International Journal of Neural Systems, Vol. 12, Nos. 3 & 4 (2002) 187–201

RELATING PRIMING AND REPETITION SUPPRESSION

A. LÖRINCZ*, G. SZIRTES and B. TAKÁCS
Department of Information Systems,
Eötvös Loránd University of Sciences, Budapest, Hungary
**alorincz@axelero.hu*

I. BIEDERMAN
Department of Psychology,
University of Southern California, Los Angeles, USA

R. VOGELS
Laboratorium voor Neuro- en Psychofysiologie,
Catholic University of Leuven, Leuven, Belgium

Received 15 October 2001
Revised 9 March 2002
Accepted 15 August 2002

We present a prototype of a recently proposed two stage model of the entorhinal–hippocampal loop. Our aim is to form a general computational model of the sensory neocortex. The model — grounded on pure information theoretic principles — accounts for the most characteristic features of long-term memory (LTM), performs bottom-up novelty detection, and supports noise filtering. Noise filtering can also serve to correct the temporal ordering of information processing. Surprisingly, as we examine the temporal characteristics of the model, the emergent dynamics can be interpreted as *perceptual priming*, a fundamental type of *implicit* memory. In the model's framework, computational results support the hypothesis of a strong correlation between perceptual priming and repetition suppression and this correlation is a direct consequence of the temporal ordering in forming the LTM. We also argue that our prototype offers a relatively simple and coherent explanation of priming and its relation to a general model of information processing by the brain.

Keywords: Hippocampus; computational model; perceptual priming; generative networks; recognition; repetition suppression.

1. Introduction

In the last two decades considerable experimental work has been directed toward exploring the possible interactions of perception and memory.[24,55] Specifically, there has been a marked interest in distinguishing declarative and non-declarative (implicit) memory processes. Empirical evidence for the two-component view goes back to the late fifties and the sixties.[11,50,57] In distinguishing memory processes, a hierarchy of the information processing is tacitly assumed. In other words, higher order levels deal with data of increasing complexity, while levels of the same rank also exist and work separately on different pieces of information (which may have the same complexity). Both of the basic categories include sub-processes that may have subtle interactions or common neuronal correlates (e.g., Ref. 21). The implicit memory class includes conditioning, skill learning and the so called priming phenomena. The term "priming" in a broader sense refers to the observation that an earlier encounter with a given stimulus can modify ("primes") the response to the same or a related stimulus. This modification can lead to an increase in efficiency of recognition which can be

measured by accuracy or reaction time. Many of the early experiments attempted to reveal the dissociable memory components. It turned out that priming itself forms a class of several, somewhat different processes, including perceptual priming, semantic (or conceptual) priming and new association priming (see, e.g., Ref. 63 and references therein). Perceptual priming is the main focus of this article. Perceptual priming is considered to be present at the lowest level of the hierarchy of the information process, that is it may subserve or accompany the layers which work on the least complex inputs. Perceptual (or repetition) priming reflects prior processing of stimulus *form* whereas the other types of priming deal mainly with stimulus *meaning*.[7–9,17,53] Although — in principle — priming tasks are easily characterized as perceptual or conceptual, these terms are not appropriate to describe all priming tasks (and the supposed processes involved).[62] To avoid any confusions from now on by "priming" we shall mean perceptual priming.

Although the overlap between implicit memory tasks and processes is still an issue of interest, the dissociation of priming from declarative memory is well founded. The main difference between priming and declarative memory processes is that the facilitation associated with priming is not conscious and may be present in all perceptional, recognition and category forming tasks. There are two lines of neurophysiological and psychological evidences that support the dissociation. First, patients with global amnesia due to bilateral medio-temporal insult exhibit more or less normal priming in many different tasks, like word recognition, word-stem completion and so on (see, e.g., Ref. 21 and references therein). Second, in healthy subjects parallel dissociations of priming and declarative memory is possible.[9,53] Remarkably, priming is not simply a "side-effect" of recognition: Under some special conditions objects become recognizable only *after* some earlier displays[60] (see also subliminal visual priming[5]).

The neuronal correlate of priming is thought to be *repetition suppression*. This belief is supported by the joint appearance of the two phenomena in many experiments (see, e.g., Refs. 4 and 13). During the experiments with repeated stimuli the neurons of the inferotemporal and frontal cortex respond with decreasing activity, while perception is facilitated. This phenomenon is barely dependent on the type of task. One of the major paradigms is called "delayed matching to sample" (DMS,[47]), in which the monkeys have to respond to the seen inputs with some delay. According to the measurements half of the corresponding cells respond weakly or do not respond at all at the second display. Given all these facts, it is still not obvious how priming may serve any memory process or how it is mediated by higher level neocortical areas.[61]

Assuming the theory of a memory system of cooperative units, the focus of our attention is on the possible binding of priming and the explicit memory functions. According to the experiments, priming manifests in some temporal change of different memory linked function, that is priming is inherently a dynamical property of our memory. To examine this temporal phenomena, we extended and modified our previously proposed[15,42,44] functional model of the hippocampus *and* its environment, the entorhinal-hippocampal loop (EHL). The new modified model is based on a minimized set of assumptions and intended to *derive* the remaining properties of the neurobiological substrate.[43,45] The aim here is to build a model, which can be considered as a general *prototype* for information processing in sensory neocortical areas. Our starting point offers a solution to the so called homunculus fallacy.[58] The proposed solution suggests generative or reconstruction networks.[41,45] Arguments related to the homunculus fallacy will be neglected here and we shall start from generative networks. The idea that generative networks may form the basis of perception was suggested first by Horn.[29] On generative networks see, e.g., Ref. 28 and references therein.

In any modeling effort a central issue is the formation of the internal representation. In particular, the handling of new incoming information is of special interest, insofar as novel information cannot be (fully) generated from the internal representation. A general scheme of information encoding would be the following:

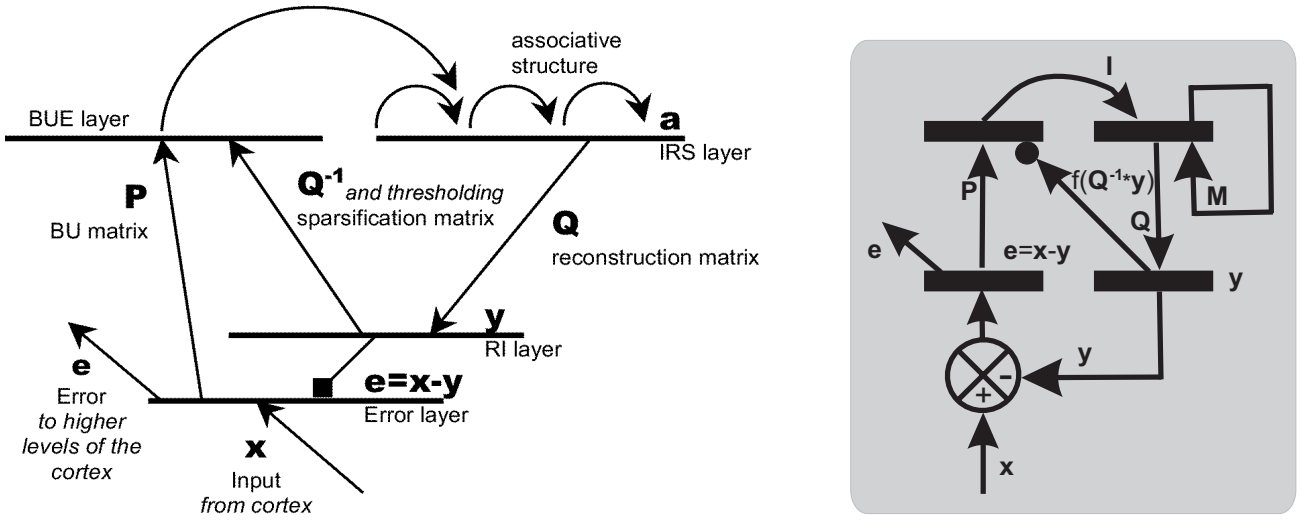
(i) noise rejection, i.e. the rejection of input portions that do not make sense according to previously learned encoded information,
(ii) analysis of the rejected noise for novelty, i.e. not-yet encoded correlations in the rejected information,
(iii) optimization of transfer of novel information, and

(iv) encoding of the novel information, i.e. adding the novel information to the encoded memory and being able to reconstruct this information.

