Brain and Mind **3:** 187–220, 2002.

Ockham's Razor at Work: Modeling of the "Homunculus"

ANDRÁS LÖRINCZ, BARNABÁS PÓCZOS, GÁBOR SZIRTES and BÁLINT TAKÁCS
Department of Information Systems, ELTE, Budapest, Hungary (E-mail: lorincz@inf.elte.hu)

(Received: 20 April 2001; in final form: 25 November 2001)

Abstract. There is a broad consensus about the fundamental role of the hippocampal system (hippocampus and its adjacent areas) in the encoding and retrieval of episodic memories. This paper presents a functional model of this system. Although memory is not a single-unit cognitive function, we took the view that the whole system of the smooth, interrelated memory processes may have a common basis. That is why we follow the Ockham's razor principle and minimize the size or complexity of our model assumption set. The fundamental assumption is the requirement of solving the so called "homunculus fallacy", which addresses the issue of interpreting the input. Generative autoassociators seem to offer a resolution of the paradox. Learning to represent and to recall information, in these generative networks, imply maximization of information transfer, sparse representation and novelty recognition. A connectionist architecture, which integrates these aspects as model constraints, is derived. Numerical studies demonstrate the novelty recognition and noise filtering properties of the architecture. Finally, we conclude that the derived connectionist architecture can be related to the neurobiological substrate.

Key words: functional modeling, generative networks, homunculus fallacy, MMI, recognition

1. Introduction

Although there are considerable anatomical differences between different parts of the six-layered sheet of the cerebral cortex, there are also great similarities (Diamond, 1979). The simplest assumption is that there is a similar function throughout this cortical mantle. It has been suggested that the common, underlying function of the cerebral cortex is *prediction* (see, e.g. Barlow, 1999), which is intimately related to universal coding (Cover, 1974; Rissanen, 1984) (for a review see Merhav and Feder, 1998). In mammals, the hippocampus is thought to be concerned with the encoding of long-term memories. In turn, the hippocampus and possibly its adjacent areas may be partially responsible for the universal coding function of the neocortex. As a consequence any theory for explaining the approximate universality of the neocortical function needs to be related to the properties of the hippocampus.

More and more information is available about the learning and coding properties of the hippocampus and the adjacent medial temporal lobe structures (such as the entorhinal, perirhinal and parahippocampal cortices). Ample clinical and experimental evidence support the view that these areas play a role in conscious recognition and recollection of facts and events, often called "declarative memory", or "episodic memory". One line of such evidence are patients with damaged medial temporal system that showed impaired recollection (Scoville and Milner, 1957; Cohen and Squire, 1980; Schachter, 1987; Cohen and Eichenbaum, 1993; Mishkin and Murray, 1994; Clark and Squire, 1998; Henson *et al*., 1999) (see also Eichenbaum, 2000 and references therein). However, it is not known, which memory function is served by the entorhinal-hippocampal system (O'Keefe and Nadel, 1978; Squire, 1992a, b; Eichenbaum *et al*., 1994; Burgess and O'Keefe, 1996). From a computational standpoint, the memory system may consist of encoding, retrieval, storage and consolidation (Squire, 1992b; Tulving, 1983; Shallice, 1988; McCarthy and Warrington, 1990). It has been demonstrated that the hippocampus is involved in more than one of these memory processes (Riedel *et al*., 1999).

Another aspect of memory is that novelty detection should precede all of the above listed memory functions, otherwise the system would not be able to properly handle the known (learned) memory elements and those that are new to the system. It is known that the encoding of novelty is distributed (e.g. Wan *et al*., 1999), for example according to arrangement or content. In spite of the seemingly simple concept of novelty, many questions arise: How should *novelty* be strictly defined? In which sense is the information novel, or, which part of the information is novel? How can we distinguish between the novel and the familiar? How can novelty be distributed? And finally, how should the system encode novel information? To answer these questions, we need to examine whether it is inevitable to suppose a separate mechanism or functional unit for novelty detection which precedes encoding. With or without a separate system, novelty detection should be as fast as the recognition itself and should be based on previous experience.

In this paper, we shall follow Ockham's razor principle (Tornay, 1938). As the principle reads, out of the set of possible hypotheses the most likely is the simplest one that is still consistent with all observations. Our central hypothesis is that in dealing with information processing models the so called "homunculus fallacy" (Searle, 1992) (described in the next section in detail) should be resolved first. Based on this single argument we show that a consistent model emerges. We shall make use of "auxiliary" assumptions when more than one solution is possible. These auxiliary assumptions do not spoil the Ockham's razor principle. Instead, they reflect the ambiguity of the model; these are the points where the predictive power of our single guiding principle on the homunculus fallacy is not fully restrictive. The reason for insisting on Ockham's razor principle is in the mathematics: In principle, non-linear systems – such as the general information processing system of the brain – can be mimicked equally well by different models. Without direct evidence, we can judge the models only by their feasibility and simplicity or their

explanation/prediction power. In turn, functional models should minimize their assumptions and should be checked by the derived properties. Some of the derived properties will be projected to the neurobiological substrate – those properties are the "predictions" of our resolution to the homunculus fallacy. Some of the derived properties are supported by results of computational neuroscience (CNS). CNS could serve as a "filter" to restrict ambiguities of the model. It is possible that our restrictions on the resolution of the homunculus fallacy will eventually crash on either this CNS filter, or directly on experimental findings. At present, we see no contradiction with either findings in neurobiology or with CNS results.

The paper is organized as follows. In Section 2 we introduce our basic assumptions and derive the corresponding architecture. Section 3 presents computational studies for demonstrative purposes. The discussion in Section 4 intends to highlight the critical points of neocortical functioning from the point of view of connectionist modeling. One such example is positive coding and properties, another one concerns the properties of the synapses. Description of correlations between the biological substrate and the model can be also found in this section. Conclusions are drawn in Section 5. Mathematical statements used to establish the model are given in the Appendix. The interested reader can find the details in the quoted references.

2. Model Description

Supposing our brain deals with representations about the environment and its own state, the problem of interpretation of these representations immediately arises. This rather philosophical issue is best described by the so called "homunculus fallacy" (Searle, 1992; Lörincz, 1997; Lörincz *et al*., 2001b). To put it in a nutshell, the fallacy states that an internal representation is still meaningless unless someone can "read" or interpret it. In turn, however, we have to find a place for this reader and define exactly how it "makes sense" to the internal representations. Since our ultimate goal is to build a connectionist system we rephrase the problem: Representation needs to be interpreted, but in a connectionist system any interpretation must assume an activity pattern as a representation which, in turn, needs to be interpreted and so on. Eventually this thought ends up in an infinite regress.

2.1. FUNDAMENTAL ASSUMPTION

In our view, the fallacy arises by construction, or by the pre-assumption that "making sense" is a procedure performed *on* the representation. The central proposal here is that a model of the interpreter can be derived by turning the fallacy upside down. We suggest that the reader is "making sense" of the representation by top–down (re)generation of the input: The input (and not the representation) "makes sense" if it can be derived from the representation. The representation

of the internal representation is the reconstructed (internally generated) input. "Making sense" is not performed *on* the representation but it is performed *by* the representation. This suggestion offers an escape from the endless regression of the homunculus fallacy.

In what follows we shall develop concepts of "making sense" in a connectionist system and shall bind these concepts to declarative memory. Some important mathematical concepts, e.g. reconstruction and novelty, have been selected for guiding the model build up. Assumptions are formulated as *propositions* and represent the lack of predictive power of our model. On the other hand, derived properties of these propositions can be seen as predictions of the model. We note that the mapping of the individual mathematical functions to the biological substrate has been given in our previous works (see Section 4). The novelty of our work is in the derivation of the mathematical functions and not the mapping of these functions. The derivation simplifies the set of assumptions of our previous works. The model is based on well founded mathematical concepts, such as maximization of information transfer. The organizing principle, which connects these mathematical concepts, has not been articulated before. For example, maximization of information transfer is a derived property of our model.

From now on a layer represents an area or a subfield of the brain. Input to (output from) the layer is the signal reaching the layer (departing from the layer) via its afferents (efferents).

2.2. CONSEQUENCES AND OTHER ASSUMPTIONS

Our starting assumption about the resolution of the fallacy implies that in a connectionist model of the brain the input is validated by a generative process, which originates from the internal representation. Horn (1977) suggested first that the underlying computation of perception is the generation (the reconstruction) of the input: "Vision is inverse graphics". Our model is closely related to such inverse (decoding, generative) networks (Hinton and Sejnowski, 1983; Hinton and Ghahramani, 1997; Roweis and Ghahramani, 1999; Rao and Ballard, 1999) including the dynamical sparse representation network introduced by Olshausen and Field (1996). Sparse coding is inspired by biology and computational efficiency, too (Barlow, 1987; Földiák, 1990; Mozer, 1991; Palm, 1992; Zemel and Hinton, 1994; Földiák and Young, 1995; Li, 1995; Dayan and Zemel, 1995; Hinton and Ghahramani, 1997; Hochreiter and Schmidhuber, 1999). The input is *valid* if the internal representation can reconstruct (generate) it. The generation process takes place in a connectionist architecture. The connection strengths of this architecture determine the result of the reconstruction (the reconstructed input) if the internal representation is given. In turn, knowledge, experiences, or *long-term memory* (LTM) are in these connections strengths: Validation happens via these connections strengths, which needs to be tuned to improve their validating capacity. Generative networks have been shown to satisfy our needs; they are able to generate data

based on the learned statistics of the given input set (see, e.g. Horn, 1977; Hinton and Zemel, 1994; Hinton and Ghahramani, 1997 and references therein).

If "making sense" is reconstruction, then reconstruction error is nonsense. The smaller the reconstruction error the more the internal representation can be trusted. In turn, validation can happen via the measurement of the reconstruction error, that is the comparison of the input and reconstructed input. Chances to reconstruct the input by a randomly chosen internal representation are negligibly small. The internal representation may be improved given the current reconstruction error: The current nonsense part of the input may undergo further analysis in order to make sense of it. One can rephrase the reconstruction principle by saying that "making sense" is a collection of processes, which diminish reconstruction error. The result is a connectionist system with relaxing reconstruction error, i.e. with relaxation dynamics.

If temporal changes of the input are not fully random but up to some extent continuity of the input may be present, then – as we shall argue below – predictive structures *at the level of the internal representation* can promote the reconstruction process. For a non-stationary input, changes of the input may contribute to the reconstruction error. Prediction structure at the level of the internal representation can "foresee" these changes and may decrease the reconstruction error. Taken to the extremes, if the internal representation at a given moment is perfect and if the predictive structure is also perfect then no more comparison between input and reconstructed input (i.e. no more input) is necessary for perfect reconstruction. Temporal changes can be learned and can be approximated by predictive networks. We shall assume that predictive structures are present at the level of the internal representation, alike to the proposal of Rao and Ballard (1997) in their Kalman-filter model of the visual cortex. Roweis and Ghahramani (1999) give a unifying view of the family of generative networks, which may also be equipped with predictive connections.

We start deriving the model from information theoretical aspects. The input-to-internal representation transformation, or mapping will be called "coding". "Decoding" is the reverse mapping. Decoding can not be perfect if information is lost in the coding process. In turn, coding of sensory information should maximize information transfer. This is a consequence of the resolution to the fallacy. We shall elaborate on this consequence after introducing the concept of noise. Coding is called "lossless" if the input can be perfectly recovered from the mapped input. We illustrate coding and decoding by the example of wavelet transformation of images (Mallat, 1998). At the image level, wavelet filtering is non-linear – it does more than change the level of illumination. In pixel space, wavelet filtering is linear and corresponds to matrix multiplication. Elements of this matrix are the parameters of the transformation and wavelet transformation is a *linear* parameterization. We shall restrict ourselves to such linear parameterizations, out of which wavelet transformation is a single example. For linear transformations, coding is performed by a matrix denoted by **P**. In turn, decoding assumes the form of a linear matrix

that will be denoted by **Q**. Coding and decoding are lossless, if **PQ** = **I**, where **I** denotes identity transformation. If equality does not hold, i.e. if learning affects parameterization of coding or decoding or both, then the whole process can be "lossy".

