effect of the dilution is therefore a renormalization of the parameters J_0 and g, which are now equal to $\tilde{J}_0 = J_0 d_0$ and $\tilde{g} = gd$. A normalization criterium similar to the one used in the text could now be defined in terms of the new parameters $\tilde{J}_0$ and $\tilde{g}$.

References

Amit, D. J. (1995). The hebbian paradigm reintegrated: Local reververations as internal representations, *Behavioral and Brain Sciences* **18**: 617–657.

Amit, D. J. & Brunel, N. (1997). Model of global spontaneous activity and local structured delay activity during delay periods in the cerebral cortex, *Cerebral Cortex* **7**: 237–252.

Amit, D. J., Sagi, D. & Usher, M. (1990). Architecture of attractor neural networks performing cognitive fast scanning, *Network* **1**: 189–216.

Amit, D. J. & Tsodyks, M. V. (1991a). Quantitative study of attractor neural network retrieving at low spike rates: I. substrate–spikes, rates and neuronal gain, *Network* **2**: 259–273.

Amit, D. J. & Tsodyks, M. V. (1991b). Quantitative study of attractor neural network retrieving at low spikes rates: II. low-rate retrieval in symmetric networks, *Network* **2**: 275–294.

Baylis, G. C. & Rolls, E. T. (1987). Responses of neurons in the inferior temporal cortex in short term and serial recognition memory tasks, *Exp. Brain Res.* **65**: 614–622.

Baylis, G. C., Rolls, E. T. & Leonard, C. M. (1987). Functional subdivision of temporal lobe neocortex, *Journal of Neuroscience* **7**: 330–342.

Buhmann, J., Divko, R. & Schulten, K. (1989). Associative memory with high information content, *Phys. Rev. A* **39**: 2689–2692.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., Woodruff, P. W. R., Iversen, S. D. & David, A. S. (1997). Activation of auditory cortex during silent lipreading, *Nature* **276**: 593–596.

Chelazzi, L., Duncan, J., Miller, E. K. & Desimone, R. (1998). Responses of neurons in inferior temporal cortex during memory-guided visual search, *J. Neurophysiol.* **80**: 2918–2940.

Chelazzi, L., Miller, E. K., Duncan, J. & Desimone, R. (1993). A neural basis for visual search in inferior temporal cortex, *Nature* **363**: 345–347.

Chelazzi, L., Miller, E. K., Lueschow, A. & Desimone, R. (1993). Dual mechanisms of short-term memory: Ventral prefrontal cortex, *Soc. Neurosci. Abstr.* **23**: 975.

Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention, *Proc. Natl. Acad. Sci. USA* **93**: 13494–13499.

Fuster, J. M. (1990). Inferotemporal units in selective visual attention and short-term memory, *J. Neurophysiol.* **64**: 681–697.

Fuster, J. M. & Alexander, G. E. (1971). Neuron activity related to short-term memory, *Science* **173**: 652–654.

Fuster, J. M., Bauer, R. H. & Jervey, J. P. (1982). Cellular discharge in the dorso-lateral prefrontal cortex of the monkey in cognitive tasks, *Exp. Neurol.* **77**: 679–694.

Fuster, J. M., Bauer, R. H. & Jervey, J. P. (1985). Functional interactions between inferotemporal and prefrontal cortex in a cognitive task, *Brain Res.* **330**: 299–307.

Fuster, J. M. & Jervey, J. P. (1981). Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli, *Science* **212**: 952–955.

Hasselmo, M. E., Rolls, E. T., Baylis, G. C. & Nalwa, V. (1989). Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey, *Experimental Brain Research* **75**: 417–429.

Hertz, J. A., Krogh, A. & Palmer, R. G. (1991). *Introduction to the theory of neural computation*, Addison-Wesley, Wokingham, U.K.

Howells, T. (1944). The experimental development of color-tone synesthesia, *J. Exp. Psychol.* **34**: 87–103.

Kühn, R. (1990). Statistical mechanics of neural networks near saturation, *in* L. Garrido (ed.), *Statistical Mechanics of Neural Networks*, Springer-Verlag, Berlin.

Lauro-Grotto, R., Reich, S. & Virasoro, M. A. (1997). The computational role of conscious processing in a model of semantic memory, *in* M. Ito, Y. Miyashita & E. T. Rolls (eds), *Cognition, Computation and Consciousness*, Oxford University Press, Oxford.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices, *Nature* **264**: 746–748.

Miller, E. K. & Desimone, R. (1994). Parallel neural mechanisms for short-term memory, *Nature* **263**: 520–522.

Miller, E. K., Erickson, C. A. & Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque, *J. Neurosci.* **16**: 5154–5167.

Miller, E. K., Li, L. & Desimone, R. (1993). Activity of neurons in anterior inferior temporal cortex during a short-term memory task, *J. Neurosci.* **13**: 1460–1478.

Miyashita, Y. & Chang, H. S. (1988). Neural correlate of pictorial short-term memory in the primate temporal cortex, *Nature* **331**: 68–70.

O'Kane, D. & Treves, A. (1992). Short and long range connections in autoassociative memory, *Journal of Physics* **25A**: 5055–5069.

Renart, A., Parga, N. & Rolls, E. T. (1999a). Backprojections in the cerebral cortex: implications for memory storage, *Neural Computation* . In press.

Renart, A., Parga, N. & Rolls, E. T. (1999b). A recurrent network model of neural mechanisms related to visual working memory. In preparation.

Rolls, E. T. (1989). Functions of neuronal networks in the hippocampus and neocortex in memory, *in* J. H. Byrne & W. O. Berry (eds), *Neural Models of Plasticity: Experimental and Theoretical Approaches*, Academic Press, San Diego, pp. 240–265.

Rolls, E. T. & Tovée, M. J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex, *J. Neurophysiol.* **73**: 713–726.

Rolls, E. T. & Treves, A. (1998). *Neural networks and brain function*, Oxford University Press, Oxford, U.K.

Shiino, M. & Fukay, T. (1990). Replica-symmetric theory of the nonlinear analogue neural networks, *J. Phys. A: Math. Gen.* **23**: L1009–L1017.

Sompolinsky, H. (1986). Neural networks with non-linear synapses and a static noise, *Phys. Rev. A* **34**: 2571–2574.

Sompolinsky, H. (1987). The theory of neural networks: the hebb rule and beyond, *in* J. L. van Hemmen & I. Morgenstern (eds), *Heidelberg Colloquium of Glassy Dynamics*, Springer-Verlag, Berlin, pp. 485–527.

Treves, A. (1990). Threshold-linear formal neurons in auto-associative nets, *J. Phys. A: Math. Gen.* **23**: 2631–2650.

Treves, A., Panzeri, S., Rolls, E. T., Booth, M. & Wakeman, E. A. (1999). Firing rate distributions and efficiency of information transmission of inferior temporal cortex neurons to natural visual stimuli, *Neural Computation* . In press.

Treves, A. & Rolls, E. T. (1994). A computational analysis of the role of the hippocampus in learning and memory, *Hippocampus* **4**: 373–391.

Tsodyks, M. V. & Feigelman, M. V. (1988). The enhanced storage capacity in neural networks with low activity level, *Europhys. Lett.* **6**(2): 101–105.

Wilson, F. A. W., Scalaidhe, S. P. O. & Goldman-Rakic, P. S. (1993). Dissociation of object and spatial processing domains in primate prefrontal cortex, *Science* **260**: 1955–1958.