We have demonstrated by computational studies that maximization of information transfer via independent component analysis (ICA,[1,6,16,36,37] for a recent review on ICA see Ref. 32) can detect novelty simply by analyzing the activity distribution of the ICA-transformed signals.[45] Computer simulations to investigate the new model show that the above described logical division of encoding emerges naturally and that some properties of perceptual *priming* can also be derived from the structure of the model. Our work is meant to be the first step towards a unified explanation of the possible relation of priming and declarative memory.

2. Model Description

In this paper we extend our recently proposed model[42,44,45] which was guided by principles of information transfer optimization and proper ordering of learning. The model, depicted in Fig. 1, is a special case of generative reconstruction networks. The system's "goal" is to reconstruct the inputs by learning their statistics. Activities are represented by lower case letters denoting vectors. Upper case letters denote the synaptic system, that is the connection matrices. This network may be viewed in terms of cortical processing layers, which receive input and provide outputs for further processing. The network is given an input ($\mathbf{x}$). The internal representation system (IRS) layer encodes the input ($\mathbf{a}$). The IRS activities are transformed into the reconstructed input ($\mathbf{y}$) via $\mathbf{Q}$. In the next step, the



(a) (b)

Fig. 1. The generalized reconstruction loop with an auxiliary noise filtering system. (a) Inputs coming from the cortex activate the bottom-up error (BUE) layer through the BU (bottom-up) filter that can be quickly tuned. Tuning makes use of statistical analysis to maximize information transfer. The output of the BUE layer, is temporally integrated at the internal representation layer (IRS). Beyond temporal integration, the associative structure of this layer may perform pattern completion, as well as other functions. Reconstruction is the task of the top-down matrix $\mathbf{Q}$. This matrix can be tuned slowly and holds elements of the long term memory. The reconstruction error ($\mathbf{e}$), that is the difference between the original and the reconstructed inputs, is processed in different ways: (i) It takes part in the fast tuning of the network (BU tuning). (ii) It becomes the signal to be processed at a higher level of the information processing system. The *sparsifying* inner loop performs noise filtering. In our simplified case — when the BUE-to-IRS transformation and the associative structure of the IRS layer is neglected — this inner loop is made of TD matrix $\mathbf{Q}$ and BU matrix of the inverse of $\mathbf{Q}$ *and* an additional nonlinear transformation. Black arrows: mostly excitatory "synapses." Black square: mostly inhibitory "synapses." (Detailed description of the model can be found in Refs. 15, 44.) (b) "Flow-chart" of the model. BUE-to-IRS transformation is the identity transformation for the sake of simplicity. M: associative structure of IRS layer. Sparsification function is $f(\cdot)$, which is acting on $\mathbf{Q}^{-1} * \mathbf{y}$.

emerging error ($\mathbf{x}$–$\mathbf{y}$) is processed via the bottom-up (BU) transformation. It results in an activity pattern at the bottom-up processed error (BUE) layer. Hence, the task of the IRS layer is temporal integration of the reconstruction error. Activity at the BUE layer is sparsified by the sparsification connections. In other words, activity has to be modified by a non-linear filtering transformation to gain a sparse form. This process makes use of the inverse of matrix $\mathbf{Q}$ and a component-wise non-linear function, which can be viewed as component-wise thresholding. The non-reconstructed part of the input, which is regarded as "noise" by the actual layer is forwarding to higher layers. It is important to emphasize that the whole construction is not arbitrary: Each part is well founded and supported by recent results in learning theory.[31,45,48] It has been shown that learning is Hebbian for each connection matrix of this network,[44] therefore the structure is biologically plausible. The working of the architecture can be understood via the following simplified equation:

$$\Delta\mathbf{a} = \mathbf{P} * (\mathbf{x} - \mathbf{Q} * \mathbf{a})\,. \tag{1}$$

where "$*$" denotes matrix-matrix (matrix-vector) multiplication. In the base model the activities of the BUE layer (right hand side of Eq. 1) are transmitted to the IRS layer through an identity matrix for simplicity (on the possible role of the transformation that works in place of the identity transformation, see Sec. 5). Activity of the IRS layer changes until input ($\mathbf{x}$) and reconstructed input ($\mathbf{y} = \mathbf{Q} * \mathbf{a}$) become equal, i.e. the IRS layer performs temporal integration. Under suitable conditions the equation stops when $\mathbf{a} = \mathbf{Q}^{-1} * \mathbf{x}$, where $\mathbf{Q}^{-1}$ denotes the inverse of matrix $\mathbf{Q}$. That is, the internal representation is determined by the "synaptic strengths" of matrix $\mathbf{Q}$. According to the model, this matrix represents the connections between the deep layer and superficial layers and can be interpreted as the locus of the long-term memory (provided that its temporal changes are negligible relative to the changes in other parts of the architecture). Fast changes for each input or input sequences can modify the BU matrix and the IRS associative structure. Such changes may modify, for example, the reconstruction speed but have little if any influence on the relaxed activities, because they are almost solely determined by matrix $\mathbf{Q}$ (see Eq. 1). Grouping the inputs to learn statistics, and generalization to unseen examples based on the learnt statistics may be the task of the associative structure in the model.

Although Eq. 1 is able to capture essential segments of memory building, it does not account for some intrinsic and possibly important features of information processing:

- Input is noisy and denoising is necessary. Denoising and sparsification are closely related.[31] Denoising built on minimization of mutual information promotes novelty detection.[45]
- A sparse representation reduces interference between components of LTM.[46]
- The resulting representation facilitates working in an "economic mode."[3]

On one hand, sparsification (denoising) and tuning of LTM require two phases for efficient processing and learning.[1,31,49] On the other hand, two-phase wake-sleep operation of the hippocampus (in the case of rodents, at least) has been found experimentally.[14] The modified equation that performs denoising is the following:[49]

$$\Delta\mathbf{a} = \mathbf{P} * (\mathbf{x} - \mathbf{Q} * \mathbf{a}) + f(\mathbf{a})\,, \tag{2}$$

where $f(\mathbf{a})$ is the so-called sparsifying term. We used the following functional form in our simulations:

$$f(\mathbf{a}) \doteq -\alpha * \frac{\mathbf{a}.}{\mathbf{a}.^2 + c^2} \tag{3}$$

where α defines the weight of the sparsifying member, c is responsible for the shape of the function, and the argument of $f(\cdot)$ indicates that the function acts separately on each component of the IRS layer (i.e. on the elements of vector $\mathbf{a}$).

In this work the role of the associative connections of the IRS layer will not be examined. However, it may be worth mentioning that both prediction and association can be embedded into the model.[2,59,64] Three aspects summarize the differences between the original model and the prototype model presented in this paper:

(i) The BU processing now consists of only one stage instead of two. The main reason is that the base model was designed to describe the hippocampal-entorhinal loop, which has a distinctive role in organizing the whole memory system (see also Sec. 5). This function has some special demands on speed

and timing. Two stages can establish fast learning by implementing the so called natural gradient[1] directly. BU learning can be slower in the prototype model.