Lossy processing may be desirable in some cases. For example, should we reconstruct the noise, which is present in the input? Or should we "loose" this noise content in our coding process and to have a noise-free internal representation? To answer this question the concept of noise should be defined. In most approximations, noise is introduced to describe the unmodeled part of the problem. This definition is not fully satisfactory in our case, because it involves a vicious circle: if unmodeled parts are not present at the internal representation then the internal representation cannot learn to represent novel information. This leads us to say that noise means "no information" (no structure, no regularity, maximal entropy). For continuous inputs noise can then be characterized by a Gaussian distribution, because this distribution has the maximal entropy. In turn, the removal of noise concerns the removal of Gaussian noise, whereas lossless coding refers to lossless coding of information, i.e. lossless coding of structure.

There is always a trade off between fidelity (reliability of noise-filtering) and efficiency. Efficiency is related to the possible speed of the whole coding-decoding process and is influenced by the capacity of the channels. Channel capacity measures the maximum rate (pulse/s, or bit/s) for the given channel. In most cases channel capacity forms a hard constraint (the increase in the number of channels is costly), so we need other tools to make information transfer efficient. According to Shannon (1948; Cover and Thomas, 1991), optimal coding can be achieved by grouping the atomic units (the symbols) of the input first, and coding these new blocks (instead of coding the symbols themselves).

PROPOSITION 2.1. *Efficient transfer of information between the coding and the decoding subsystems is important and coding in the brain is based on grouping of symbols.*

If optimization of information transfer results in a code whose length is close to the Shannon limit (i.e. codeword length is $l(x) \approx log_k(1/p(x))$, where k is the number of the letters of the alphabet) then the emerging internal representation is sparse (Cover and Thomas, 1991): long codes are infrequent, while short codes correspond to few "ones" in a binary system (or a few large components in a continuous representation). A code is called *instantaneous* (or prefix-free code) if no codeword is a prefix of any other codeword. Under this condition, the end of the codeword is immediately recognizable and, in turn, it can be decoded without reference to future codewords. At this stage we define the term "event" to make easier to describe the functioning of our model and to relate the model to declarative memory.

DEFINITION 2.2. Events *are spatio-temporal patterns that can be described by instantaneous coding.*

For a system that transfers related information in a parallel fashion, and if queuing of information is costly, then instantaneous coding becomes a necessity. If the channel capacity rate is limited in this multi-channel system then the transfer of redundant information is to be minimized: Minimizing the mutual information among these channels becomes desirable. At the same time this minimization *maximizes* the mutual information (MMI) between the input and its coded form (Papoulis, 1984), providing further support for our argument on lossless coding of information. From the aspect of MMI, independent component (IC) analysis (ICA) (Jutten and Herault, 1991; Comon, 1994) (see Appendix)[1] is the solution. IC analysis results in minimized mutual information (MMI) between processing channels. Grouping of symbols in hierarchical coding system may concern both the grouping of parallel channels and the grouping of blocks of subsequent inputs.

We note that ICA can also be related to the above mentioned wavelet analysis. Wavelet analysis is known for its denoising capacity, i.e. its capacity to remove Gaussian noise. This property can be generalized to ICA (Hyvärinen *et al*., 1999). Denoising of the ICA components can be approximated as thresholding of the low amplitude ICA components. That is, upon performing ICA transformation, the transformed components of low amplitude can be considered as noise and are to be diminished. Interestingly, it can be shown that this threshold corresponds to the sparsifying term of the relaxation equation of Olshausen and Field (1996, 1997; Hyvärinen *et al*., 1999).

In forming sparse representation *novelty detection* is an indirect consequence: Transformation of novel inputs typically yields representations with different statistical properties as compared to previously known ones (Parra *et al*., 1995; Schölkopf *et al*., 1999; Roberts, 2000; Lörincz *et al*., 2001). Therefore this property can be also used as a constraint (sparse prior) on the probability distribution of the internal representation amplitudes. Hence, further sparsification of the emerging independent components may serve two goals: fidelity and efficiency can be assured together (Hyvärinen *et al*., 1999).

Up to this point we made no assumption about encoding beyond that transfer of information needs to be maximized and that Gaussian noise needs to be removed. In information processing there are other types of noises beyond Gaussian noise. One particular noise component is the presence or absence of information. If information is missing then the procedure of *filling in* the missing part is called pattern completion in machine vision. Pattern completion tacitly assumes that the

[1] ICA and MMI express the assumption of the algorithm and the result of the algorithm, respectively. These concepts are generalizations of decorrelation, i.e. the removal of second order correlations. Second order correlation may be removed in many different ways. Some of them can minimize higher order correlations, too. This is the goal of ICA algorithms.

input is made of positively correlated components: The presence of a component increases the probability of the presence of another component. For example, if the hair covers part of the eye, the missing part of the eye could be filled in. Such positive components should be found in the code in order to filter out non-Gaussian noise. Another view of pattern completion is that components of spatio-temporal patterns could be missing and can be inferred by experience. Such missing components can be seen as missing sub-events from a set of interrelated events. In our view, hierarchy of events, spatio-temporal regions, pattern completion or conditions of a true statement can be seen as related concepts.

2.3. DERIVATION OF THE ARCHITECTURE

According to our basic assumption, the task of the internal representation is to reconstruct inputs. In order to investigate other immediate consequences, let us define a simple reconstruction network, in which both the coding and decoding processes are linear transformations. Let $x \in \mathbb{R}^n$ denote the input of the reconstruction network. The emerging internal representation vector is ($\mathbf{a} = (a_1, \ldots a_i, \ldots a_m)^T \in \mathbb{R}^m$ where superscript T denotes matrix transposition). The top–down matrix $\mathbf{Q}$ forms the reconstruction vector ($\mathbf{y} = \mathbf{Qa}$). The reconstruction error ($\mathbf{x} - \mathbf{y}$) is processed by the bottom-up matrix $\mathbf{P}$ via the following equation:

$$\dot{a}_i = \eta[\mathbf{P}(\mathbf{x} - \mathbf{Qa})]_i \tag{1}$$

where dot denotes temporal differentiation. Vector $\mathbf{a}$ will also be called the activity (vector or pattern) and its components as activities. The corresponding architecture is shown in Figure 1a.

Lossless coding-decoding in a simple reconstruction network involves that bottom–up ($\mathbf{P}$) and top–down ($\mathbf{Q}$) matrices are of full rank and invert each other $\mathbf{QP} = \mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix. Reconstruction networks exhibit the following properties.

- If the internal representation is of the same dimension as the input and both $\mathbf{P}$ and $\mathbf{Q}$ matrices can be inverted, then the relaxed internal representation is determined *solely* by the input and by matrix $\mathbf{Q}$ ($\mathbf{a} = \mathbf{Q}^{-1}\mathbf{x}$ and $\mathbf{x} = \mathbf{y}$), irrespective of matrix $\mathbf{P}$ (Note that the internal representation converges if matrix $\mathbf{QP}$ is positive definite. See Appendix).
- Reconstruction networks work iteratively. First, the internal representation generates the (expected) reconstructed input. In the second step error between the expected input and the actual input arises. This is then used to correct the values of the internal representation.
- The top–down generative matrix (matrix $\mathbf{Q}$) that determines the relaxed activities can be considered as the *long-term memory* of the system. This is a key function of the system in the resolution of the homunculus fallacy.

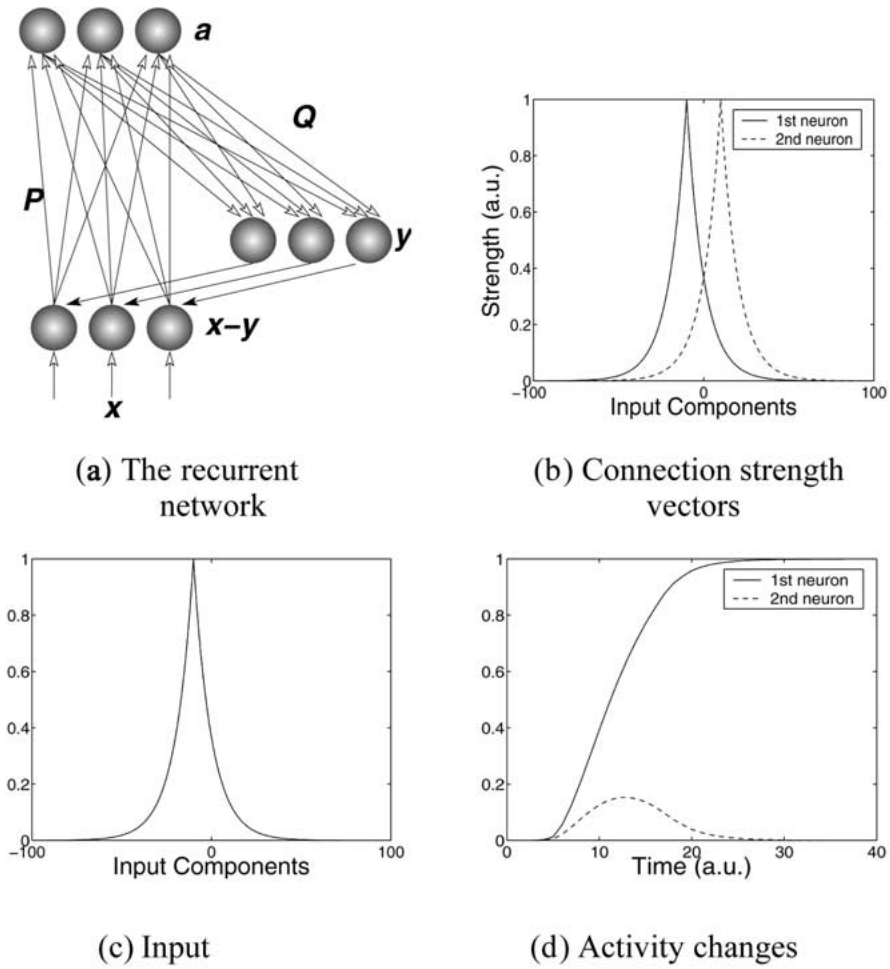

Figure 1. Reconstruction network and reconstruction dynamics.
(a) Reconstruction network. Input, reconstructed input and reconstruction error are computed in a dynamic network. The internal representation (vector **a**) is used to reconstruct the input by means of matrix **Q**. The reconstructed input is subtracted from the input to form the reconstruction error. The reconstruction error is used to correct the internal representation via matrix **P**. Open and solid arrows represent inhibitory and excitatory synapses, respectively. (b) Two connection strength vectors of 201 dimensions. Vector components are connected with a solid line that looks like a smooth function. (c) Let the input be equal with one of the connection strength vectors (memory element). (d) The temporal evolution of the activities of the computational units. The reconstruction principle gives rise to competition between the units. Reconstruction can be slow as shown in the figure. The formation of the internal representation is one-step feedforward process provided that matrix **P** and matrix **Q** invert each other (**PQ** = **I**, where matrix **I** is the identity matrix).

– If the reconstruction network has relaxed to $\mathbf{a}_{old} = \mathbf{Q}^{-1}\mathbf{x}_{old}$ and a new input $\mathbf{x}_{new}$ arrives then the initial activity changes are determined by $\mathbf{P}(\mathbf{x}_{new} - \mathbf{x}_{old})$ irrespective of matrix $\mathbf{Q}$.

Reconstruction networks for brain modeling have been criticized for their slowness (Koch and Poggio, 1999): Human data show that frontal cortex can analyze complex scenes within 150 ms (Thorpe *et al.*, 1996). The speed of this process is in favor of feedforward networks and is a real challenge for iterative schemes (see relaxation for non-inverting coding and decoding in Figure 1d). Fortunately, it is easy to see that speed is not a real drawback for our model: If BU and TD matrices invert each other then the correct internal representation is formed in a feedforward manner; correct internal representation is formed before it would undergo any validation for the $\eta = 1$ case. To see this in a special case, assume that the iteration starts with zeroed internal representation and with zeroed reconstructed input. Then (1) input is provided, (2) error becomes equal to the input, (3) coding gives rise to the internal representation, (4) reconstruction inverts coding, (5) error disappears, which validates the internal representation. Similarly, for non-zero internal representation, error is also eliminated in one iteration. In turn, feedforward networks and generative networks will have the same one-step (bottom–up processing) delay in the forming of the internal representation.

