Latent Attractor Selection for Variable Length Episodic Context Stimuli with Distractors

Simona Doboli
Computer Science Department
Hofstra University
Hempstead, NY 11549

Ali A. Minai
Complex Adaptive Systems Laboratory
ECECS Department
University of Cincinnati
Cincinnati, OH 45221

Abstract

Latent attractor networks have been proposed as a possible mechanism for representing episodic context in the hippocampus [5], and as general purpose models of episodic context-dependent encoding in neural networks [3]. These are recurrent neural networks with attractors that never fully manifest themselves, but bias the network's response to external stimuli. While each attractor in the original latent attractor model was triggered by unique *context patterns* specific to the context, this model was later extended to the case where contexts were triggered progressively by the sequential presentation of several stimulus patterns without regard to order, simulating the more realistic situation where a context is identified by a sequentially scanned combination of landmarks. In this paper, we describe a network model that can select among contexts identified by overlapping sequences of different lengths, even if the relevant stimulus patterns are interspersed among patterns irrelevant to context selection.

I. INTRODUCTION

In most realistic situations facing a cognitive system, the meaning of stimuli depends on context. It is possible to distinguish between two types of contexts, depending on how far in the past the context information is given *relative to the current time*. The first one — which we call *Type I Context* — comprises the cases where the system's response at time t depends on stimuli presented in an immediately preceding time window. Examples of such context occur in applications such as speech processing and word recognition. It is embodied in autoregressive models or finite-state machines and can be learned by recurrent neural networks where past states are fed back to the network [12, 14, 13].

The second type of context, which we term *Type II* or *episodic context*, arises when the information identifying context is given transiently at *a particular time* — typically at the beginning of an episode. Examples of such context occur in the recognition of spatial environments, social situations and in task planning, etc., where the context identified at the beginning of the episode continues to be in force for its entire duration. For example, upon entering a room, one recognizes it based on the presence of certain objects and/or persons, setting the context for future behavior in that room even if the identifying stimuli disappear subsequently. Episodic context dependence is more difficult for neural networks because the information on context is specified at a fixed time which grows increasingly remote from the present. This requires the system to "latch" information which is difficult to achieve by simple recurrence [1, 13].

The hippocampal region of the brain in rodents appears to construct distinct representations of similar — even identical — environments based on episodic context [21, 22]. We have previously proposed a class of networks called *latent attractor networks* to explain how episodic context-dependent representations can arise in a hippocampus-like system without resorting to off-line or external biasing [19, 5, 3, 6]. Latent attractor networks are recurrent neural networks with competitive firing that embed patterns of activity as attractors using associative Hebbian learning. However, the recurrent connections are not strong enough to sustain the activity patterns autonomously. Each attractor is associated with a specific external stimulus pattern called a *context pattern*. When the context pattern is presented to the network, it disproportionately activates neurons that are supposed to be active in the associated attractor. This then produces a stable bias onto this set through the recurrent connections so that subsequent external inputs — not explicitly associated with any particular attractor — also produce response patterns whose activity lies mainly in the active set of the chosen attractor, thus conditioning the system's responses to stimuli by the original context pattern long after the pattern itself is gone. This situation persists until an external stimulus associated with another context/attractor is presented to the network [5].

In the paradigm described above, the context patterns are unitary stimuli (e.g., looking at the number on a door or seeing a single identifying landmark). However, in realistic situations, context is not set by a single stimulus, but by a conjunction of stimuli (e.g., objects in a room). For each episode with the same context, these stimuli would typically be apprehended in different order depending on

their location and the viewer's actions, and may be interspersed with other irrelevant stimuli, termed *distractors*. Furthermore, each individual stimulus may be part of the identifying combination for multiple contexts; it is the combination as a whole that indicates the specific context. Thus, as the context-setting stimuli are scanned, a unique context identification would only emerge gradually rather than instantaneously, and, until the context is uniquely identified, the system's response may be compatible with several choices.