(ii) Temporal deconvolution is the assigned role of the dentate gyrus in the original model. Temporal deconvolution is neglected here.

(iii) In the modified model the role of the BU part of the internal loop is noise filtering, which replaced the original suggestion for a predictive structure. Note that the recurrent collaterals of the IRS layer are also able to maintain a directed predictive structure.[51]

Our computer simulations concern the effect of changes in **Q** (the LTM) and the **P** (the filter) on the memory system and the reconstruction dynamics. In particular, the influence of the incremental tuning induced by new inputs will be examined.

3. Methods

Temporal behavior and the dynamics of the model were studied using two different types of inputs. The first dataset consists of sound mixtures. We have chosen natural (animals, forest, ...) and artificial (music, machines) acoustic samples, which were downloaded from freely available internet sources (e.g., Ref. 67). This choice allows controlled manipulation of mutual information between inputs and bottom-up processed signals. Two different input sets were made of sounds, each was a mix of three samples (Fig. 2). Two random mixing matrices were used to make linear mixtures of the samples. The resulting mixtures were cut into training and testing sets. In both cases the sets contained 4000–4000 instances (inputs) of the mixtures.

The impact of the independence on the system's behavior was investigated in the following fashion. The three original samples were corrupted by adding

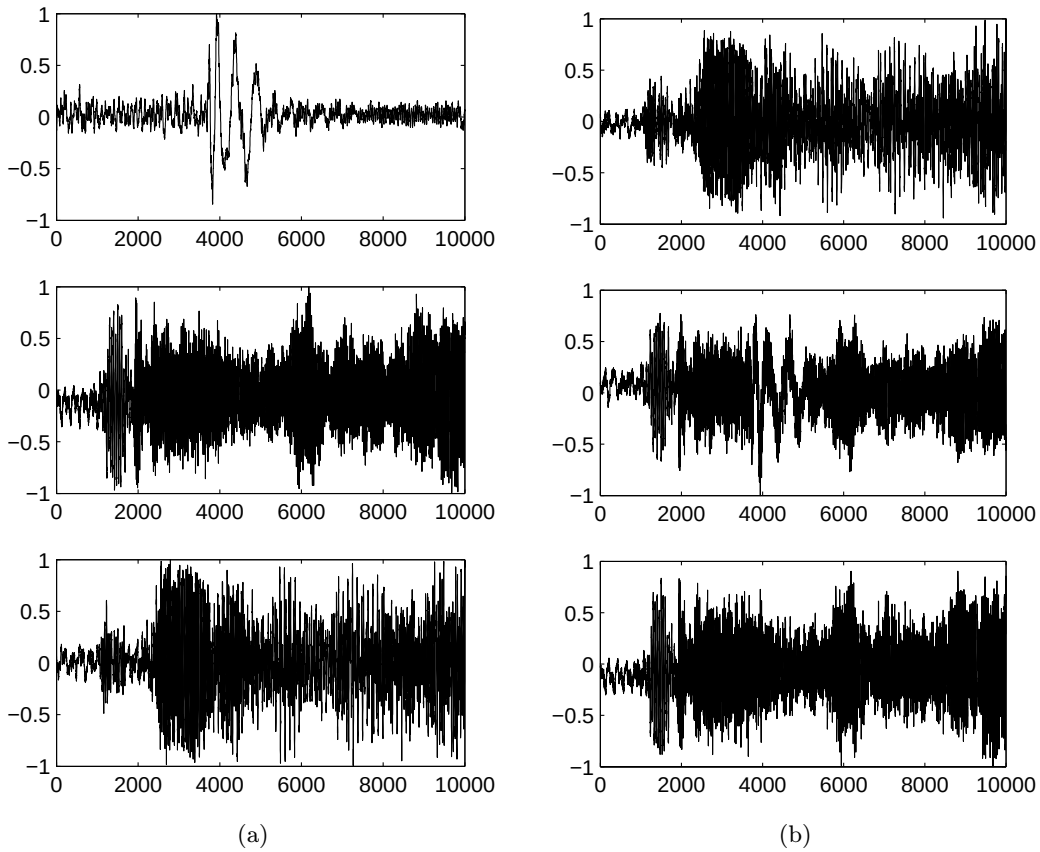


Fig. 2. The original and the mixed sounds. For simulations we used the linear mixture (B column) of three, randomly chosen and independent sound samples (A column). The components of the mixing matrix were random and were kept fixed.

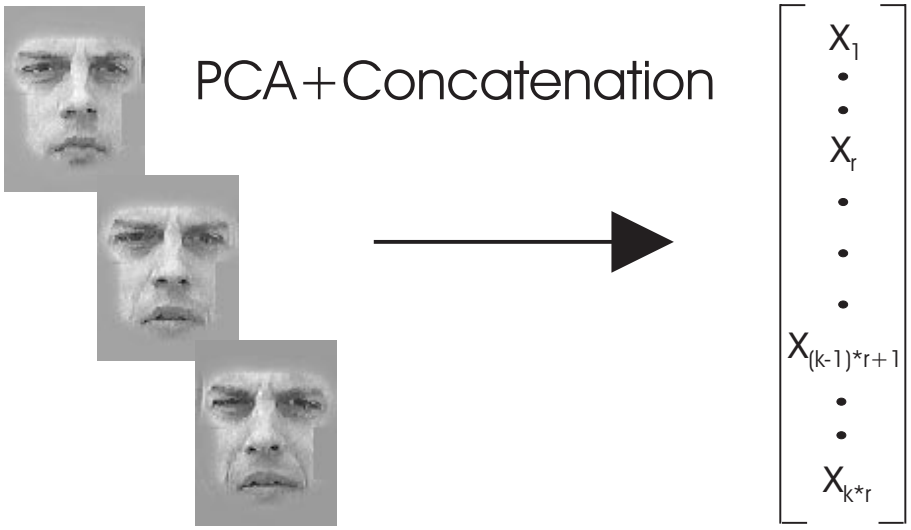


Fig. 3. The input made of facial expressions. Streams of bitmap images of size 72×96 pixels showing different emotions were used. Inputs were formed by concatenating the pixel values of three non adjacent images from the sequences. Starting from the nth image of the sequence the $n+3$rd and the $n+6$th images were chosen to form a three image long input. Dimension of the concatenated vectors is $3 \times 72 \times 96 = 20736$. Principal component analysis (PCA) was employed to reduce the dimensionality of the vectors to 1584. The emerging compressed vectors formed the inputs for ICA analysis. An example sequence of images belonging to facial expression *disgust* is shown. Parameters: Number of concatenated images: $k = 3$, reduced dimension of one image: $r = 528$.

a fourth sound sample in a parameterized way: $S_i^* = (1-w) * S_i + w * S_4$, where $i = 1$, 2 and 3, S_i denotes the ith original sound, $w = 0, 0.05, \ldots, 0.25$ is the weight of the additional sound.