In a generative framework the optimal information transfer should be present in the whole iteration loop: There is no way to assure it only in one part (for example, in BU tuning) of the loop and ignoring this demand in other parts. As it was shown before, $\mathbf{Q}$ is strongly constrained because of its central role in forming the internal representation. However, information optimization can take advantage of the freedom in choosing $\mathbf{P}$. In order to minimize the mutual information between the parallel processing channels, $\mathbf{P}$ should represent the so called *separation matrix* of IC analysis. It has been previously shown that tuning of $\mathbf{P}$ can be local in the reconstruction architecture (Lörincz and Buzsáki, 2000). The tuning can be fast if matrix $\mathbf{P}$ is decomposed into two distinct transformation, a whitening transformation and a separation transformation (For a short explanation about whitening and separation transformations see Appendix). The output of the whitening transformation, the whitened signal, is provided by the whitening layer, whereas the separation transformation works on the whitened signal and emits the separated signal. Such two-step ICA is called "ICA transformation with natural gradient learning" (Amari *et al.*, 1996; Karhunen *et al.*, 1997).

If $\mathbf{P}$ is optimal, $\mathbf{Q}$ needs to be tuned to $\mathbf{P}^{-1}$ for a one step reconstruction. The following quadratic cost function gives rise to the correct delta rule:

$$J = \frac{1}{2}\mathbb{E}\|\mathbf{x} - \mathbf{Q}\mathbf{a}\|^2 \tag{2}$$

where $\mathbb{E}$ denotes the expectation value. Taking the negative gradient of this cost function with respect to Q we have the following upgrade rule:

$$\Delta\mathbf{Q} = \alpha(\mathbf{x} - \mathbf{y})\mathbf{a}^T \tag{3}$$

where T denotes transposition, $\alpha > 0$. This rule can be used as a stochastic gradient upgrade that converges under certain conditions imposed on learning rate α (see, e.g. Haykin, 1999 and references therein). This stochastic gradient method minimizes the mean square error of the reconstruction (Karhunen *et al.*, 1995). It can be seen that this update rule contains both the reconstruction error and the internal representation and, in turn, it is Hebbian. After learning, the average reconstruction error is 0.

In our reconstruction network the input to the whitening stage is the input to the network minus the reconstructed input, i.e. the input is equal to the reconstruction error. Thus, the output of the separation layer is the bottom–up processed error (BUE), which is used to correct the internal representation. In turn, BUE needs to be integrated. We assume that this temporal integration occurs at a separate layer, which will be called as the IRS (internal representation of the system) layer. This choice is compelling because of the need to remove Gaussian noise *and* the need to remove non-Gaussian noise (i.e. pattern completion). We shall return to this point later. The architecture and its working are depicted in Figure 2: The separation layer is called the layer of the bottom–up processed error (BUE), or BUE layer.

Temporal integration can be executed by a simple recurrent network: Inputs of each neurons are made of two components; the output of the respective components of the BUE layer and the output of the IRS neuron itself. In turn, the internal representation and its correction will be added and temporal integration occurs. Such recurrencies are called "recurrent collaterals" and can serve other purposes beyond temporal integration. Recurrencies can be directed to other neurons of the same layer and may indicate "who can be the next". Recurrent collaterals will be considered as the "internal predictive model" of the architecture. Moreover, the same recurrent collaterals may fix the problem of missing components: If some continuity is present in the input flow then indications for "who can be next" and indications for "who should be present at this moment" overlap.

PROPOSITION 2.3. *We assume that the internal model makes use of positive coding for the requirement of the removal of the non-Gaussian noise and pattern completion for missing components.*

Positive coding means that representation makes use of positive values exclusively and that only positive activities can be passed to subsequent layers (Charles and Fyfe, 1998). Activities of internal representation can spread quickly along the directed associative structure (especially in the absence of error correcting bottom–up processing) and can grow quickly because of the self-references. Therefore, these activities need to be limited:

PROPOSITION 2.4. *Normalization for average activity occurs at the internal representation at each time instant.*

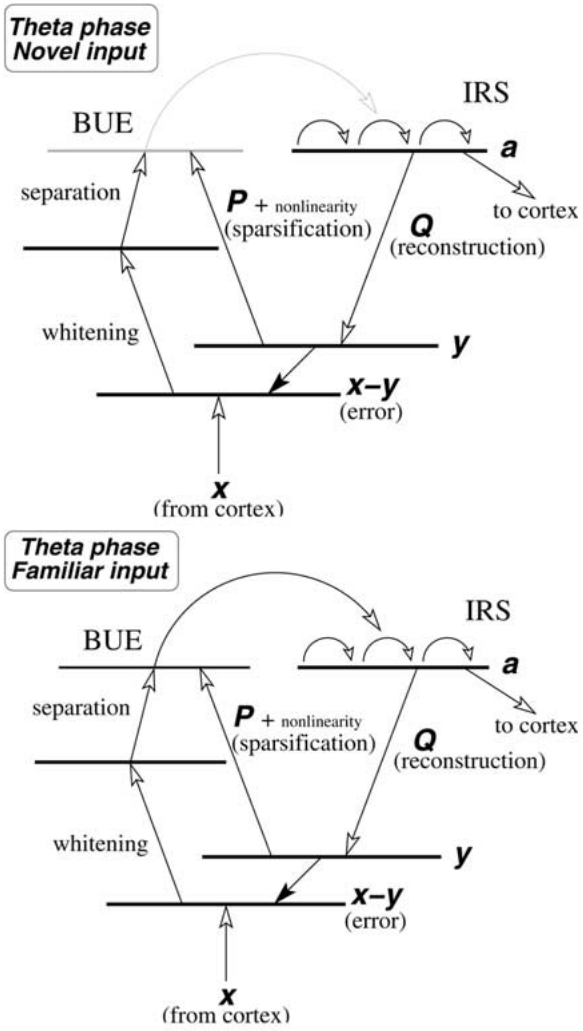

Figure 2. Operation during the non-linear *theta* phase.
Derived connectionist architecture. Theta phase refers to the corresponding neurobiological phase. Bottom–up processing: whitening and separating transformation. Bottom–up input: reconstruction error. Top–down processing: reconstruction using the internal representation. Notation: **x**: input, **y**: reconstructed input, **a**: internal representation, **Q**: top–down generative matrix, **P** + nonlinearity: bottom–up sparsifying transformation (See Equation 4 or Appendix Equation 10), open arrows: excitatory connections, black arrows: inhibitory connections. *Top: Novel input.* Sparsification gates the output of the bottom–up error processing (BUE) layer, activities of the layer of the internal representation system (IRS) are not corrected, bottom–up error correcting transformation (whitening and separation that together form matrix **P**) is optimized. The recurrent collateral system of the whitening layer (not shown, it is not effective during this phase) is being encoded. *Bottom: Familiar input.* Same as above with optimized (sparse) BUE output. Sparsification does not stop BUE output but performs denoising. The model gives rise to an associative structure at the whitening layer. This associative structure is not shown in this figure (See text for more details).

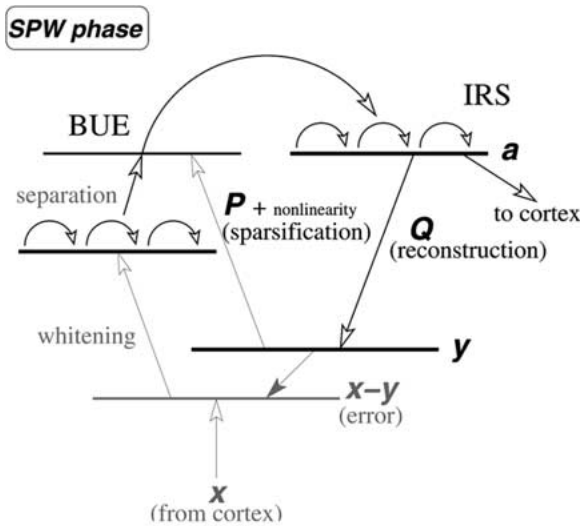

Figure 3. Operation during the linear sharp wave *SPW* phase.
Derived connectionist architecture. SPW phase refers to the corresponding neurobiological phase. For notations see Figure 2. One difference between this figure and Figure 2 is the role of the associative structure at the whitening layer. This associative structure is inactive and *learns* temporal order during the theta phase. The same associative structure is active and replays sequences during the SPW phase. Encoding of the top–down long–term memory matrix **Q** occurs during this phase. Sparsification of synaptic connections of the separating matrix may occur during this phase.

Removal of Gaussian noise, on the other hand, occurs via sparsification as determined by the Olshausen–Field equation (Olshausen and Field, 1996; Hyvärinen *et al.*, 1999):

$$\dot{\mathbf{h}} = \mathbf{P}(\mathbf{x} - \mathbf{y}) + S(\mathbf{h}) \tag{4}$$

where h denotes the MMI vector, $S(.)$ is a non-linear penalty function acting on the components of the argument separately. It then follows that non-linearly processed activities of the MMI components ($S(\mathbf{h})$) direct the sparsification of components of the activities at the BUE layer ($\dot{\mathbf{h}}$). We have two choices for sparsification: (i) straight feedback from the IRS layer to the BUE layer and (ii) indirect feedback from a transformed form of the IRS activities to the BUE layer. Recall that activities at the the IRS layer (components of vector **a**) perform positive coding. In turn, activity components of vector **h** are not represented directly in the network and option (i) requires learning the inverse of the **h** to **a** transformation. Considering case (ii) the only option is the reconstructed input **y**, which is indeed, a transformed form of **h** by construction. Transformation of **y** to **h** requires the ICA matrix belonging to internal representation **a**.

PROPOSITION 2.5. *Let there be a connection system between the reconstructed input and the BUE layer. The matrix of these connections is equal to the joined two*

step transformation of whitening and separation. This processing channel between the reconstructed input and the BUE layer will be referred to as "the sparsifying system".

Now let us assume that the input is novel. By novel we mean that the input, or part of the input has not been encountered before and that the new part entails some structure. The structural part of the input should be transferred, whereas the noise content of the input should be rejected. Because the input is novel, it is not optimized for information transfer and the activity vector at the BUE layer will not be sparse. To go further we need to make another assumption:

PROPOSITION 2.6. *Normalization (for the sum of the absolute values of components) of activities occur at the BUE layer.*

In turn, non-sparse activities will be small and will be strongly affected by denoising. Thus, the output of the BUE layer will be blocked by sparsification. If coding is positive, if recurrent collaterals of the IRS layer spread these positive activities, if the average activity in the IRS layer is given, and if no error correction takes place in the absence of BUE layer outputs, then activity distribution will not be sparse at the IRS layer either and the normalized activities of the IRS layer will be small. Contribution from top–down processing will be reminiscent to the *average* input. Apart from this small contribution, the reconstruction error becomes equal to the input. This is optimal for fast maximization of information transfer in bottom–up processing (Karhunen *et al.*, 1997; Hyvärinen, 1999).

The non-linearity of the BUE layer modifies (tunes) the two-layer processing between reconstruction error and the BUE layer. Bottom–up processed reconstruction error (now, almost equal to the input) is getting sparse. When it becomes sparse, sparse activities gradually overcome the sparsification effect and error correction begins. Thus, optimization of whitening and separation occurs mostly in the absence of reconstruction and, in turn, these transformations are trained according to the statistics of the input(s). Optimization of BU matrices may need further tuning that we shall discuss shortly. If upon tuning the product of the three matrices **QRP** (where **R** represents the BUE to IRS transformation (not shown)) is not positive definite that could, in principle, make Equation 4 divergent. However, the assumed normalizations at the BUE and the IRS layers constrains the differential Equation 1 and stability is ensured. One explanatory example is given in Figure 4. When the loop operates, training of matrix **Q** may be carried on. Hebbian training of matrix **Q** diminishes reconstruction error. In turn, training of matrix **Q** stops when **QRP** is equal to the identity matrix.