In previous studies, we have considered the gradual activation of latent attractors by the presentation of a stimulus sequence including distractor stimuli [8, 9]. However, it was implicitly assumed that each stimulus sequence contained the same number of relevant stimuli. This is, of course, a rather artificial assumption, and we now report on a model that removes it. As noted in our earlier work, the problem of gradual convergence in response to temporally presented stimuli is relevant to tasks other than context selection, e.g., representation of hierarchical information structures in neural networks.

Previous studies have considered the gradual activation of attractors by sequences of stimuli [15, 2]. However, in these cases, the order of stimuli is fixed, while we focus on the equally — perhaps more — natural case where the order is explicitly irrelevant. This is closer in spirit to the problem of encoding and recovering hierarchical information structures in modular neural networks [10, 11]. In these networks, higher level categories are represented by the simultaneous activation of lower level concepts represented by activity in different modules. However, these networks learn specially designed hierarchical patterns [10, 11]. We do not use such patterns.

II. PROBLEM DEFINITION

The network is presented with η different external stimulus sequences in discrete time. Each sequence, S^q, of length n_q

$$S^q = C_1^q[R^q]C_2^q[R^q]...[R^q]C_{m_q}^q[R^q] \quad q = 1, 2, ..., \eta \quad (1)$$

begins with a sub-sequence of r_q patterns – *the context sequence* – comprising m_q context patterns, C_i^q, interspersed with varying numbers of non-context patterns, indicated by $[R^q]$. The last $[R^q]$ is a sub-sequence of $n_q - r_q$ non-context stimuli termed the *regular sequence*. The context patterns, C_i^q are drawn from a set of patterns called the *context set*, $C = \{C_k\}$, and the remaining patterns are drawn from the set, $R = \{R_k\}$. Both C_k, $R_k \in I$, where I is the input space of dimension N_I. We use binary stimulus patterns, so $I \equiv \{0,\ 1\}^{N_I}$ For purposes of simulation, patterns in both sets are generated randomly, but mutually exclusive.

A total of ν possible contexts are defined, each context, c^k, specified by a unique set of μ_k context patterns drawn randomly without repetition from C. Each episode sequence, S^q, has a unique context, c^{k_q}, drawn from among the $\{c^k\}$. The context sequence for S^q includes the $m_q = \mu_{k_q}$ context patterns for c^{k_q} in random order mixed with $n_q - m_q$ non-context patterns. The regular sequence patterns for S^q are chosen randomly from R.

At the beginning of an episode, as context patterns are presented, each context pattern identifies the correct attractor with increasing specificity until, at the end of the context sequence, the attractor is uniquely identified. Correspondingly, the network activity should gradually become confined to the active set of the correct latent attractor and remain confined during the presentation of the regular stimulus patterns.

III. METHOD

The core of the system is a latent attractor network composed of two layers: the *response layer*, L_R, and the *intermediary layer*, L_H. The input patterns are projected to the response layer through the *stimulus layer*, L_S. An additional layer called the *biasing layer*, L_B, receives input from the stimulus layer and projects back to the response layer.

The *stimulus layer*, L_S, has N_S neurons that project the input stimuli to the *response layer*, L_R. The connections from L_S to L_R are set randomly with probability p_S of connection. Only K_S neurons in the input layer are active at one time. The *biasing layer* L_B has N_B neurons that receive input from the stimulus layer L_S and project to the response layer. The input connections from L_S to L_B layer are chosen randomly, with a high probability of connection p_{B1}. The output connections from L_B to L_R layer are also chosen randomly with a high probability of connection (p_{B2}).