Beside the example of sounds for easily analyzible structure, we have chosen movies of facial expressions for the second database. Video movies were collected in collaboration with the Department of Psychiatry of the Medical School of Budapest. The videos display the six basic emotions (anger, happiness, disgust, surprise, sadness and fear). The database contained about 200 video movies of 40 people. The number of movies per person varied between 2 and 12. Each movie consisted of 15–50 frames of the development of facial expressions. Bitmap images of size 72×96 underwent digitization, scaling, transformation, and gray scale normalization. The inputs were made as follows. Starting from the nth image of a frame sequence the $n+3$rd and the $n+6$th images were chosen to form a three image long input. Each bitmap was transformed to vectors and the three vectors of the three images were concatenated. The concatenation is motivated by both information theory and neurobiology. First, grouping the signals improves coding efficiency.[18,45] Second, from the point of view of reverse engineering the receptive fields, it is known that ICA filters on image *sequences* (rather than on *single* images) which is in reasonable agreement with receptive fields of simple cells of the primary visual cortex.[25] Finally, it has been suggested that gamma waves form temporal sequences in the hippocampus.[35,39] The dimension of the concatenated inputs ($3 \times 72 \times 96 = 20736$) was reduced to 7–8% (to 1584) by principal component analysis (PCA,[30,26]) (Fig. 3). Following this procedure we created an input database for every emotion.

The databases belonging to the different facial expressions were divided into a training set and a test set. The system's behavior was studied using Eq. 1. It was assumed that tuning of BU and TD matrices takes place on different time scales and could be considered independently. For both the sound and the facial expression input sets we followed the next procedure to imitate the distinction of fast BU learning and slow TD learning of our model in a reproducible manner:

(i) The training set was used for *initial training.* ICA matrix was developed by the FastICA algorithm.[33] The ICA (or filtering)

matrix and its inverse were used as **P** and **Q**, respectively.

(ii) For each test input the relaxed activity values "**a**" were computed and the time of relaxation was measured by the iteration cycle number and was *averaged* over the training set. (The stopping condition of the iteration was fixed for all runs.)

(iii) Part of the test set was selected and was added to the training set. The new training samples were excluded from the test set. A new ICA matrix was developed for the new training set. The filtering matrix **P** was replaced by the new ICA matrix. Matrix **Q** was kept fixed ($\mathbf{Q} * \mathbf{P} \neq \mathbf{I}$ where **I** denotes the identity matrix).

(iv) Back to (ii).

The FastICA algorithm was chosen, because this algorithm has no tunable parameters that could misguide the computations. The FastICA algorithm was used in *batch* mode, because the expected changes for single inputs could have been subject to accidental factors.

4. Results

In the course of the computer runs we examined the temporal behavior of the differential equation Eq. 2. The impact of familiar inputs on the temporal behavior was studied. Our test results can be summarized as follows:

- The activity distribution is super-Gaussian after the initial training. Super-Gaussian representations are called sparse representations. Increasing the the training set made the activity distribution sharper on the test set, with the activity distribution shifted further away from a Gaussian distribution. The relative probability of higher activity components decreases.
- Thresholding increases the relaxation speed parallel with the increase of the ratio of familiar inputs in the training set (Fig. 4).
- Parallel with the speed up of relaxation, relaxed average activities of the IRS layer become smaller (Fig. 5).
- While the average activity of the "neurons" of the internal representation layer (IRS) decreased, in some cases an *increase* of activity can also be experienced.

It is important to note that the sparsification term in Eq. 2 accelerates relaxation, but the reconstruction error is smaller *without* this term. The main reason of the deterioration of reconstruction is that the TD matrix was not tuned and BU and TD matrices do not invert each other. The denoising made by the BU matrix is not optimal for a non-inverting TD matrix.

The relaxation equation that governs the dynamics of the system has two variable parameters (see Eq. 2). Computer runs were executed using parameters from a broader range ($0.01 \leq c \leq 0.41$ and $0.01 \leq \alpha \leq 0.21$). According to these computer

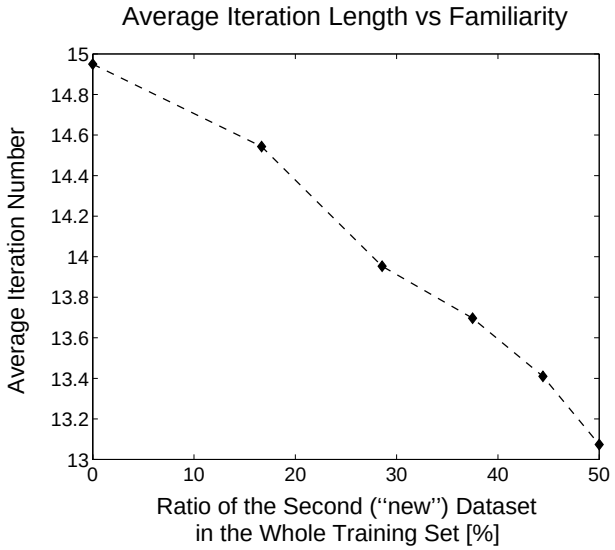


Fig. 4. Priming. Effects of *familiarity* of the inputs on the reconstruction were studied by gradually changing the training set. The results shown on this and the next two figures were developed by computer runs using the sound sample database. Inputs from the test set were included into the training set in a gradual fashion. Inputs in the training set can be viewed as *familiar* inputs. Relaxation time became shorter when the ratio of familiar inputs in the whole training set increased. Note that the increase of the training set concerned only the training of the BU matrix, TD matrix was untouched. The shortening relaxation time can be interpreted as the model of perceptual *priming*. Horizontal axis: The relative ratio of new training inputs (selected from the test set) and added to the training set. Vertical axis: Average iteration number on 300 test inputs. The stopping condition for the iteration was the same for all computer runs. (Parameters of Eq. 2: $\alpha = 0.05$ $c = 0.18$ $\eta = 0.001$.)

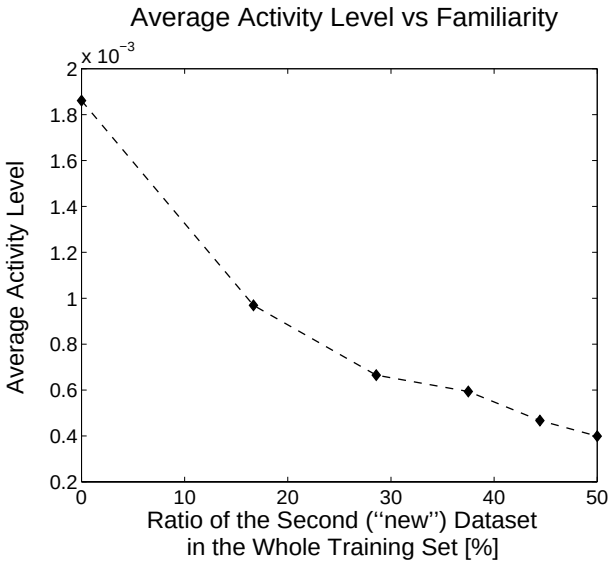


Fig. 5. Repetition suppression. The effect of ICA transformation and the sparsifying process on the activity of the internal representations as a function of the ratio of the familiar inputs in the training set. The larger the relative size of the set of familiar inputs within the training set the smaller the average activities. This activity decrease can be considered as a model of *repetition suppression*. Horizontal axis: The relative ratio of new training inputs (selected from the test set) to the whole training set. Vertical axis: Average relaxed activities on 300 test inputs. (Parameters of Eq. 2: $\alpha = 0.05$, $c = 0.18$, $\eta = 0.001$.)