Recall, that the sparsifying system works on the ICA representation of the IRS activities (See the last part of Equation 4). Thus, when the whitening and separation BU transformation path (BU transformation, for short) is tuned the following parts of the system are "available": (a) the not-yet-tuned BUE to IRS transformation,

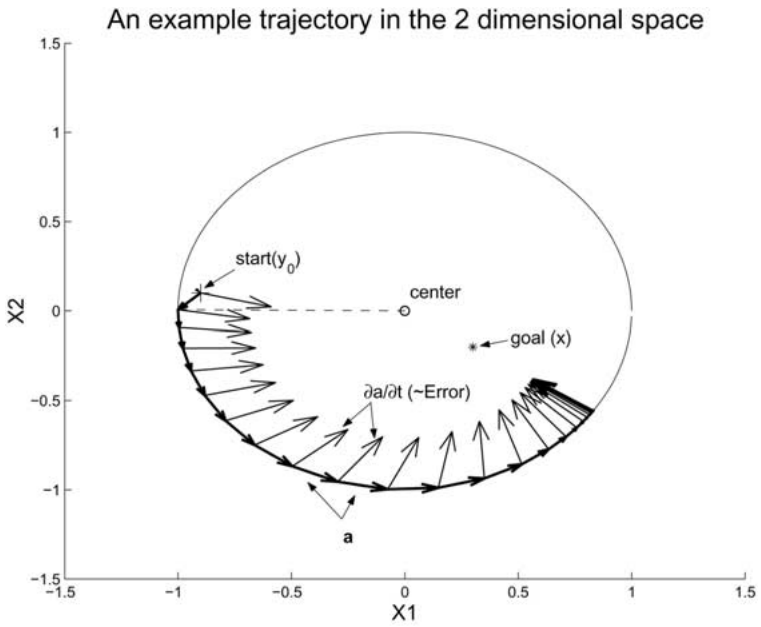

Figure 4. Illustration of the effect of normalization on reconstruction dynamics.
As a two dimensional example, one trajectory of the temporal evolution of Equation 1 can be seen. The task of the system is to reconstruct the original input (**x**, with $|\mathbf{x}| \neq 1$). The simplest case is when **P**, **R** and **Q** are identity matrices, because the internal representation (**a**) of the original input and the reconstructed input (**y**) are exactly the same ($\mathbf{y} = \mathbf{Qa} = \mathbf{Ia} = \mathbf{a}$). The relaxation starts from a random position ($\mathbf{y}_0$ with $|\mathbf{y}_0| \neq 1$). According to the equation, the dynamics is relaxed if the difference between the original input and the reconstructed input has disappeared. However, if the process is subject to normalization, then the difference is never zeroed out, except when the goal is also on the unit circle. The normalization takes place at two levels. First, the internal representations are always re-projected onto the unit circle. Second, the changes of the internal representations ($\frac{\delta \mathbf{a}}{\delta t} \propto$ reconstruction error) are also normalized in each step. The general case is more complex. The properties of matrix **PRQ** determine the equation's behavior. If the matrix is neither positive (semi)-definite nor negative (semi)-definite (that is it may have eigenvalues of different signs), then normalization of the bottom–up error may also be required.

(b) the not-yet-tuned LTM (**Q**) and (c) the sparsifying system. The later one is based on the ICA representation of the IRS activities, which, however, is formed by the not-yet-tuned LTM. In turn, the sparsifying system (in which the information flow is bottom up as well) and the BU transformation may differ. What will happen to the sparsifying system? Natural order of learning is that the optimized BU transformation can overcome the sparsification system and modification of the sparsifying system is not necessary as long as the LTM is not changed: The sparsifying system should somehow refer to the LTM. If LTM changes then the sparsifying system would follow it. The problem that immediately arises is the

learning rule of sparsification matrix. The intriguing outcome of our architecture is that the learning rule for this rather particular matrix could be Hebbian.

Let us investigate the case when BU transformation is "new", but the BUE to IRS transformation and the LTM are "old". What rules will properly train the sparsifying system? Given the loop structure, Hebbian training may consist of two parts. The input to this sparsifying transformation and its non-linearly processed output (which sparsifies the BUE activities) form one part. The input could be positive (for positive coding), whereas the output should have zero average (a necessary condition for ICA). The other part of learning can be provided by accidental coincidences between output and input via travelling through various pathways of the rest of the loop, i.e. the BUE – to – IRS transformation and the TD transformation. Note that synaptic learning is open for about 50 ms, which is circa twice as long as the "signal round trip time" in the HC-EC loop. The stronger these various pathways from output to input, the more effective this part of the learning may be. Note, that both the BUE – to – IRS and the TD transformations correspond to the "old" information. In turn, the various pathways can be represented by a single term determined by the loop structure. This single term is the corresponding (transposed) component of the inverse of the "old" BU matrix. Thus the learning rule can be written as

$$\Delta P \propto [P_{old}^T]^{-1} + \mathbf{y} f^T(\mathbf{P}\mathbf{y}) \tag{5}$$

where f denotes a non-linear function, which provides output with zero mean. This learning equation includes the original ICA learning algorithm suggested by Bell and Sejnowski (1995). It has favorable properties in our architecture: it is slower than the "natural gradient" and is not influenced by the new BU transformation. The learning rule of Equation 5 can keep a sparsifying system, which belongs to the (old) LTM.

Summing up the above described elements a natural ordering can be outlined in optimizing the information transfer of novel inputs in generative networks:

1. optimize BU transformation (whitening and separation), while preventing the modification of $\mathbf{Q}$, that is the decoding system.
2. optimize decoding using the delta-rule.
3. ensure that novel information cannot be mixed with noise, which could spoil the LTM (i.e. tune the sparsifying system only if the LTM is tuned).

The assumption of instantaneous coding, i.e. an *event based coding*, makes possible to fully describe the internal model by its actual state and reference to past is not needed. This assumption in a hierarchical system can be interpreted as a *self-consistent approximation*: If the assumption is false then the computational power (i.e. computational units and connections between them) is not enough to develop independent coding. Coding is imperfect and higher levels are needed to correct that. The dependence at the actual level could be represented by the connections of the IRS layer. This can be the second role of the recurrent collateral system: Not yet encoded information can be reconstructed for a short time (i.e. in a transient

fashion) during relaxation. The two functions, i.e. pattern completion and transient reconstruction of not-yet-encoded information might be related somehow.

It may be worth noting that efficient tuning of the BU transformation requires a linear and a non-linear phase (Karhunen *et al.*, 1997). For novel inputs non-linearity is guaranteed by the sparsifying system. On the other hand, a separate linear phase is required for tuning the LTM matrix **Q**. The linear phase is a replay phase dealing with learned input sequences. Such sequence learning in CA3 recurrent collaterals was first suggested by Levy (for a recent review, see Levy, 1996). Learning is governed by Hebbian learning between the reconstructed error (available during the theta phase) and the rehearsed linearly processed IRS activities available during the SPW phase. This two-phase learning may correspond to the E- and L-LTP processes (Squire and Kandel, 1999). The full architecture and its two-phase operation modes for novel and for familiar inputs are shown in Figure 2 and Figure 3.

3. Qualitative Demonstrations

Our computer simulations serve demonstrative purposes. In these demonstrations independent component analysis has been applied on temporally concatenated inputs. That is why the resulting components are called temporal independent components (TICs), while the extended version of ICA is the TIC analysis (TICA). TICA is known in the literature. Hateren and Ruderman (1998) have shown that filters produced by TICA are in better accord with receptive fields of the primary visual cortex than those of ICA when natural image sequences (instead of individual images) are considered. It has been shown that all three major types of receptive fields of the primary visual cortex emerge in TICA provided that properties of LGN signals are taken into account (Lörincz *et al.*, 2001; Szatmáry and Lörincz, 2001). We shall present the computational simulations on different examples. Our goal is to demonstrate the generality of the method.

In the first computer study facial expressions were examined (Figure 5). (Database was collected in collaboration with Dr. Lajos Simon, Department of Psychiatry, Semmelweis Medical School, Budapest.) The database consists of 200 temporal sequences (movies) of 10–20 s duration displaying the six major facial expressions from about 40 subjects. Faces were rescaled to identical eye distances and eye brows to mouth distances. Hair and possible beard areas were filtered out. Also, each movie frame underwent PCA compression (see, e.g. Haykin, 1999 and references therein). The entire database was used for calculating the PCA covariance matrix. PCA compression lowered the data dimension by a factor of 5. Networks for each facial expression category were trained separately. Three subsequent PCA compressed frames were concatenated for ICA training. The subsequent frames spanned 60 ms. Tests were conducted on unseen movies including the movies from the "same category" (i.e. the TICA basis belonging to "anger" was used for novel movies on "anger"). This case is called *familiar*. Test were conducted on (unseen) movies from "different category" (i.e. the TICA basis belonging to

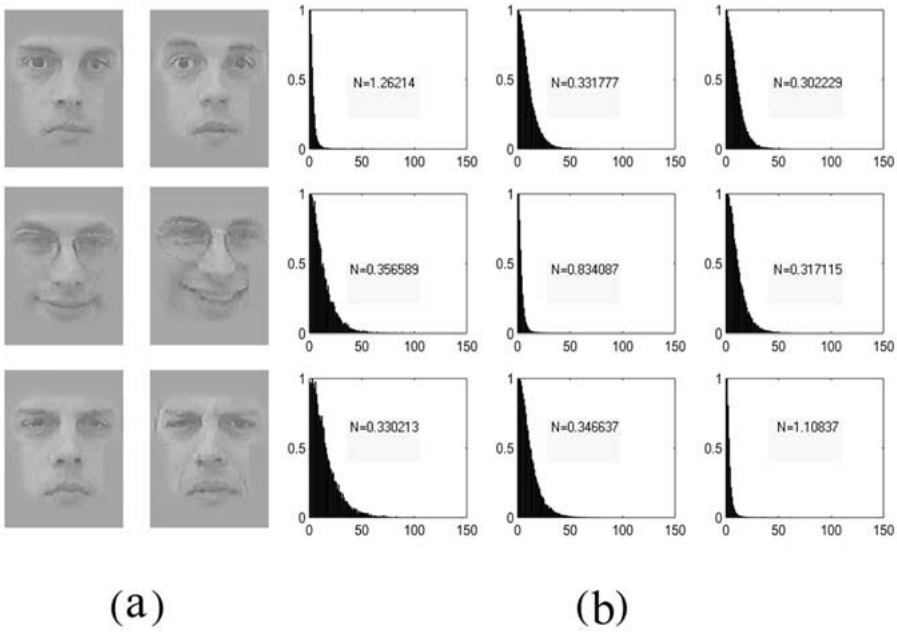

Figure 5. Facial expressions.
(a) Sample frames from "movies" on surprise, smile, and anger. (b) Database was cut into training and into test movies. Training was performed for the three facial expressions separately. In the test phase, test (not yet seen) movies were used. Test movies were from the category of the same facial expression or from a different category. Distribution of TICA outputs in the test movies are shown. Diagonal: test on movies from the same category (familiar inputs), Off-diagonal: test on movies from different facial expression categories (novel inputs). Negentropy values are shown within the subfigures.

"anger" was used for the movies on "surprise"). This case is called *novel*. Activity distributions are computed by normalizing the activities into the same region for each tests. We have found that activity distributions for within-category test movies were much narrower. Quantitative results about the form of the distributions are given by the negentropy values within each subfigure.

In another experiment four acoustic recordings (music, animal sounds, forest sounds, etc.) of a few seconds were cut by half and mixed. The first halves were used for training whereas the second halves were used for testing. These part of the testing formed the familiar set. The novel set was made of similar and identically mixed sounds. Similar results were found in all examples.