The response layer, with N_R neurons, also receives a disynaptic recurrent connection through the *intermediary layer*, L_H with N_H neurons. The latent attractors are stored in the recurrent connections between the L_R and L_H layers. There are M attractors, each comprising two binary patterns, one in layer L_R and the other in layer L_H. The patterns have G_R and G_H active neurons, respectively in L_R and L_H, called the *active sets* of the attractor. The connections between L_R and L_H layers are chosen randomly with probability of connections p_R (L_H to L_R) and p_H (L_R to L_H). The attractors are stored in the connections using a clipped binary Hebbian rule first proposed by Willshaw: The connections between neurons active in the two patterns of any attractor are set to high values, while the rest are set to low values [23]. In this way, the M pairs of patterns are set as attractors or fixed points in the in the 2-layer network. The attractors are called latent because they are never fully activated.

The network activity is determined as follows. The excitation to a layer L_R neuron, i, at time t is given by:

$$y_i^R(t) = \sum_{j \in L_S} w_{ij}^{SR} x_j(t) + g_i(t) \sum_{j \in L_H}$$

$$w_{ij}^{HR} z_j(t-1) + g_{bias} \sum_{j \in L_B} w_{ij}^{BR} u_j(t-1) \qquad (2)$$

where $w_{ij}^{(.)}$ denote connection weights, $x_j(t)$ is the jth bit of the external stimulus patterns at time t, $z_j(t)$ is the output of neuron $j \in L_H$, $u_j(t)$ is the output of neuron $j \in L_B$, $g_i(t)$ is the (modifiable) recurrent gain of neuron i, and g_{bias} is the gain from the biasing layer L_B to the response layer L_R.

The synaptic input to a layer L_H neuron, i, is given by:

$$y_i^H(t) = \sum_{j \in L_R} w_{ij}^{RH} v_j(t) \qquad (3)$$

where $v_j(t)$ is the output of $j \in L_R$.

Firing in both L_R and L_H is competitive: The output of the K_R (K_H) most excited neurons in L_R (L_H) at time t is set to 1, while the rest of the neurons output 0. This is a K-winner take all competitive firing rule. The number of neurons allowed to fire - K_R and K_H respectively - are much smaller than the size of the attractors - G_R and G_H.

The input onto a layer L_B neuron i is:

$$y_i^B(t) = \sum_{j \in L_S} w_{ij}^{SB} x_j(t) \qquad (4)$$

where $x_j(t)$ is the output of neuron $j \in L_S$. Neurons in L_B fire when their synaptic input exceeds a threshold level, θ_B.

Latent attractors are associated with stimulus sequences as follows: The connections between L_S and L_R layers are modified such that each context pattern in a sequence excites primarily neurons in the active set of the corresponding attractor in the L_R layer.

The system is faced with the following three problems: (1) the network's state should not become confined to any attractor until it has received enough evidence to uniquely identify the context (i.e. until all context patterns have been presented at the stimulus layer), (2) the activity of the network should not change during the presentation of distractor inputs, and (3) the length of a context (μ_k) should not influence its recognition.

The first problem is addressed by a process called *incremental competitive positive feedback* [8, 9]. The stability of any attractor in the network is controlled by its recurrent gain, g_i, which sets the relative strength of the recurrent excitation versus the external one. For a latent attractor to be stable, neurons in its active set must be above a minimum limit of g_i [6]. When g_i is small compared to the strength of the external stimulus, activity in the network is dictated by the external, feed-forward pathway. Neurons which receive stronger excitation by the stimulus will tend to win the competition for firing among L_R neurons, independent of the distribution of activity in the network. If g_i are large, the recurrent path dominates and the network's activity is determined by the competitive firing between attractors in L_R and L_H.

In our system, all g_i are set to a small value at the beginning of an episode, so that attractors that are associated with the early context patterns are likely to be activated a bit more than others due to feed-forward association. As the presentation of context patterns proceeds, g_i for neurons that belong to the active sets of attractors with more current activity is increased gradually, priming these attractors for possible persistence if reinforced by subsequent context stimuli. Thus, at each stage, activity is distributed among those attractors that are consistent with the context stimuli received thus far. As each new context stimulus is presented, some of these candidate attractors are reinforced further at the expense of others until, finally, only one is left. When the stimulus is not a context pattern, it causes no significant change in the bias for any attractor, and the biasing neurons keep activity distributed among the attractors as at the previous step. The recurrent gain is not allowed to change during presentation of irrelevant inputs. Since these stimuli do not represent any positive or negative reinforcement about the correct context, they should not modify the balance of activity in the network, by allowing the recurrent gain to vary.