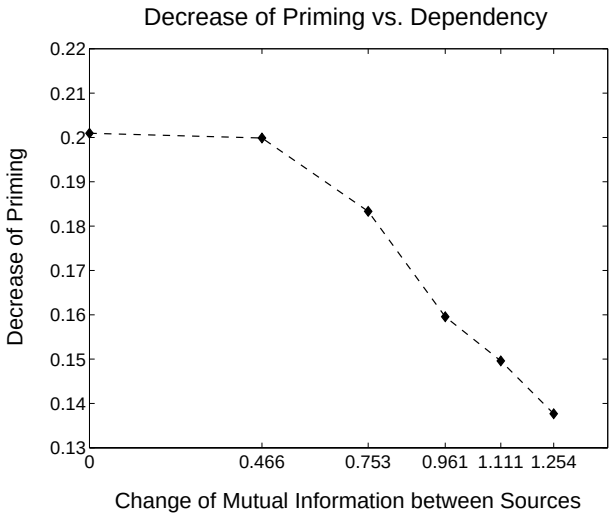


Fig. 6. Effect of statistical dependency between BUE components. Inputs belonging to the BU ICA matrix $\mathbf{P}$ were contaminated as described in the main text. (Note that $\mathbf{P} \neq \mathbf{Q}^{-1}$.) The priming effect is a monotonic decreasing function of the amount of contamination. Priming may completely disappear at higher values. Experiments were conducted by gradually mixing a fourth sound sample into the inputs. The decrease of independence resulted in the decrease or complete disappearance of the priming. Horizontal axis: the relative mutual information of the sound mixture (In mathematical form: $\Delta I_M = I_M(corrupted) - I_M(original)$, where, $I_M(corrupted)$ denotes the mutual information of the corrupted mixture and $I_M(original)$ the mutual information of the original mixture.) Vertical axis: Relative change of the decrease of iteration number until stopping condition for Eq. 2 is reached. (In mathematical terms: $(\max(N) - \min(N))/(\max(N))$, where N denotes the average iteration number for the test input set.)

runs, results show mild sensitivity to these parameters within the range studied.

Some differences were found between the two input sets. We have investigated the possibility that these differences are due to the mutual information (MI) of the ICA components. We studied the effects of degrading independence in a well controlled manner. This goal is feasible for the acoustic signals where the original signals are recovered by the ICA transformation. The three sound samples were contaminated by mixing an additional (fourth) sound sample (see Sec. 3). The results are shown in Fig. 6. According to the simulations the priming effect may weaken or may even disappear with the increase of mutual information between components at the BUE layer. The following expression was used to estimate the independence of the sources:

$$I_M = \sum_{i=1}^{3} \mathbf{H}(\mathbf{S}_i) - \mathbf{H}_{\text{joint}}(\mathbf{S}) \qquad (4)$$

I_M denotes the mutual information, $\mathbf{S}_i$ is the ith component of the original source, $\mathbf{S}$ is the set of the sources, $\mathbf{H}$ notes the differential entropy (for an excellent introduction, see, e.g., Ref. 18).

For facial expressions the independence of the components is not ensured (the inputs are not mixed from independent sources). Moreover, the size of the whole data set is small. And yet, using this imperfect data base the average total activity still exhibited repetition suppression. In a few cases, components of the IRS layer *increased* their activity when the training set was increased. The explanation of this finding — according to our results — is that sparsification is improved by the enlargement of the

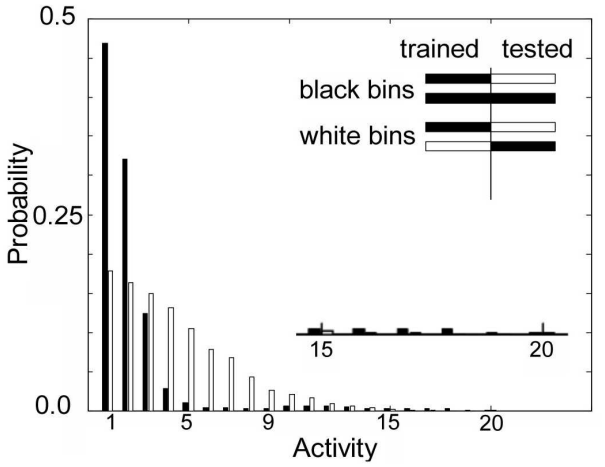


Fig. 7. Repetition suppression and repetition enhancement for facial expressions. Increased familiarity resulted in a sharper distribution. Results depict histogram without denoising. The *white* histogram belongs to the simulations with novel inputs: the training and the test input sets are disjunct and have different statistical properties. The black histogram is for the familiar inputs: some elements in the training set share the statistical properties of the test set (but no training input was used for testing, naturally). It can be observed that the probability of higher activity for some units may increase (see inset). This latter phenomenon can be interpreted as *repetition enhancement*. Denoising further increases the magnitude of both effects (not shown).

training set against the test set. At a cognitive level this phenomenon would correspond to the build up of a new memory element. The phenomenon is general, repetition suppression and repetition enhancement can be seen both for the non-sparsified activities (Fig. 7(a)) and for the the sparsified activities (Fig. 7(b)). The effect is more pronounced for the latter case. Figure 7 shows the results for the facial expression database. At the level of electrophysiology the phenomenon of increased activity may correspond to the *repetition enhancement* found in many experiments.[8–9,17,27,34,56] Our interpretation is based on the model. It may be important to note that the interpretation of the experimental results is far from obvious at present (see, e.g., Ref. 22 and references therein).

5. Discussion

Learning steps warranted by the model are as follows:

(i) Novelty is signaled by non-sparse BUE activity pattern.
(ii) BUE output is gated by sparsification.
(iii) "Noise" is estimated and information transfer is maximized in the BUE processing channel via ICA. Denoising (sparsification) cannot withhold high amplitude components of the BUE layer and information (structure) is transferred to the IRS layer.
(iv) Structured information is encoded into TD long-term memory.

In turn, the architecture learns structure and rejects noise. Both cognitive and neurophysiological experiments show that neurons in the neocortex respond with less and less activity in the case of repeated stimuli (e.g., Ref. 54). This repetition decrement is often called "repetition suppression" in the primate literature.[19,66] The characteristic features of this decrease are the following:

- Stimulus-specific.
- The response decreases in a monotonic fashion with the number of repetitions.
- It can be maintained for a relatively longer time period.
- It is experienced during passive fixation, in anesthesia or after cholinergic blockade. Based on these findings it is assumed that this phenomenon is automatic and is an intrinsic property of cortical neurons.
- Probably in most cases no explicit attention is required.

These features are emerging properties of novelty detection, noise rejection and denoising in our architecture. The form of novelty, however, is probably polymorphous and there is increasing evidence that different brain areas share the task of recognizing different aspects of novelty *within* the same scene. It is known that regions of the rat brain associated with discriminating the novelty or familiarity of an individual item/object differ from those responding to spatial arrays of objects/scenes.[65] Perirhinal cortex and area TE of the temporal lobe are more significantly activated by pictures of novel rather than familiar objects. In contrast, the hippocampus is not differently activated by the two input types. However, in the case of images of novel *arrangements* of familiar items postrhinal cortex and subfield CA1 of

the hippocampus produce significantly greater activation than for familiar arrangements. There is significantly less activation in the dentate gyrus and subiculum and the perirhinal cortex and area TE are not differentially activated. It thus seems that the encoding of novelty — the origin for priming and repetition suppression in our model — is distributed. In turn, we suggest that our model may serve as a prototype model for neocortical sensory processing areas, including the EHL loop.