The result of sparsification is shown for facial expressions. There are drastic changes on novel inputs (Figure 7); outputs are strongly diminished. Novel input do not pass the BUE layer and cannot be reconstructed. In turn, novel input – from the standpoint of hierarchical processing – appears as *reconstruction error* and will be again processed in this or in higher areas.

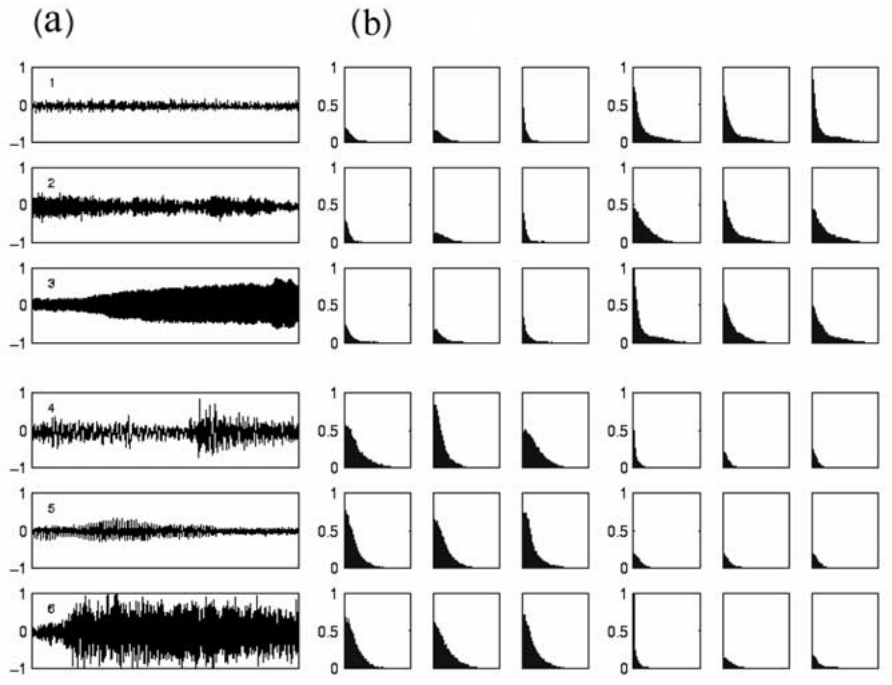

Figure 6. The case of acoustic signals.
(a) Samples are about 250 ms in length and are from different sources (e.g. music, sounds in a forest or sounds of a whale). The first three samples belong to the first mixture, while the others belong to the other mixture. (b) TIC output distributions. The training set for the TICA matrix was created from a mix of three samples out of the six signals. Mixing of the other three samples made the "novel" inputs. Embedding depth is 16. The number of TICs is thus 3 x 16 = 48. Here – in contrast to Figure 5 – the histograms of randomly selected individual outputs (9 out of the 48) are shown. Diagonal (off-diagonal) blocks of histograms represent familiar (novel) tests.

4. Discussion

We voted for generative networks as opposed to feedforward schemes to save the "homunculus" without falling into the trap of the "homunculus fallacy". Ordinary generative networks, however, are poor candidates for "making sense" of the internal representation. This can be seen by the following argument: Assume that the internal representation and the input are of the same dimension. Assume further that bottom-up and top-down processing are linear and are of full rank. Then, a generative network is able reconstruct every possible input with zero error. In other words, every input "makes sense" for an ordinary generative network. How then could this system "recognize" that the actual input "does not make sense"? The first step of "making sense" is to recognize "familiarity" as opposed to "novelty". Unfamiliar inputs can be novel or may contain only noise. In a hierarchical system, noise at one level may be grouped to form input that contains novel information at

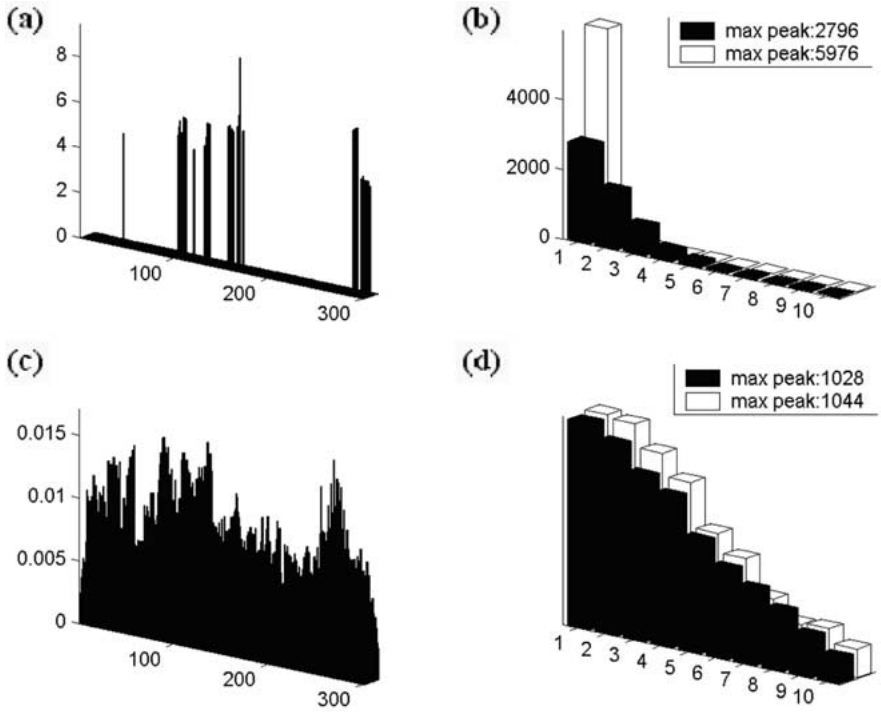

Figure 7. Sparsified activities of facial expressions.
(a) Sparsified internal representation on familiar test movies. Vertical axis: activities, horizontal axis: number of movie frames, in depth axis: number of computational units. Same for subfigure (c). (b) Distributions (histogram) of activities. Black: without sparsification, white: with sparsification. Values of maximal activities are given in the subfigure. Maximal activities belong to the 20th block of the histogram (not shown). (c) Sparsified internal representation on novel test movies. Note the change of the vertical scale. (d) Same as subfigure (b) but for novel test movies.

another level. Optimization of information transfer requires that frequent symbols be represented by short codes (low levels in the hierarchy), whereas less frequent symbols be represented by longer codes (higher levels in the hierarchy). ICA and TICA are optimal to distinguish noise and novelty. Approximate denoising of IC outputs takes place by diminishing small amplitude outputs (Hyvärinen *et al.*, 1999). In turn, noise filtering is local in IC coordinates. Reconstruction is performed by high activity IC outputs. Nothing is lost if information is filtered out. This information will appear as reconstruction error and can be encoded at higher levels. In other words, reconstruction error appears because of the denoising process. This reconstruction error can be analyzed by higher levels, as suggested in Rao and Ballard (1999). Our simulations imply that the recognition of novelty is "easy" and can be instantaneous in optimized generative networks. We found that a novel (not yet seen, not yet optimized) input gives rise to distinctly different statistical properties at the level of internal representation.

It is also an important property of our model that it works on temporally concatenated input series. Computational considerations, such as entropy minimization, factor analysis, or sparse coding lead to representational schemes that optimize information transfer (Attneave, 1954; Barlow, 1961; Dong and Atick, 1995; Field, 1987). Interestingly, neurobiology also supports the idea of doing ICA on temporal information: Temporal response of LGN cells (Wimbauer *et al*., 1997) are such that TICA processing is possible in the primary visual cortex (Szatmáry and Lörincz, 2001).

In this paper, we put forth the idea that long-term memory is represented by the connection strengths of the top-down processing. In turn, channel capacity constraint forms the underlying reason for maximizing information transfer and that independent component analysis of (spatial, temporal, modularity-wise) sequences of symbols need to be considered.

Furthermore, the concept that the ICA process is performed in a generative network allows for "deriving" several important details of the complex hippocampal–entorhinal region (Lörincz, 1998; Lörincz and Buzsáki, 2000). Processing of temporal sequences in the hippocampus, which is the focal point for the learning of long-term explicit memory, has been suggested (Lisman and Idiart, 1995; Lisman, 1999). Distances on dendritic trees of principal cells measured from the soma in the CA3 and CA1 sub-fields can be interpreted as *time* on the 10–100 ms time scale. The reason is that there are propagation delays along the dendritic trees (Henze *et al*., 1996). According to simulations of CA3 principal cells of the hippocampus, information from remote synapses on the dendritic tree is *not* damped compared to synapses close to the soma (Jaffe and Carnevale, 1999) and thus TICA, indeed, becomes possible in the hippocampus.

The present and previous works (Lörincz, 1998; Lörincz and Buzsáki, 2000; Chrobak *et al*., 2000) differ in the first place in the derivation method. Previous works stated that connectionist constraints of ICA fit the hippocampal–entorhinal loop. Here we intended to minimize assumptions *and* derive the architecture that can learn and generate inputs. As we started with a particular resolution of the homunculus fallacy on "making sense", new considerations beyond maximization of information transfer have emerged. Though the presented architecture is almost the same as it was in our previous work, the theoretical framework underwent some important changes and less assumptions have been made. We have shown that the architecture features: 1, correct ordering of recognition of novelty 2, analyzing novelty 3, optimizing information transfer of novelty and 4, encoding novelty. The minimization of our assumptions gave rise to more stringent constraints and a definite general role (sparsification of BUE activities) have emerged for the entorhinal afferents of the CA1 layer. The role of this area was not defined in the previous work.

Individual histograms on familiar inputs (see Figure 6) are most reminiscent to findings in the prefrontal cortex, the inferior temporal visual cortices (see, e.g. Baddeley *et al*., 1997; Treves *et al*., 1999 and references therein), and the hippo-

campus. Individual histograms sometimes show both broad and sharp parts and best characterized by mixed Gaussian and exponential distribution. Qualitatively similar histograms have been reported in Treves *et al*. (1999). Our computations are thus in good accord with available firing data.

There are two questionable properties of IC analysis, sparse coding and other neuronal models presented to date. The first one is that IC outputs can be either positive or negative. There are different ways to think of this property. According to Olshausen and Field (1996, 1997) – whose model is also subject to this problem – the real signal corresponds to the deviation from the average firing rate. For example, a currently not firing neuron corresponds to a negative signal, whereas if the firing rate is at its maximum then it is a positive signal. Another interpretation can be given by extending ICA and TICA analysis to positive coding networks (Charles and Fyfe, 1998). In this case, outputs of the generative networks are constrained to positive values. No negative output is transferred by the network, but reconstruction is still the main goal of the internal representation. The concept modifies the second order correlations of the representation but may leave other aspects unchanged. This issue has been addressed, for example in Charles and Fyfe (1998).

The other problem is as follows. IC analysis, similar to most networks, require both positive and negative synapses. Negative synapses in reality, however, do not exist and this point could form a major obstacle for all of the modeling ideas on factorization and independent component analysis. Very few efficient exceptions are known, one example is the non-negative matrix factorization (NMF) (Lee and Seung, 1999, 2001). According to Lörincz and Buzsáki (Lörincz and Buzsáki, 2000) synapses with *effective* negative signs may be understood as follows. A pyramidal cell may have excitatory connections with other pyramidal cells in subsequent layers. The same pyramidal cell innervates inhibitory neurons. In the feed-forward regulatory system (Buzsáki, 1984) afferent volleys directly activate the inhibitory neuron that, in turn, reduces the probability of firing of the principal cells. We note that the number of inhibitory neurons is about 10% of principal cells. From the point of view of rate coding, the effect of inhibitory neurons on each pyramidal cell represents a (non-specific) negative input whereas the effect of excitatory afferents represents a positive input. Let us assume that two principal cells (layer 1 cell and layer 2 cell) belonging to subsequent layers are connected by feedforward inhibitory paths. Let us call the axon terminal(s) of these paths on the layer 2 cell as indirect inhibitory terminals. The small number of inhibitory neurons allows "summing up" the synaptic strengths from the cell of layer 1 to the cell of layer 2. If the synaptic strength of (non-specific) indirect inhibitory axon terminals overrides the synaptic strength of the direct excitatory terminals then one may talk about negative synaptic strength, or negative connection strength between these neurons. The tunability of the excitatory connections means that the effective connection strength is also tunable.