The equation governing the modulation of recurrent gain is [9]:

$$\hat{g}_i(t) = g_{min} + \frac{g_{max} - g_{min}}{(1 + e^{-\alpha(a_l(t) - \beta)})}$$

$$d_i(t) = \hat{g}_i(t) - g_i(t-1)$$

$$g_i(t) = \begin{cases} \hat{g}_i(t) & if\ |d_i(t)| < \Delta g_{max} \\ g_i(t-1) + \Delta g_{max} sgn(d_i(t)) & else \end{cases} \qquad (5)$$

where α is a rate of change parameter, β is an offset parameter, l is the index of the attractor for which i is in the active set, $a_l(t)$ is the total number of active neurons in the L_R active set of attractor l at time t, g_{min} and g_{max} are the minimum and maximum possible values of the recurrent gain. Thus, the gain is $\hat{g}_i(t)$, but the absolute change in gain is bounded by Δg_{max} ($\Delta g_{max} > 0$).

The modulation of recurrent gain on individual neurons is motivated by several biological considerations:

1. Projections to neurons in most cortical regions are segregated on the dendritic tree, making the selective modulation of gain on input from individual sources quite feasible [17].

2. It is well known that, in the hippocampal region, which is the basis for our model, animals are especially attentive at the beginning of an episode, as indicated by the change in the EEG theta rhythm. This leads to, for example, greater spike synchronization, lower firing latency, and other phenomena [20].

3. In the granule cells of the dentate gyrus, which, we hypothesize, corresponds roughly to our layer L_R, there is both anatomical and physiological evidence [16, 18] of an intricate and highly specific system of excitability modulation based on motivation and attention [20].

The second problem - the preservation of the network's state during irrelevant inputs - is addressed by the functionality of the biasing layer. The role of the biasing layer L_B is to sustain the level of activity in candidate latent attractors during the presentation of irrelevant patterns in the context sequence. Each neuron in L_B corresponds to one context pattern. When a context pattern is presented at the stimulus layer L_S, the associated neuron in the biasing layer becomes active. The active biasing neuron, in turn, projects back a higher excitation to the L_R neurons in the active sets of the attractors with which that context pattern is associated. Thus, in between context patterns, the activity in the latent attractors tends to be preserved until a new context pattern is presented at the input. The activity of the biasing neurons is reset after a latent attractor has been fully activated (i.e. at the end of a context sequence). The biasing layer plays the role of a short-term memory by sustaining the effect of context patterns until the context sequence is complete.

The weights of the connections between the biasing layer and the response layer (w_{ij}^{BR}) are set as follows:

$$w_{ij}^{BR} = \begin{cases} 1/\mu_k & , \quad i,\ j \in c^k \\ 0 & , \quad otherwise \end{cases} \tag{6}$$

The weights from a biasing neuron to the active set of an attractor in the L_R layer are normalized by the length of the context set (μ_k) which is associated with that attractor. Thus, neurons in the active set of an attractor corresponding to a context (c^k) of length μ_k, receive connections from μ_k biasing neurons. The weight of these connections is inverse proportional with the length of the context. The rule normalizes the total synaptic input coming from the biasing layer into a L_R neuron. At the end of a context sequence only the active set of the correct attractor will receive full excitation from the biasing layer. The normalization rule addresses the third problem of the system - independent context recognition behavior with respect to the length of a context μ_k. The rule ensures that shorter contexts have the same chance as longer ones to be recognized. A biological motivation for the normalization rule is the limited number of synaptic sites which constrains the total change in synaptic strength by the number of connections a neuron receives.