According to the aforementioned properties, repetition suppression has been suggested to form the neuronal base of priming.[19] The goal of our work was to study the temporal behavior of our previously presented model. According to the simulations, when inputs were presented repeatedly to the system, the emerging dynamics strongly resembles (or can be interpreted as) perceptual priming. That is, instead of directly modeling this as a particular implicit memory process, a change in activity that could readily be interpreted as priming emerged from the system properties as a "by-product." It is worth noting that other types of priming may also emerge from the model: The associative structures at the level of IRS or at the level of the reconstructed input may be perfect candidates for this goal. It is thus possible that semantic or other higher level forms of priming[23] have different origins than implicit priming. An important finding in research on priming is that in Alzheimer's disease most forms of declarative memory are degraded but perceptual priming is still evident.[20] This finding also supports the hypothesis that the neuronal correlates of the different cognitive functions are separate both in anatomy and physiology. The model outlined in this paper exhibits all the marked properties of the collaborating memory subsystems and supports speculations on the assumed relation of priming as a cognitive event and the repetition suppression as its neuronal correlate. The BU tuning of the loop represents the response of the system for repeated stimuli. This process accelerates the dynamics of the system (Fig. 8) and can be interpreted as perceptual priming. During perception of repeated stimuli the TD connections are slowly tuned to form the *inverse* of the BU matrix. The TD matrix can be considered as the LTM, or part of the LTM of the model. From the point of view of priming, the "quality" of the LTM seems to be secondary. What matters is that the BU matrix should represent minimized mutual information. Activity

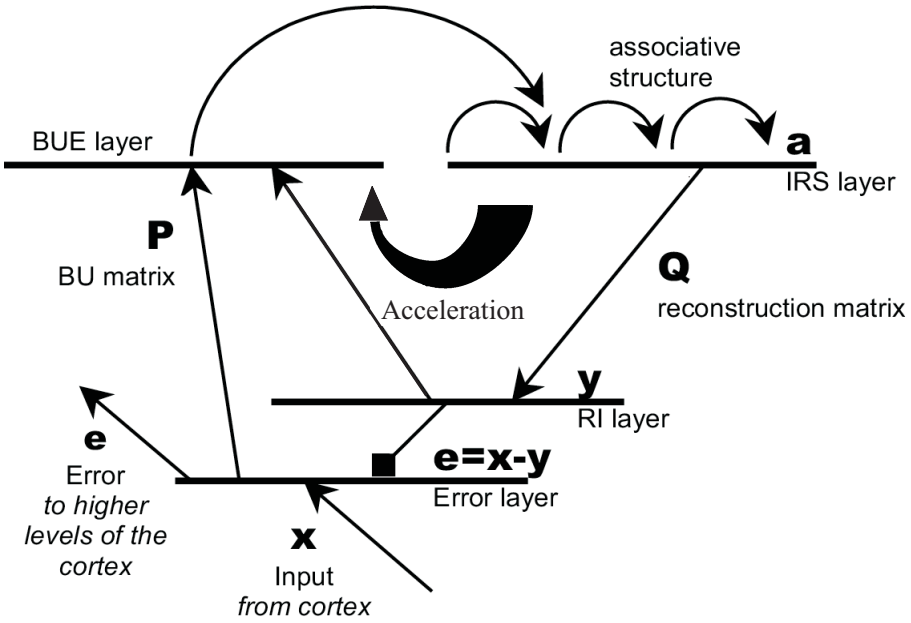


Fig. 8. Priming in the model IRS layer. The relaxation of the dynamics becomes faster and the sparsification increases upon tuning the BU matrix (**P**) and leaving the TD matrix (**Q**) intact. This effect appears in the model and according to our conjecture it corresponds to perceptual priming. The conjecture is supported by the natural separation of time scales of BU analysis and TD encoding of the model. (For further discussions, see Sec. 5.)

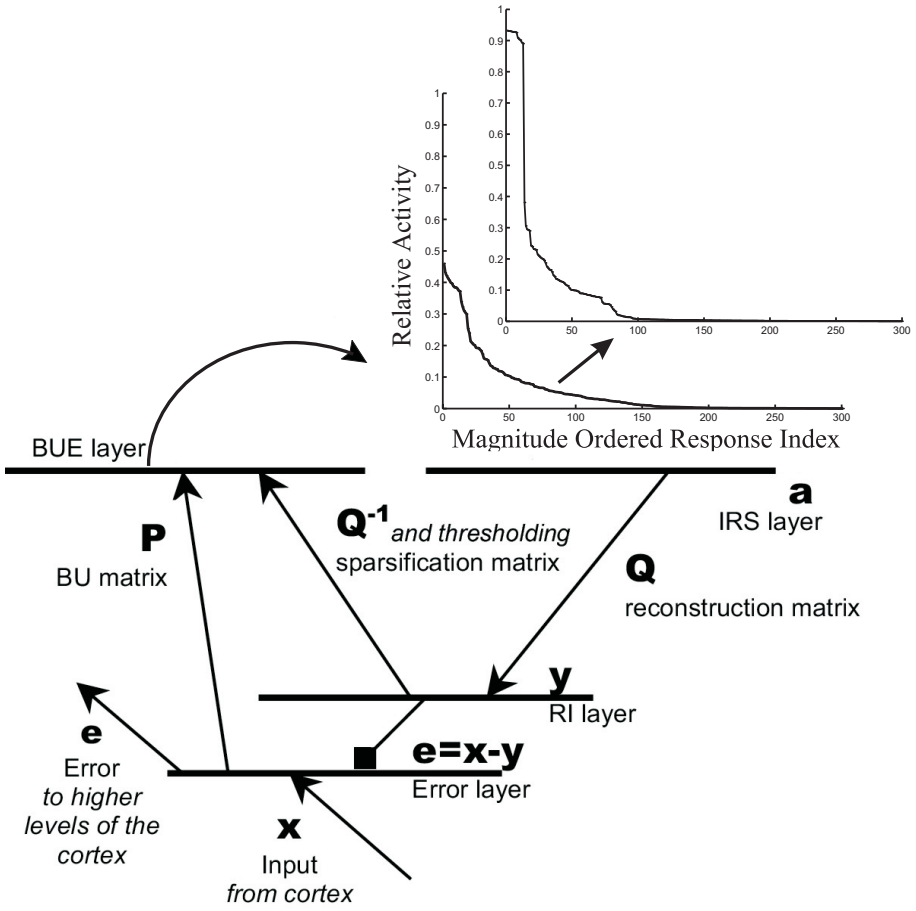


Fig. 9. Repetition suppression and repetition enhancement in the IRS layer. The ensemble activity at the integrating layer decreases upon tuning of the BU matrix (**P**) and leaving the TD matrix (**Q**) intact. This is a natural property of the model. We conjecture that it corresponds to the effect of *repetition suppression*. An example depicted in the top-right insets at the IRS layer. The inset shows the magnitude ordered activities of one component of the IRS layer for the facial expression database. Upon the tuning of the BU matrix, the response for most of the inputs became weaker (repetition suppression). Also, the response of the depicted component of the IRS layer became stronger for a few inputs (repetition enhancement), which can not be seen with this resolution, but is demonstrated on the inset of Fig 7.

changes upon repetition can be seen in Fig. 9 for a single computational unit of the IRS layer.