The derived model – at this stage – does not provide information about coding in the IRS layer. The coding hypothesis, as it was mentioned before, constrains us to consider that the representation is made of "ones" and "zeros" with a low number of "ones". This case corresponds to coupled Markov models (Brand *et al.*, 1997; Brand, 1999; Rezek *et al.*, 2000). We note that for static inputs an approximation of the interactive activation and competition (IAC) model (McClelland, 1981; McClelland and Rumelhart, 1981, 1982; Cohen and Grossberg, 1983; Grossberg, 1988), can be developed by assuming that subsets of neurons undergo hard competition (alike to Markov nets, and winner–takes–all networks) and there is coupling between such subsets. Increasing the time window of observation, transformed signals over unit time may correspond to rate coding. Then the system may be better characterized as a linear dynamical system (Ghahramani and Hinton, 1996). Less restrictive role can be provided by non-negative matrix factorization and by searching positive correlating components. In this case, another type of noise, missing information could be filled in via the recurrent collateral system of the IRS layer.

Mathematical frameworks that are closest to the conjectured internal coding system include computational theories that aim to minimize coding cost (Zemel and Hinton, 1994; Hinton and Zemel, 1994; Pajunen, 1998). They are based on the principle of minimum description length, minimum transmission length and possibly as a generalization, the complexity approximation principle (Solomonoff, 1964; Wallace and Boulton, 1968; Rissanen, 1978; Vovk and Gammerman, 1999). Baddeley (1996) has raised the point that the origin of sparse coding could be simply savings in energy consumption. Energy consumption, coding efficiency and maximized generalization capability considerations may, indeed, play a role in the coding and decoding strategy of the neocortical hierarchy. Low complexity coding and decoding (Hochreiter and Schmidhuber, 1999) could be one of the candidates. Nevertheless, beyond the complexity of coding and beyond the generalization capabilities, behavioral aspects of learning may also play a key role. Such behavioral effects may effectively blur the underlying coding scheme: Low level perceptual learning (Sáry *et al.*, 1994; Schultz *et al.*, 1999) and the possibly related phenomenon of categorical perception show that learning gives rise to resource allocation to decision surfaces. The enhancement of "between category discrimination" measured in psychophysical discrimination tests (Harnad, 1987) is a clear example of such resource allocation. However, the accompanying decrease of "within-category discrimination" (Harnad, 1987; Livingstone *et al.*, 1998; Csató *et al.*, 2000) involves that resources are finite and sometimes resource reallocation is a better description instead of resource allocation. These experiments indicates that coding is also controlled by behavioral success. That is statistics based coding is only one aspect of the coding strategy developed by evolution.

Our resolution to the homunculus fallacy is definitely not the only one way to answer the basic question of interpretation. One may simply deny the fallacy

by considering the brain as a feedforward input–output system (Dennett, 1991). Another possibility is to consider ART-like architectures (Grossberg and Carpenter, 1993), which resonate when there is match between input and representation. All of them are viable options. There are more than one modeling possibility for non-linear systems. The choice determines the derivable properties. In our case, the proposition on "making-sense" provided a rather straightforward route to order learning steps from the recognition of novelty through the filtering of noise and finally to the encoding of information. ART, on the other hand, deals with correct ordering of learning by introducing external means, such as the vigilance parameter. The major theoretical difference between the two models is how they solve the stability–plasticity. Alas, ART fails to explain the success of linear information maximization principles.

5. Conclusions

We have made an effort to minimize our assumptions concerning the model of the information encoding of the "homunculus" and derive the other properties. Here, the goal was to minimize assumptions *and* derive networks that can reconstruct their inputs *and* have appropriate ordering of recognition of novelty, analyzing novelty, optimizing information transfer of novelty and encoding novelty. The derived architecture has minor differences when compared to the previous architecture (Lörincz and Buzsáki, 2000). Following the Ockham's razor principle resulted in a minimized number of assumptions, which have constrained the model building. The difference is in the assignment of the functional role of the direct connections from the layer of the reconstructed input to the layer of the bottom–up processed error. Here, these connections must serve sparsification purposes.

We have derived the architecture by starting from the concept of "making sense". We have conjectured that auto-associators formed by generative networks offer a solution to shortcut the fallacy accompanying the concept of "making sense". Optimization of generative networks should be concerned with channel capacity constraints and the optimization of information transfer. The derivation of the architecture relies heavily on these properties. The derived architecture incorporates recent ideas on sparse coding and predictive coding. The working of the architecture ensures that "making sense" starts by the recognition of novelty followed by the optimization of the encoding of novelty. "Making sense" is seen as a top–down process, whereas novelty recognition is a bottom–up process in the model.

As we noted, the model can be mapped to the loop formed by the hippocampus and the entorhinal cortex (Lörincz, 1998; Lörincz and Buzsáki, 1999, 2000; Chrobak *et al.*, 2000). This agreement means some kind of success to the model owing to the small set of starting assumptions. Derived functional roles of areas and sub-fields as well as the particular Hebbian learning rules of those areas and

those sub-fields are still falsifying predictions if the architecture model is to be considered as the model of the entorhinal-hippocampal loop.

The resolution to the homunculus fallacy via generative networks is the main single assumption of the model. This single assumption does not constrain fully the derivable properties. When more than one option was possible, the uncertainties were constrained by "propositions". These "propositions" may be resolved by new arguments or could be discarded and replaced by other propositions. This was, indeed, the case when the starting principle was changed here from the maximization of information transfer (Lörincz and Buzsáki, 2000) to the more general principle of generative networks. These propositions represent the flexibility of the model. On the other hand, derived properties can be seen as consequences of the main single assumption of the model. For this reason, we call our approach Ockham's razor modeling.

6. Appendix

6.1. BASIS OF INDEPENDENT COMPONENT ANALYSIS

Independent component analysis is now a well established paradigm of signal processing. The seminal works (Jutten and Herault, 1991; Comon, 1994; Laheld and Cardoso, 1994; Bell and Sejnowski, 1995; Amari, 1996; Karhunen *et al.*, 1997) described different approaches, while the review of Hyvärinen (1999) presents a common framework and gives a detailed description on the most popular algorithms. From a neurobiological aspect, individual neurons can be viewed as individual processing channels. Mutual information between processing channels may be decreased, e.g. by decorrelation. Decorrelation removes second order correlations, which results in zero mutual information only, if no higher order correlations exist. In general, however, these higher order correlations do exist. This is best explained by the simplified example of Figure 8. Assume two processing channels. The output of two processing channels is depicted in a plane. The left sub-panel shows the original distribution of the outputs. For the sake of simplicity, uniform distribution within a rhombus-shaped rectangle is assumed.

The removal of second order correlation is equivalent to the coordinate transformation of principal component analysis (PCA). If the PCA axes undergo normalization, then the coordinate transformation is called whitening. In the presented example, whitening transforms the input into a rhombus and the coordinate axes are directed along the diagonals of this rhombus (middle sub-panel of Figure 8). The removal of second order correlation in this example involves that if one of the components of the input is known by us then we are left uncertain about the sign of the other component, for example. However, if one of the input components takes its maximal possible value, e.g. the vertical coordinate is equal to 1.0 then we "know" that the other component should be exactly zero. This knowledge is not in contradiction with the uncertainty about the sign of the other component,

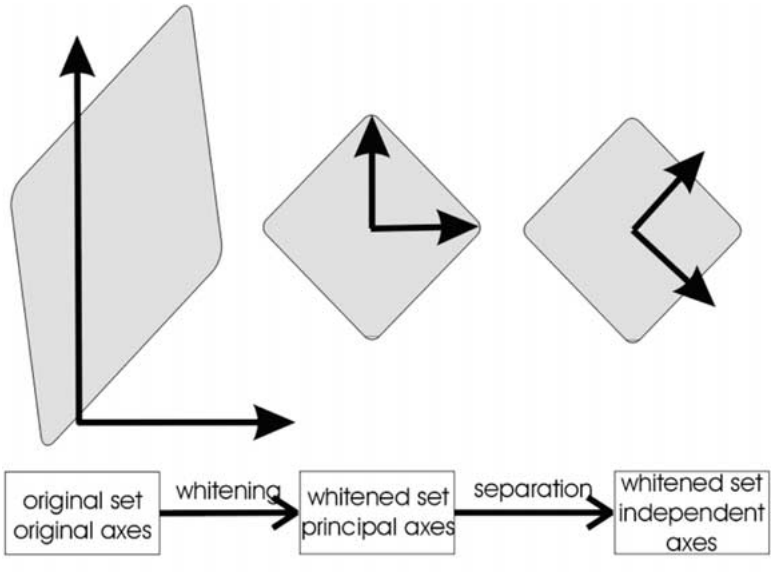

Figure 8. Whitening, separation and independent components.
Left-hand side: Outputs of two processing channels are distributed evenly within a limited space of a rhombus shape. Any output corresponds to a point within the rhombus. The orthogonal projections to the two coordinate axes recovers the output. The input-to-output transformation can be augmented by linear transformations. A linear transformation modifies the length and the direction of the coordinate axes. Middle: Whitening moves, rotates, re-scales the coordinate system, removes second order correlations and provides outputs of equal variance. Right-hand side: Independent components – in this case – are provided by a 45 degree rotation of the coordinate axes.

which, in this particular case, is irrelevant. Thus, higher order correlations persist and mutual information may be lowered. Rotation of the coordinate axes by 45 degrees (right hand side sub-panel of Figure 8) provides a new coordinate system and a new representation of the inputs to the two processing channels. In this new representation, information about one of the coordinates provides no information about the other. In this special case, the minimal possible value of mutual information, that is 0.0, is achieved in the 45-degree rotated coordinate system and thus the components are independent. In general, it is not essential to completely eliminate mutual information. In turn, minimization of mutual information may not lead to independent components. Nevertheless, algorithms minimizing mutual information are called independent component analysis.

ICA can be formalized as a generative model (Bell and Sejnowski, 1995) by assuming an external system made of independent sources subject to a non-linear function and noise, in turn, mixed by a mixing matrix.

$$\mathbf{y} = g(\mathbf{s}) \tag{6}$$

$$\mathbf{x} = \mathbf{M}\mathbf{y} + \xi \tag{7}$$

where (**s**) has n components (in our figure it has 2 components) and denotes the supposed independent sources, $g(.)$ denotes a component-wise non-linearity. **M** is an n by n (in our case a 2 by 2) mixing matrix (unknown for the learning system) and **y** is the observation corrupted by a Gaussian noise source ξ, where ξ has n components. Memoryless noise source is assumed. Values of the sources are "sampled" and are indexed as superscripts. S_1 and S_2 are random variables with probability mass functions p_1 and p_2. The random variables can take values that are denoted by lower-case letters. Mutual information ($I(S_1, S_2)$) of the two variables measures the reduction of uncertainty of one of the components due to the knowledge of the other (Cover and Thomas, 1991). Mathematically

$$I(S_1; S_2) = \sum_{k,l=1}^{K,L} p(s_1^{(k)}, s_2^{(l)}) \log \frac{p(s_1^{(k)}, s_2^{(l)})}{p(s_1^{(k)}) p(s_2^{(l)})} \tag{8}$$

where $p(s_1^{(k)}, s_2^{(l)})$ denotes the joint probability that variable S_1 takes value $s_1^{(k)}$ and that variable S_2 takes the value $s_2^{(l)}$ at the same time. Probabilities $p(s_1^{(k)})$ and $p(s_2^{(l)})$ are defined in a similar fashion without considering the other variable.