IV. SIMULATION RESULTS

Simulation were done using a four layer latent attractor network with the following parameters: $N_S = 400$, $K_S = 40$, $p_S = 0.4$, $N_B = 20$, $p_{B1} = 0.9$, $p_{B2} = 0.9$, $N_R = 2000$, $G_R = 200$, $K_R = 40$, $p_R = 0.4$, $N_H = 500$, $G_H = 50$, $K_H = 45$, $p_H = 0.8$. There are $M = 10$ attractors embedded in the connections between L_R and L_H layers. The modulation rate for recurrent gain g_i is $\alpha = 0.3$ and $\beta = 25$. The gain of the L_B to L_R projection is $g_{bias} = 24$.

The context set C has 20 distinct patterns, from which $\nu = 5$ context sets are selected. Each c^k consists of μ_k distinct patterns picked randomly without repetition from C. The length of each sequence is chosen uniformly in the interval $\mu_k^{min} = 2$ and $\mu_k^{max} = 6$. Context patterns in distinct c^k's are not mutually exclusive, but context sets completely included in another context set are excluded. Each c^k set of context patterns is associated with a randomly chosen attractor: The connections from L_S to L_R layers are potentiated such that patterns in c^k are associated with the neurons in the active set of the appropriate attractor. Also, each individual context pattern is associated with a neuron in the biasing layer through Hebbian potentiation of the connections from the stimulus layer L_S and the L_B layer. In turn, each biasing neuron, provides excitation to neurons in the active sets of those latent attractors whose context sequences include that context pattern.

At the beginning of each sequence, S^q, the recurrent gain for all L_R neurons is set to a low value. Depending on how many context groups are simultaneously stimulated by the incoming context patterns from c^{k_q}, the activity in L_R and L_H is distributed among the excited attractors. The recurrent gain of neurons in these attractors goes up, while that of other neurons decreases. At the end of a context sequence, only one attractor is consistent with the whole set of context stimuli in c^{k_q}, and almost all activity should be concentrated in its active set.

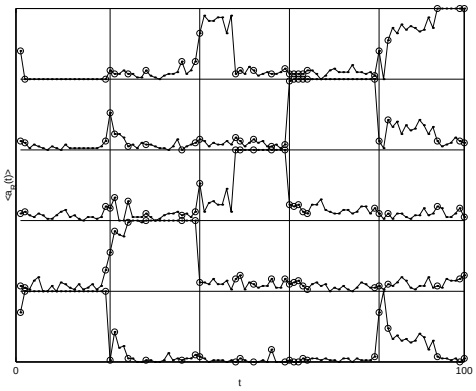

Figure 1: The activity level in the selected L_R attractors with respect to time. Every $r_q = 20$ time steps a different context sequence starts. The activity is normalized with respect to K_R. The time steps when context patterns are presented is denoted with circles.

In the first set of simulations, the context patterns are presented in a random order at the stimulus layer interleaved with irrelevant patterns. The total length of the context sequence is fixed to $r_q = 20$, but the position of the m_q context patterns and of the $r_q - m_q$ irrelevant patterns is not fixed. Figure 1 shows the result of a single network simulation, during one presentation of each context sequence. Each graph represents the normalized activity within the active set of an attractor in the L_R layer. It can be seen that, for each context sequence, the activity in only one of the attractors goes up steadily. In all other attractors the activity might increase for a few time steps, but it finally shuts down. In between consecutive context patterns, the activity is spread approximately equally be-

tween the candidate attractors. Figure 2 shows the results when the simulation is repeated with the same 5 context sequences but with the context patterns presented in different order each time. Each context sequence is followed by 10 regular patterns. It is clear that the activity remains confined within the chosen attractor even though the regular patterns have no association with any attractor. It can be seen that sometimes a wrong attractor almost wins the competition (the spikes) in the middle of a context sequence, but it is finally shut down.