Below, four points that we consider relevant for further modeling are listed:

(i) The interpretation of event related fMRI, PET, etc., measurements is not obvious. We have shown that input activities (activities at the afferent synapses) of an area and output activities from that area (neuronal firing) may differ in a significant manner during different phases of learning. This issue is, indeed, unresolved yet.[4,22,40]

(ii) The transformation between the BUE layer and the IRS layer was not modeled. The identity matrix of our simplified model could be replaced by any kind of transformation that preserves information (i.e, any matrix of full rank could be inserted). As far as ICA is considered, it is intriguing that the resulting components are excellent for filtering noise (i) simultaneously, (ii) in a distributed manner, and (iii) using local information. In spite of these useful properties, ICA has a serious limitation: Its components have only limited interpretation power. One way to

overcome this limitation is to consider transformations other than the identity between BUE and IRS layers. As described above, the identity transformation assures temporal integration. In this case, the only reason for building associations is to make the system do temporal predictions. Though it sounds reasonable, a more adequate choice of transformation may provide the internal representations with improved interpretation power. According to our current thinking, interpretation could be better served by a layer that *breaks* information into parts. For example, let us consider the case of a steady input and let us make that input a face. Parts of the face are mouth, eyes, nose, hair, etc. In this case, it makes sense to build associations between components, e.g. for the completion of partially occluded faces. Interestingly, there exist models,[10] which suggest that the representation of faces is not in terms of its obvious parts: special filter values can also meet the requirements. Fortunately, the nature of these parts does not effect the possible extension of our proposed model.

(iii) The associative structure (not included in our model yet) may take part in different tasks, than representing spatial or temporal associations. It has been suggested that these structures may play a key role in categorization.[2,38,59]

(iv) According to the model, the hippocampus and its environment may participate at least in two forms of memory formations, namely in priming and in the encoding of long term memory. Nevertheless it is possible that the loop is concerned in other memory formation tasks as well.[52]

(v) We feel that the nature of novelty is still obscure and it is an open question how we discriminate among parts of a scene along the line of properties like "part of the context" or "being an independent object." Although the hippocampus is allegedly not concerned with detecting single (independent) stimuli,[12] it is not fully understood under what circumstances we emphasize one aspect and ignore others. To illuminate this problem regarding novelty as a special aspect, one example may be given. Recognizing a face and recognizing a facial expression are separate tasks of probably separate areas, and the level of their interaction is not known. It is possible that the same original scene may be the subject of two different transformations, which result in two different, invariant (regarding differing properties) representations. (A not too far analogy would be the pre-wired parvo and magno pathways that convey different information content of the same scene.) The novelty content from one aspect may be irrelevant in forming the other representation and *vice versa*. In our example, if one region is involved in detecting faces, novel *expressions* have no (or small) impact on its working. In this regard, the hippocampal region might work at different novelty contents belonging, for example, to the same scene.

6. Conclusions

We have treated a family of generative networks. The generative network has four basic layers, the input layer, the layer of the reconstructed input, the layer of bottom-up processed error, and the layer of the internal representation system. Two properties of the system should be mentioned in the first place:

(i) Bottom-up information transfer is optimized.
(ii) Bottom-up information is denoized.

We have demonstrated by means of computer experiments on sound and image databases, that these properties lead to important features. One feature is related to neuronal signals and can be given the interpretation of repetition suppression and repetition enhancement on the neuronal level. The same process gives rise to accelerated dynamics and this acceleration can be given the interpretation of perceptual priming on the cognitive level. These features support the view that repetition suppression and priming are related to each other.[19] We have noted that within-layer associative structures are not included into the model and that these structures may also play a role in implicit memory.

Acknowledgments

This work was partially supported by a subproject to HFSP grant RG0035/2000–B and by the Hungarian National Science Foundation (Grant No. OTKA 32487).

References

1. S. L. Amari, A. Cichocki and H. H. Yang 1996, "A new learning algorithm for blind signal separation," *Advances in Neural Information Processing Systems*, eds. D. Touretzky, M. Mozer and M. Hasselmo (Morgan Kaufmann, San Mateo, CA), 757–763.
2. P. Aszalós, Sz. Kéri, Gy. Kovács, Gy. Benedek, Z. Janka and A. Lörincz 1999, "Generative network explains category formation in Alzheimer patients," *Proceedings of IJCNN*, IEEE Catalog Number: 99CH36339C.
3. R. Baddeley 1996, "An efficient code in V1?" *Nature* **381**, 560–561.
4. P. A. Bandettini and L. G. Ungerleider 2001, "From neuron to BOLD: New connections," *Nature Neuroscience* **4**, 864–866.
5. M. Bar and I. Biederman 1998, "Subliminal visual priming," *Psychological Science* **9**, 464–469.
6. A. J. Bell and T. J. Sejnowski 1995, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation* **7**, 1129–1159.
7. I. Biederman and E. E. Cooper 1991, "Evidence for complete translational and reflectional invariance in visual object priming," *Perception* **20**, 585–593.
8. I. Biederman and E. E. Cooper 1991, "Priming contour-deleted images: Evidence for intermediate representations in visual object recognition," *Cognitive Psychology* **23**, 393–419.
9. I. Biederman and E. E. Cooper 1992, "Size invariance in visual object priming," *Journal of Experimental Psychology: Human Perception and Performance* **18**, 121–133.
10. I. Biederman and P. Kalocsai 1997, "Neurocomputational bases of object and face recognition," *Philosophical Transactions of the Royal Society London: Biological Sciences* **352**, 1203–1219.
11. J. Brown 1958, "Some tests of the decay theory of immediate memory," *Quaterly J. Exp. Psychology* **10**, 12–21.
12. M. W. Brown and J. P. Aggleton 2001, "Recognition memory: What are the roles of the perirhinal cortes and hippocampus?" *Nature Reviews Neuroscience* **2**, 1–11.
13. R. L. Buckner, J. Goodman, M. Burock, M. Rotte, W. Koutstaal, D. Schacter, B. Rosen and A. M. Dale 1998, "Functional-anatomic correlates of object priming in humans revealed by rapid presentation event-related fmri," *Neuron* **20**, 285–296.
14. G. Buzsáki 1989, "A two-stage model of memory trace formation: A role for 'noisy' brain states," *Neuroscience* **31**, 551–570.
15. J. J. Chrobak, A. Lörincz and G. Buzsáki 2000, "Physiological patterns in the hippocampo-entorhinal cortex system," *Hippocampus* **10**, 457–465.
16. P. Comon 1994, "Independent component analysis — A new concept?" *Signal Processing* **36**, 287–314.
17. E. E. Cooper, I. Biederman and J. E. Hummel 1992, "Metric invariance in object recognition: A review and further evidence," *Canadian Journal of Psychology* **46**, 191–214.
18. T. Cover and J. Thomas 1991, *Elements of Information Theory* (John Wiley and Sons, New York, USA).
19. R. Desimone 1996, "Neural mechanisms for visual memory and their role in attention," *Proc. Natl. Acad. Sci.* **93**, 13494–13499.
20. D. Fleischman and J. Gabrieli 1998, "Repetition priming in normal aging and Alzheimer's disease: A review of findings and theories," *Psychology and Aging* **13**, 1, 88–119.
21. J. D. Gabrieli 1998, "Cognitive neuroscience of human memory," *Annual Review of Psychology* **49**, 87–115.
22. I. Gauthier 2000, "The ups and downs of familiarity," *Current Biology* **10**, R753–R756.
23. S. Gotts and D. C. Plaut, "The impact of synaptic depression following brain damage: A connectionist account of 'access' and 'degraded-store' semantic impairments," Manuscript in preparation. http://www.cnbc.cmu.edu/gotts/accdeg/.
24. P. Graf and D. L. Schachter 1985, "Implicit and explicit memory for new associtions in normal subjects and amnesic patients," *J. Exp. Psychology: Learning, Memory and Cognition* **11**, 501–518.
25. J. H. Hateren and D. L. Ruderman 1998, "Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex," *Proc. R. Soc. London* **B265**, 2315–2320.
26. S. Haykin 1999, *Neural Networks: A Comprehensive Foundation* (Prentice Hall, New Jersey, USA).
27. R. Henson, T. Shallice and R. Dolan 2000, "Neuroimaging evidence for dissociable forms of repetition priming," *Science* **287**, 1269–1272.
28. G. E. Hinton and Z. Ghahramani 1997, "Generative models for discovering sparse distributed representations," *Philosophical Transactions of the Royal Society* **B352**, 1177–1190.
29. B. K. P. Horn 1977, "Understanding image intensities," *Artificial Intelligence* **8**, 201–231.
30. H. Hotteling 1933, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology* **24**, 417–441, 498–520.
31. A. Hyvärinen 1999, "Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation," *Neural Computation* **11**, 1739–1768.