The task of ICA is to find an "unmixing" matrix that separates the original (independent) sources. That is, the task of ICA is to invert Equation 6. ICA can be interpreted also as a linear model with non-Gaussian noise. In this case $g(.)$ is simply an identity matrix and noise v is not Gaussian. Depending on the non-linear function it can be sub-Gaussian (narrow distribution with long tails) or super-Gaussian (very broad compared to the tails). For the sake of generality we will talk about IC analysis (learning) that provides outputs with minimized mutual information (MMI).

From a computational point of view it is advantageous to process ICA in two stages (Amari *et al.*, 1996; Karhunen *et al.*, 1997). These two stages are the two steps of Figure 8 depicted in the middle and right hand side sub-figures. The first stage is called whitening: the second order moments are eliminated and the resulting distributions are normalized. The second stage computes the separated coordinate axes. Both stages can be learned by local (Hebbian-like) rules (Cardoso and Laheld, 1996; Hyvärinen and Oja, 1997).

6.2. FASTICA ALGORITHM (HYVÄRINEN AND OJA, 1997)

This algorithm describes a linear transformation for minimizing mutual information between components by maximizing negentropy of the components of the output. Negentropy is defined as:

$$J(S) = H(p_{gauss}) - H(p_S) \tag{9}$$

where $H(p_S)$ is the entropy of the data distribution (i.e. S), $H(p_{gauss})$ is the equivalent entropy of a Gaussian distribution with equal mean and covariance as p_S.

Negentropy measures the "distance" between the data distribution and the Gaussian distribution with the same mean and variance. Note, that minimization of negentropy and the optimization of sparsity, in general, correspond to different ICA transformations.

6.3. SPARSIFICATION AND DENOISING PROCEDURES

Olshausen and Field (1996, 1997; Olshausen, 1996) suggested a non-linear procedure to sparsify redundant representations. This relaxation equation assumes a linear transformation between output and input and contains a sparsification term. The set of equations we used for sparsification are as follows:

$$\dot{a}_i = \eta[\mathbf{P}(\mathbf{x} - \mathbf{P}^{\ddagger}\mathbf{a})]_i - \alpha \frac{a_i}{a_i^2 + c^2} \tag{10}$$

for all output components i, where $\mathbf{P}$ is the learned unmixing matrix, ‡ denotes pseudoinverse, $\mathbf{P}^{\ddagger}$ is the right pseudoinverse of $\mathbf{P}$, $\mathbf{PP}^{\ddagger} = \mathbf{PM} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. $\mathbf{x}$ is the actual input vector, and $\mathbf{a}$, the "internal" (or hidden) representation, which is to be sparsified, $\mathbf{y} = \mathbf{P}^{\ddagger}\mathbf{a}$ is the reconstruction vector. Equation 10 can be derived from the minimization of the following cost function:

$$J_P = \frac{1}{2}(\mathbf{x} - \mathbf{P}^T\mathbf{a})^T(\mathbf{P}^T\mathbf{P})^{-1}(\mathbf{x} - \mathbf{P}^T\mathbf{a}) + f(\mathbf{a}) \tag{11}$$

where $f(.)$ is a component-wise non-linearity. The derivative of function $f(.)$ according to its argument is proportional to $\frac{a_i}{a_i^2+c^2}$. The two parameters α and c can modify sparsification. We used the same values ($\alpha = 0.1$; $c = 0.1$) in every simulation. η was chosen as 0.1. Introducing the term $(\mathbf{P}^T\mathbf{P})^{-1}$ implies the assumption that the separated components have superimposed Gaussian (normal) noise. *This assumption seems more natural* than the similar assumption of Olshausen and Field (Olshausen, 1996) about components of the not–yet–separated inputs. This slight generalization can also be used to illustrate the connection between sparse code shrinkage (Hyvärinen *et al.*, 1999) and the differential equation formulation of sparse coding (Olshausen and Field, 1996). Equation 11 can be used to derive a differential relaxation equation somewhat different from that of (Olshausen and Field, 1996). A basically identical equation served the derivation in sparse code shrinkage consideration (Hyvärinen *et al.*, 1999).

The model of Rao and Ballard (1997) can also be related by including the correlation matrix into Equation 11:

$$J_{P\pm} = \frac{1}{2}(\mathbf{x} - \mathbf{P}^T\mathbf{a})^T(\mathbf{P})^{-1}\Sigma^{-1}(\mathbf{P}^{-1})^T(\mathbf{x} - \mathbf{P}^T\mathbf{a}) + f(\mathbf{a}) \tag{12}$$

Here, Σ denotes the correlation matrix computed between the components of the transformed input. The transformed output concerns the minimization of

mutual information between the transformed components. If the input is a mixture of independent sources then the correlation matrix Σ becomes diagonal upon minimization of mutual information.

Acknowledgments

Enlightening discussions with Irving Biederman, György Buzsáki, and Rufin Vogels are gratefully acknowledged. We are most grateful to John Bickle for calling our attention to recent results B- and L-LTP. Thanks are also due to the unknown referees for their useful comments. This work was partially supported by Hungarian National Science Foundation (Grant No. OTKA 32487). We thank Edgar Körner and to the Future Technology Research Laboratory of Honda for the generous support of a parallel project on the representation of visual information. Movies on city and highway traffic were kindly provided by Honda Future Technology Research, Offenbach, Germany with permission from Werner von Seelen, Ruhr-Universität Bochum, Institut für Theoretische Biologie.

References

Amari, S., Cichocki, A. and Yang, H., 1996: A new learning algorithm for blind signal separation, in *Advances in Neural Information Processing Systems*, Morgan Kaufmann, San Mateo, CA, pp. 757–763.

Attneave, F., 1954: Some informational aspects of visual perception, *Psychological Review* **61**, 183–193.

Baddeley, R., 1996: An efficient code in v1? *Nature* **381**, 560–561.

Baddeley, R., Abbott, L., Booth, M., Sengpiel, F., Freeman, T., Wakeman, E. and Rolls, E., 1997: Responses of neurons in primary and inferior temporal visual cortices to natural scenes, *Proc. Roy. Soc. London* **B264**, 1775–1783.

Barlow, H., 1961: *Sensory Communication*, MIT Press, Cambridge, MA, w.a. rosenblith edition, pp. 217–234.

Barlow, H., 1987: *Learning receptive fields*, volume IV of *Proceedings of the IEEE 1st Annual Conf on Neural Networks*, IEEE Press, U.S.A., pp. 115–121.

Barlow, H., 1999: Cerebral cortex, in C. Koch and J. Davis (eds), *The MIT Encyclopedia of the Cognitive Sciences*, MIT Press, Cambridge, MA, pp. 111–113.

Bell, A. and Sejnowski, T., 1995: An information-maximization approach to blind separation and blind deconvolution, *Neural Computation* **7**, 1129–1159.

Brand, M., 1999: Voice puppetry, in *Proceedings of Siggraph 99*, ACM Press, New York, pp. 21–28.

Brand, M., Oliver, N. and Pentland, A., 1997: Coupled hidden Markov models for complex action recognition, in *Proceedings of IEEE CVPR97*, IEEE Press, pp. 994–999.

Burgess, N. and O'Keefe, J., 1996: Neuronal computation underlying the firing of place cells and their role in navigation, *Hippocampus* **7**, 1–15.

Buzsáki, G., 1984: Feed-forward inhibition in the hippocampal formation, *Prog. Neurobiol.* **22**, 131–153.

Cardoso, J. and Laheld, B., 1996: Equivalent adaptive source separation, *IEEE Trans. on Signal Proc.* **44**, 3017–3030.

Charles, D. and Fyfe, C., 1998: Modelling multiple cause structure using rectification constraints, *Network: Computations in Neural Systems* **9**, 167–182.

Chrobak, J., Lörincz, A. and Buzsáki, G., 2000: Physiological patterns in the hippocampo-entorhinal cortex system, *Hippocampus* **10**, 457–465.

Clark, R. and Squire, L., 1998: Classical conditioning and brain systems: The role of awareness, *Science* **280**, 77–81.

Cohen, M. and Grossberg, S., 1983: Absolute stability of global pattern formation and parallel memory storage by competitive neural networks, *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-13**, 815–826.

Cohen, N. and Eichenbaum, H., 1993: *Memory, Amnesia and the Hippocampal System*, MIT Press, Cambridge, MA.

Cohen, N. and Squire, L., 1980: Preserved learning and retrention of pattern-analyzing skill in amnesia: Dissociation of knowing how and knowing that, *Science* **210**, 207–210.

Comon, P., 1994: Independent component analysis – A new concept? *Signal Processing* **36**, 287–314.

Cover, T., 1974: Universal gambling schemes and the complexity measures of Kolmogorov and Chaitin, Technical Report 12, Department of Statistics, Stanford University, Stanford, CA.

Cover, T. and Thomas, J., 1991: *Elements of Information Theory*, John Wiley and Sons, New York.

Csató, L., Kovács, G., Harnad, S., Pevtzow, R. and Lörincz, A., 2000: Category learning, categorization difficulty, and categorical perception: Computational and behavioral evidence, preprint.

Dayan, P. and Zemel, R., 1995: Competition and multiple cause models, *Neural Computation* **7**, 565–579.

Dennett, D., 1991: *Consciousness Explained*, Little Brown, Boston, MA.

Diamond, I., 1979: The subdivision of the neocortex: A proposal to revise the traditional view of sensory, motor, and association areas, in J. Sprague and A. Epstein (eds), *Progress in Psychobiology and Physiological Psychology*, volume 8, Academic Press, New York, pp. 1–43.

Dong, D. and Atick, J., 1995: Temporal decorrelation: A theory of lagged and nonlagged responses in the lateral geniculate nucleus, *Network* **6**, 159–178.

Eichenbaum, H., 2000: A cortical-hippocampal system for declarative memory, *Nature Reviews, Neuroscience* **1**, 41–50.

Eichenbaum, H., Otto, T. and Cohen, N., 1994: Two functional roles of the hippocampal memory system, *Behavioral and Brain Sciences* **17**, 449–518.

Field, D., 1987: Relations between the statistics of natural images and the response properties of cortical cells, *Journal of the Optical Society of America* **A4**, 2379–2394.

Földiák, P., 1990: Forming sparse representation by local anti-hebbian learning, *Biological Cybernetics* **64**, 165–170.

Földiák, P. and Young, M., 1995: Sparse coding in the primate cortex, in M. Arbib (ed.), *The Handbook of Brain Theory and Neural Networks*, MIT Press, Cambridge, MA, pp. 895–898.

Ghahramani, Z. and Hinton, G., 1996: Parameter estimation for linear dynamical systems, Technical Report CRG-TR-96-2, University of Toronto, Toronto, http://www.gatsby.ucl.ac.uk/zoubin/papers.html.

Grossberg, S., 1988: Competitive learning: From interactive activation to adaptive resonance, in S. Grossberg (ed.), *Neural Networks and Natural Intelligence*, MIT Press, Cambridge, MA.

Grossberg, S. and Carpenter, G., 1993: Normal and amnesic learning, recognition, and memory by a neural model of cortico-hippocampal interactions, *Trends in Neurosciences* **16**, 131–137.

Harnad, S., 1987: *Psychophysical and Cognitive Aspects of Categorical Perception: A Critical Overview*, chapter 1, Cambridge University Press, New York.

Hateren, J. and Ruderman, D., 1998: Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proc. R. Soc. London B* **265**, 2315–2320.

Haykin, S., 1999: *Neural Networks: A Comprehensive Foundation*, Prentice Hall, New Jersey.

Henson, R., Rugg, M., Shallice, T., Josephs, O. and Dolan, R., 1999: Recollection and familiarity in recognition memory: An event-related functional magnetic resonance imaging study, *Journal of Neuroscience* **19**, 3962–3972.

Henze, D., WE, W.C. and Barrionuevo, G., 1996: Dendritic morphology and its effects on the amplitude and rise-time of synaptic signals in hippocampal ca3 pyramidal cells, *J. Comp. Neurology* **369**, 331–344.

Hinton, G. and Ghahramani, Z., 1997: Generative models for discovering sparse distributed representations, *Philosophical Transactions of the Royal Society B* **352**, 1177–1190.