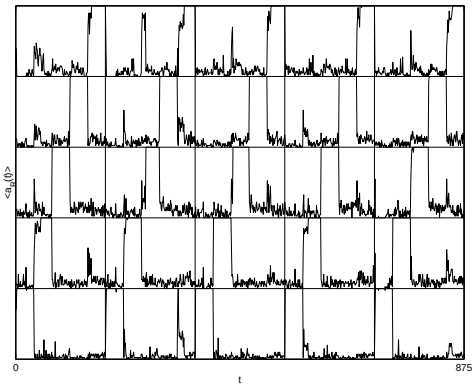

Figure 2: Five repeats of the run in Figure 1, with different context pattern order each time. Each context sequence is followed by a regular sequence of 10 patterns.

In the second set of simulations is presented the mean activity in L_R context attractors over five different networks, five different order presentations of the same context sequence. In Figure 3, the mean activity during the presentation of the context sequence is plotted as a function of time. The level of activity slowly goes up as more evidence is presented during the context sequence. The mean activity is averaged over all context sequences, independent of their length. Figure 4 shows the mean activity in the correct attractor with respect to the length, m_q, of the context set. The mean is averaged over the regular sequence patterns - the last 10 irrelevant patterns at the end of a context sequence. The plot shows that the correct attractor is active at the end of a context sequence and that it remains stable in face of noisy external inputs, independent of the length of the context set.

V. CONCLUSIONS

We have demonstrated a mechanism by which a latent attractor can be activated progressively by a set of context-setting stimuli presented sequentially in a random order interspersed with a variable number irrelevant stimuli. The system is able to select and activate the right attractor progressively even though individual input patterns are associated with multiple attractors. The system overcomes the disrupting effect of the irrelevant patterns by trying to maintain the state of attractor participation until a new relevant pattern is encountered. This requires a subtle gain control scheme consistent with experimental

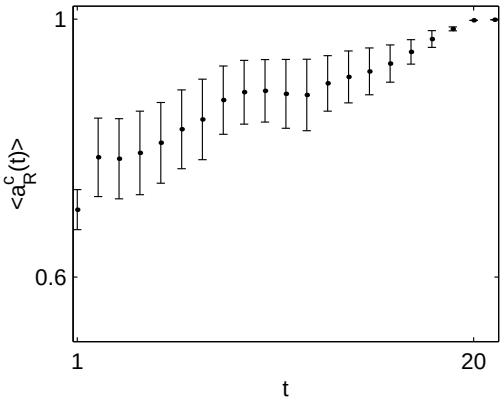

Figure 3: The graph presents the mean activity in the context attractors of the L_R layer as it varies in time, during the presentation of the context sequence.

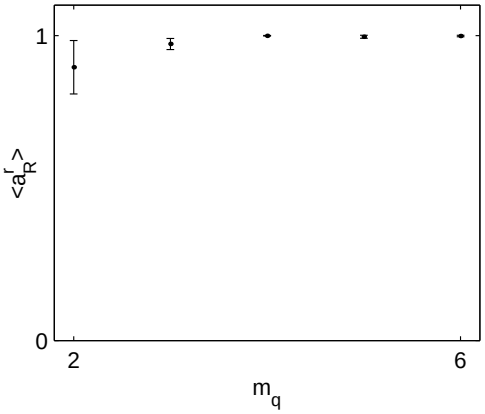

Figure 4: The plot shows the mean and the standard deviation of the activity in the correct attractor in the L_R layer with respect to the length of a context set (m_q). The mean is computed over five different networks, five different order presentations of the context patterns, and over the last 10 regular patterns at the end of each context sequence.

evidence in the hippocampus, but whose precise neural correlates remain to be fully explored. This will be discussed in future papers.

References

[1] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. on Neural Networks*, vol. Vol. 5, No. 2, pp. 157–166, 1994.

[2] G. Bradski, G.A. Carpenter and S. Grossberg. STORE working memory networks for storage ad recall of arbitrary temporal sequences. *Biological Cybernetics* 71:469–480, 1994.