32. A. Hyvärinen 1999, "Survey on independent component analysis," *Neural Computing Surveys* **2**, 94–128.
33. A. Hyvärinen and E. Oja 1997, "A fast fixed-point algorithm for independent component analysis," *Neural Computation* **9**, 1483–1492.
34. T. W. James, G. K. Humphrey, J. S. Gati, R. S. Menon and M. A. Goodale 2000, "The effects of visual object priming on brain activation before and after recognition," *Current Biology* **10**, 1017–1024.
35. O. Jensen and J. E. Lisman 1996, "Hippocampal CA3 region predicts memory sequences: Accounting for the phase precession of place cells," *Learning and Memory* **3**, 279–287.
36. C. Jutten and J. Herault 1991, "Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture," *Signal Processing* **24**, 1–10.
37. J. Karhunen, E. Oja, L. Wang, R. Vigario and J. Joutsensalo 1997, "A class of neural networks for independent component analysis," *IEEE Trans. on Neural Networks* **8**, 487–504.
38. Sz. Kéri 2001, "Are Alzheimer's disease patients able to learn visual prototypes?" *Neuropsychologia* **39**, 1218–1223.
39. J. E. Lisman 1999, "Relating hippocampal circuitry to function: Recall of memory sequences by reciprocal dentate-CA3 interactions," *Neuron* **22**, 233–242.
40. N. K. Logothetis, J. Pauls and M. Augath, T. Trinath and A. Oeltermann 2001, "Neurophysiological investigation of the basis of the fMRI signal," *Nature* **412**, 150–157.
41. A. Lörincz 1997, "Towards a unified model of cortical computation II: From control architecture to a model of consiousness," *Neural Network World* **7**, 137–152.
42. A. Lörincz 1998, "Forming independent components via temporal locking of reconstruction architectures: A functional model of the hippocampus," *Biological Cybernetics* **79**, 263–275.
43. A. Lörincz and Gy. Buzsáki 1999, "Proceedings of IJCNN, Computational model of the entorhinal-hippocampal region derived from a single principle," IEEE Catalog Number: 99CH36339C, ISBN: 0-7803-5532-6, JCNN2136.PDF, Washington, 9–16.
44. A. Lörincz and Gy. Buzsáki 2000, "Two-phase computational model training long-term memories in the entorhinal–hippocampal region," *The Parahippocampal Region: Implications for Neurological and Psychiatric Dieseases* (Academy of Sciences, New York), *Annals of the New York Academy of Sciences*, eds. H. E. Scharfman, M. P. Witter and R. Schwarz, 911 (New York), 83–111.
45. A. Lörincz, B. Szatmáry, G. Szirtes and B. Takács 2000, "Recognition of novelty made easy: Constraints of channel capacity on generative networks," *Connectionist Models of Learning, Development and Evolution*, NCPW6, ed. R. French (Springer-Verlag, London), 73–82.
46. M. McCloskey and N. J. Cohen 1989, "Catastrophic interference in connectionist networks: The sequential learning problem," *The Psychology of Learning and Motivation*, ed. G. H. Bower, **24** (San Diego, CA, Academic Press, Inc.), 109–164.
47. E. K. Miller and R. Desimone 1994, "Parallel neuronal mechanisms for short-term memory," *Science* **263**, 520–522.
48. B. A. Olshausen 1996, "Learning linear, sparse factorial codes," MIT AI Lab, A. I. Memo, 1580, C.B.C.L. Paper No. 138.
49. B. A. Olshausen and D. J. Field, 1996, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature* **381**, 607–609.
50. L. R. Peterson and M. J. Peterson 1959, "Short-term retention of individual verbal items," *J. Exp. Psychology* **58**, 193–198.
51. R. P. N. Rao and D. H. Ballard 1999, "Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects," *Nature Neuroscience* **2**, 79–87.
52. G. Riedel, J. Micheau, A. G. M. Lam, E. v. L. Roloff, S. J. Martin, H. Bridge, L. de Hoz, B. Poeschel, J. McVulloch and R. G. M. Morris 1999, "Reversible neural inactivation reveals hippocampal participation in several memory processes," *Nature Neuroscience* **2**, 898–905.
53. H. L. Roediger and K. McDermott 1993, "Implicit memory in normal human subjects," *Handbook of Neuropsychology*, eds. F. Boller and J. Grafman, **8** (Elsevier, New York), 63–131.
54. D. L. Schacter, N. M. Alpert, C. R. Savage, S. L. Rauch and M. S. Albert 1996, "Conscious recollection and the human hippocampal formation: Evidence from positron emission topography," *Proc. Natl. Acad. Sci.*, USA, **93**, 321–325.
55. D. L. Schacter, C.-Y. P. Chiu and K. N. Ochnser 1993, "Implicit memory: A selective review," *Annu. Rev. Neurosci* **16**, 159–182.
56. D. L. Schacter, E. Reiman, A. Uecker, M. R. Polster, L. S. Yun and L. A. Cooper 1995, "Brain regions associated with retrieval of structurally coherent visual information," *Nature* **376**, 587–590.
57. W. B. Scoville and B. Milner 1957, "Loss of recent memory after bilateral hippocampal lesions," *J. Neurol. Neurosurg. Psychiatry* **20**, 11–21.
58. J. R. Searle 1992, *The Rediscovery of Mind* (Bradford Books, MIT Press, Cambridge, MA).
59. Sz. Kéri, Gy. Benedek, Z. Janka, P. Aszalós, B. Szatmáry, G. Szirtes and A. Lörincz 2002, "Categories, prototypes and memory systems in Alzheimer's disease," *Trends in Cognitive Science* **6**, 132–136.
60. M. J. Tovee, E. T. Rolls and V. S. Ramachan-

dran 1996, "Rapid visual learning in neurones of the primate temporal visual cortex," *Neuroreport* **7**, 2757–2760.
61. C. J. Vaidya, J. D. E. Gabrieli, M. Verfaellie, D. Fleischman and N. Askari 1998, "Font-specific priming following global amnesia and occipital lobe damage," *Neuropsychology* **12**, 1–10.
62. C. J. Vaidya, J. D. E. Gabrieli, M. M. Keane, L. A. Monti, H. Gutierrez-Rivas and M. M. Zarella 1997, "Evidence for multiple mechanisms of conceptual priming on implicit memory tests," *J. Exp. Psychol. Learn. Mem. Cogn.* **23**, 1324–1343.
63. A. D. Wagner and W. Koutstaal, *Priming, Encyclopedia of the Human Brain*, ed. V. S. Ramachandran (San Diego, Academic Press), in press.
64. G. Wallis 1994, "Neural mechanisms underlying processing in the visual areas of the occipital and temporal lobes," Department of Experimental Psychology, Oxford Univeristy.
65. H. Wan, J. P. Aggleton and M. W. Brown 1999, *J. Neurosci.* **19**, 1142–1148.
66. C. L. Wiggs and A. Martin 1998, "Properties and mechanisms of perceptual priming," *Current Opinion on Neurobiology* **8**, 227–233.
67. www.soundcentral.com.