Hinton, G. and Sejnowski, T., 1983: Optimal perceptual inference, in *Proc. of the IEEE Computer Society Conf. on Vision and Pattern Recognition*, IEEE Computer Society, New York, pp. 448–453.

Hinton, G. and Zemel, R., 1994: Autoencoders, minimum description length and Helmholtz free energy, in J. Cowan, G. Tesauro and J. Alspector (eds), *Advances in Neural Processing Systems*, volume 6, Morgan Kaufmann, San Mateo, CA, pp. 3–10.

Hochreiter, S. and Schmidhuber, J., 1999: Lococode performs nonlinear ica without knowing the number of sources, in *Proceedings of the ICA'99*, Aussois, France, pp. 149–154.

Horn, B., 1977: Understanding image intensities, *Artificial Intelligence* **8**, 201–231.

Hyvärinen, A., 1999: Survey on independent component analysis, *Neural Computing Surveys* **2**, 94–128.

Hyvärinen, A., Hoyer, P. and Oja, E., 1999: Sparse code shrinkage: Denoising by nonlinear maximum likelihood estimation, in *Advances in Neural Information Processing Systems 11 (NIPS*98)*, MIT Press, pp. 1739–1768.

Hyvärinen, A. and Oja, E., 1997: A fast fixed-point algorithm for independent component analysis, *Neural Computation* **9**, 1483–1492.

Jaffe, D.B. and Carnevale, N.T., 1999: Passive normalization of synaptic integration influenced by dendritic architecture, *J. Neurophysiol* **82**, 3268–3285.

Jutten, C. and Herault, J., 1991: Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture, *Signal Processing* **24**, 1–10.

Karhunen, J., Oja, E., Wang, L., Vigario, R. and Joutsensalo, J., 1997: A class of neural networks for independent component analysis, *IEEE Trans. on Neural Networks* **8**, 487–504.

Karhunen, J., Wang, L. and Joutsensalo, J., 1995: Neural estimation of basis vectors in independent component analysis, in *Proceedings of the 1995 IEEE International Conference on Neural Networks*, Perth, Australia, pp. 995–1000,

Koch, C. and Poggio, T., 1999: Predicting the visual world: Silence is golden, *Nature Neuroscience* **2**, 9–10.

Laheld, B. and Cardoso, J., 1994: Adaptive source separation with uniform performance, in *Signal Processing VII: Theories and applications. Proceedings of EUSIPCO-94, Edinburgh, UK* (September), volume 2, 183–186.

Lee, D. and Seung, H., 1999: Learning the parts of objects by non-negative matrix factorization, *Nature* **401**, 788–791.

Lee, D. and Seung, H., 2001: Algorithms for non-negative matrix factorization, in *Advances in Neural Processing Systems*, volume 13, Morgan Kaufmann, San Mateo, CA, pp. 556–562.

Levy, W., 1996: A sequence predicting CA3 is a flexible associator that learns and uses context to solve hippocampal-like tasks, *Hippocampus* **6**, 579–590.

Li, Z., 1995: A theory of visual motion coding in the primary visual cortex, *Neural Computation* **7**, 705–730.

Lisman, J., 1999: Relating hippocampal circuitry to function: Recall of memory sequences by reciprocal dentate-ca3 interactions, *Neuron* **22**, 233–242.

Lisman, J. and Idiart, M., 1995: A mechanism for storing 7ą2 short-term memories in oscillatory subcycles, *Science* **267**, 1512–1514.

Livingston, K. andrews, J. and Harnad, S., 1998: Categorical perception effects induced by category learning, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **24**, 732–753.

Lörincz, A., 1997: Towards a unified model of cortical computation II: From control architecture to a model of consiousness, *Neural Network World* **7**, 137–152.

Lörincz, A., 1998: Forming independent components via temporal locking of reconstruction architectures: A functional model of the hippocampus, *Biological Cybernetics* **79**, 263–275.

Lörincz, A. and Buzsáki, G., 1999: Computational model of the entorhinal-hippocampal region derived from a single principle, in *Proceedings of IJCNN* (July 9–16), IEEE Catalog Number: 99CH36339C, ISBN: 0-7803-5532-6, Washington.

Lörincz, A. and Buzsáki, G., 2000: Two–phase computational model training long–term memories in the entorhinal–hippocampal region, in H. Scharfman, M. Witter and R. Schwarz (eds), *The Parahippocampal Region: Implications for Neurological and Psychiatric Dieseases*, volume 911 of *Annals of the New York Academy of Sciences*, New York Academy of Sciences, New York, pp. 83–111.

Lörincz, A., Szatmáry, B. and Kabán, A., 2001a: Sign-changing filters similar to cells in primary visual cortex emerge by independent component analysis of temporally convolved natural image sequences, *Neurocomputing* **38–40**, 1437–1442.

Lörincz, A., Szatmáry, B., Szirtes, G. and Takács, B., 2001b: Recognition of novelty made easy: Constraints of channel capacity on generative networks, in R. French (ed.), *Connectionist Models of Learning, Development and Evolution*, Springer-Verlag, London, pp. 73–82.

Mallat, S., 1998: *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, CA.

McCarthy, R. and Warrington, E., 1990: *Cognitive Neuropsychology*, Academic Press, San Diego.

McClelland, J., 1981: Retrieving general and specific information from stored knowledge of specifics, in *Proceedings of the Third Annual Meeting of the Cognitive Science Society*, pp. 170–172.

McClelland, J. and Rumelhart, D., 1981: An interactive activation model of context effects in letter perception: Part 1 an account of basic findings, *Psychological Review* **88**, 375–407.

McClelland, J. and Rumelhart, D., 1982: An interactive activation model of context effects in letter perception: Part 2 the contextual enhancement effect and some tests and extensions of the model, *Psychological Review* **89**, 60–94.

Merhav, N. and Feder, M., 1998: Universal prediction, *IEEE Trans. Inform. Theory.* **IT-44**, 2124–2147.

Mishkin, M. and Murray, E., 1994: Stimulus recognition, *Current Opinion in Neurobiology* **4**, 200–206.

Mozer, M., 1991: Discovering discrete distributed representations with iterative competitive learning, in R. Lippmann, J. Moody and D. Touretzky (eds), *Advances in Neural Processing Systems*, volume 3, Morgan Kaufmann, San Mateo, CA, pp. 627–634.

O'Keefe, J. and Nadel, L., 1978: *The Hippocampus as a Cognitive Map*, Clarendon Press, Oxford.

Olshausen, B., 1996: Learning linear, sparse factorial codes, A.I. Memo 1580, MIT AI Lab. C.B.C.L. Paper No. 138.

Olshausen, B. and Field, D., 1996: Emergence of simple-cell receptive field properties by learning a sparse code for natural images, *Nature* **381**, 607–609.

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

Pajunen, P., 1998: Blind source separation using algorithmic information theory, in C. Fyfe (ed.), *Proceedings of Independence and Artificial Neural Networks*, ISCS Academic Press, pp. 26–31.

Palm, G., 1992: On the information storage capacity of local learning rules, *Neural Computation* **4**, 703–711.

Papoulis, A., 1984: *Probability, Random Variables and Stochastic Processes*, 2nd edition, McGraww-Hill, New York.

Parra, L., Deco, G. and Miesbach, S., 1995: Statistical independence and novelty detection with information preserving nonlinear maps, *Neural Computation* **8**(2), 260–269.

Rao, R. and Ballard, D., 1997: Dynamic model of visual recognition predicts neural response properties in the visual cortex, *Neural Computation* **9**, 721–763.

Rao, R. and Ballard, D., 1999: Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects, *Nature Neuroscience* **2**, 79–87.

Rezek, I., Sykacek, P. and Roberts, S., 2000: Coupled hidden Markov models for biosignal interaction modelling, Technical report PARG-00-5, Oxford University, Oxford, UK to appear in 2000 IEE Special Issue Proceedings on Advances in Medical Signal and Information Processing.

Riedel, G., Micheau, J., Lam, A., Roloff, E., Martin, S., Bridge, H., Hoz, L., Poeschel, B., McVulloch, J. and Morris, R., 1999: Reversible neural inactivation reveals hippocampal participation in several memory processes, *Nature Neuroscience* **2**, 898–905.

Rissanen, J., 1978: Modeling by shortest data description, *Automatica* **14**, 465–471.

Rissanen, J., 1984: Universal coding, information, prediction and estimation, *IEEE Trans. Inform. Theory* **IT-30**, 629–636.

Roberts, S., 2000: *Novelty Detection Using Extreme Value Statistics*, draft, Oxford University, Oxford, UK, http://www.robots.ox.ac.ukparg.

Roweis, A. and Ghahramani, J., 1999: A unifying review of linear gaussian models, *Neural Computation* **11**, 305–345.

Sáry, G., Vogels, R. and Orban, G., 1994: Orientation discrimination of motion-defined gratings, *Vision Res.* **34**, 1331–1334.

Schachter, D., 1987: Implicit memory: History and current status, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **13**, 501–518.

Schlitz, C., Bodart, J., Dubois, S., Dejardin, S., Michel, C., Roucoux, A., Crommelinck, M. and Orban, G., 1999: Neuronal mechanisms of perceptual learning: Changes in human brain activity with training in orientation discrimination, *NeuroImage* **9**, 46–62.

Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A. and Williamson, R., 1999: Estimating the support of a high-dimensional distribution, Technical Report 99–87, Microsoft Research.

Scoville, W. and Milner, B., 1957: Loss of recent memory after bilateral hippocampal lesions, *Journal of Neurol. Neurosurg. Psychiatry* **20**, 11–21.

Searle, J., 1992: *The Rediscovery of Mind*, Bradford Books, MIT Press, Cambridge, MA.

Shallice, T., 1988: *From Neuropsychiology to Mental Structure*, Cambridge Univ. Press, New York.

Shannon, C., 1948: A mathematical theory of communication, *Bell Sys. Tech. Journal* **27**, 379–423 and 623–656.

Solomonoff, R., 1964: A formal theory of inductive inference, part i, *Information and Control* **7**, 1–22.

Squire, L., 1992a: Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory, *J. Cog. Neurosci.* **4**, 232–243.

Squire, L., 1992b: Memory and the hippocampus: A synthesis of findings with rats, monkeys, and humans, *Psychol. Rev.* **99**, 195–231.

Squire, L. and Kandel, E., 1999: *Memory: From Mind to Molecules*, Scientific American Press, New York.

Szatmáry, B. and Lörincz, A., 2001: Independent component analysis of temporal sequences subject to constraints by lgn inputs yields all the three major cell types of the primary visual cortex, *J. of Comp. Neurosci.* **11**, 241–248.

Thorpe, S., Fize, D. and Marlot, C., 1996: Speed of processing in the human visual system, *Nature* **381**, 520–522.

Tornay, S., 1938: Ockham: Studies and selections.

Treves, A., Panzeri, S., Rolls, E., Booth, M. and Wakeman, E., 1999: Firing rate distributions and efficiency of information transmission of inferior temporal cortex neurons to natural visual stimuli, *Neural Computation* **11**, 601–631.

Tulving, E., 1983: *Elements of Episodic Memory*, Clarendon Press, Oxford.

Vovk, V. and Gammerman, A., 1999: Complexity approximation principle, *Computer Journal* **42**, 318–322.

Wallace, C. and Boulton, D., 1968: An information theoretic measure for classification, *Computer Journal* **11**, 185–194.

Wan, H., Aggleton, J. and Brown, M., 1999: Different contributions of the hippocampus and perirhinal cortex to recognition memory, *J. Neurosci.* **19**, 1142–1148.

Wimbauer, S., Wenish, O., Miller, K. and van Hemmen, J., 1997: Development of spatio-temporal receptive fields of simple cells: I. Model formulation, *Biological Cybernetics* **77**, 456–461.

Zemel, R. and Hinton, G., 1994: Developing population codes by minimizing description length, in J. Cowan, G. Tesauro and J. Alspector (eds), *Advances in Neural Processing Systems*, volume 6, Morgan Kaufmann, San Mateo, CA, pp. 11–18.