[3] S. Doboli, A.A. Minai and P.J. Best. Generating smooth context-dependent representations. *Proc. of IJCNN'1999*, 1999.

[4] S. Doboli, A.A. Minai, and P.J. Best. A latent attractors model of context-selection in the dentate gyrus-hilus system. *Neurocomputing* 26-27:671–676, 1999.

[5] S. Doboli, A.A. Minai and P.J. Best. Latent attractors: a model for context-dependence place representations in the hippocampus. *Neural Computation* 12:1009–1043, 2000.

[6] S. Doboli and A.A. Minai. Network capacity for network attractor computation. *Proc. IJCNN'2000* 222-228, 2000.

[7] S. Doboli, A.A. Minai, and P.J. Best, A comparison of context-dependent hippocampal place codes in 1-layer and 2-layer recurrent networks, *Neurocomputing*, 3-33:353–358, 2000.

[8] S. Doboli, A.A. Minai, Progressive attractor selection in latent attractor networks, *Proc. IJCNN'2001*, Washington, USA, 2001.

[9] S. Doboli, A.A. Minai, Latent attractor selection in the presence of irrelevant stimuli, Proc. IJCNN'2002, Hawaii, USA, 2002.

[10] D.R.C. Dominguez. Information capacity of a hierarchical neural network. *Phys. Rev. E* 58:4811–4815, 1998.

[11] V.S. Dotsenko. Hierarchical model of memory. *Physica A*, 410–415, 1986.

[12] J.L. Elman, "Finding structure in time," *Cognitive Science.*, vol. 14, pp. 179–211, 1990.

[13] P. Frasconi and M. Gori, "Computational capabilities of local-feedback recurrent networks acting as finite-state machines," *IEEE Trans. on Neural Networks*, vol. Vol. 7, No. 6, pp. 1521–1525, 1996.

[14] C.L. Giles, C.B. Miller, D. Chen, H.H. Chen, G.Z. Sun, and Y.C. Lee, "Learning and extracting finite state automata with second-order recurrent neural networks," *Neural Computation*, vol. 4, pp. 393–405, 1992.

[15] S. Grossberg and C. W. Myers. The resonant dynamics of speech perception: Interword integration and duration-dependent backward effects. *Psychological Review*, 2000.

[16] Z.-S. Han, E.H. Buhl, Z. Lörinczi, and P. Somogyi, A high degree of spatial selectivity in the axonal and dendritic domains of physiologically identified local-circuit neurons in the dentate gyrus of the rat hippocampus. *Eur. J. Neurosci.* 5, 395–410, 1993.

[17] M.E. Hasselmo, E. Schnell, and E. Barkai. Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in hippocampal region CA3. *J. Neurosci.*, 15:5249–5262, 1995.

[18] M.B. Jackson and H.E. Scharfman, Positive feedback from hilar mossy cells to granule cells in the dentate gyrus revealed by voltage-sensitive dye and microelectrode recording. *J. Neurophysiol.* 76, 601–616, 1996.

[19] A.A. Minai and P.J. Best. Encoding spatial context: A hypothesis on the function of the dentate gyrus-hilus system. *Proc. of IJCNN'1998* 587-598, 1998.

[20] E.I. Moser. Altered inhibition of granule cells during spatial learning in an exploration task. *J. Neurosci.* 16:1247-1259, 1996.

[21] Quirk, G.J. and Muller, R.U. and Kubie, J.L. The firing of hippocampal place cells in the dark depends on the rat's recent experience *J. Neurosci.* 10:2008-2017, 1990.

[22] Rotenberg, A. and Muller, R.U. Variable place-cell coupling to a continuously viewed stimulus: Evidence that the hippocampus acts as a perceptual system *Phil. Trans. R. Soc. Lond. B*, 352:1505-1513, 1997.

[23] D. Willshaw, O.P. Buneman and H.C. Longuet-Higgins. Non-holographic associative memory. *Nature* 222:960-962, 1969.