# How the Brain Formulates Memory: A Spatio-Temporal Model

**Jun Hu**
Institute for Infocomm Research, A*STAR, Singapore
**Huajin Tang**
College of Computer Science, Sichuan University, Chengdu, China
Institute for Infocomm Research, A*STAR, Singapore
**K. C. Tan**
Department of Electrical and Computer Engineering, National University of Singapore, Singapore
**Haizhou Li**
Institute for Infocomm Research, A*STAR, Singapore
Department of Electrical and Computer Engineering, National University of Singapore, Singapore
School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, Australia

***Abstract*—Memory is a complex process across different brain regions and a fundamental function for many cognitive behaviors. Emerging experimental results suggest that memories are represented by populations of neurons and organized in a categorical and hierarchical manner. However, it is still not clear how the neural mechanisms are emulated in computational models. In this paper, we present a spatio-temporal memory (STM) model using spiking neurons to explore the memory formulation and organization in the brain. Unlike previous approaches, this model employs temporal population codes as the neural representation of information and spike-timing-based learning methods to formulate the memory structure. It explicitly demonstrates that the complex spatio-temporal patterns are the internal neural representations of memory items. Two types of memory processes are analyzed and emulated: associative memory, i.e., spatio-temporal patterns driven by intra-assembly connections, and episodic memory, i.e., temporally separated spatio-temporal patterns linked by inter-assembly connections. Our model will provide a computational substrate based on low-level neural circuits for developing neuromorphic cognitive systems with wide applications.**

## I. Introduction

Memory is an extremely complex brain-wide process, which is an indispensable part of species intelligence. Over the past few decades, researchers have devoted significant efforts to modeling memory mechanisms, in particular, internal representation of memory and memory organization in the brain.

The first question related to memory representation is how information is encoded in the nervous system. As a traditional coding scheme, rate coding assumes that the most important information about a stimulus is described by the firing rates of sensory neurons. However, rate codes fail to describe rapidly varying real-world stimuli. Recent experimental studies show that spike timing makes sense in visual [1], auditory [2], olfactory [3] pathways and hippocampus [4] in various neuronal systems [5]. It has been reported that precisely timed spikes play a pivotal role in the integration process of cortical neurons [6]. In addition, studies of population coding suggest that information can be encoded by clusters of cells rather than single cells [7]. Population coding has been found to exist throughout the nervous system. Visual features and natural sounds are encoded with population codes in the visual cortex [8] and auditory cortex [9], respectively. In addition, temporal population coding has been found to be capable of encoding visual stimuli invariantly and related to memory [10], [11]. We believe memory coding is achieved by combining temporal codes and population codes.

With development in large-scale ensemble recording techniques, network-level functional coding units termed *neural assembly* (or population) have been identified in the hippocampus [12]. Moreover, a recent study on population response patterns in monkey inferior temporal cortex suggests that external stimuli can be represented by responses of neural populations, and encoded memory patterns are organized in a hierarchy structure [13], [14].

The organization of memory is closely associated with the learning process in the nervous system. Spike-timing-dependent plasticity (STDP) and other spike-timing based learning schemes are thought to be involved in the formation of neural assemblies and associative memory [15], [16]. In addition, different learning algorithms using spiking neuron have been proposed to study hetero-association [17]–[21]. However, the formulation and organization of auto-associative and episodic memories (Fig. 1) by virtue of temporal coding and learning remain underexplored.

In this work, we propose a hierarchically structured spiking neural network model, called spatio-temporal memory (STM), which is able to study the formation of neural assemblies and the organization principle. In this model, sensory information travels upwards along the hierarchical network during the bottom-up information processing (Fig. 2). With a spike-timing based learning algorithm during the storing phase, the model maps sensory information into neural assembly

Corresponding author: Huajin Tang (E-mail: htang@scu.edu.cn).

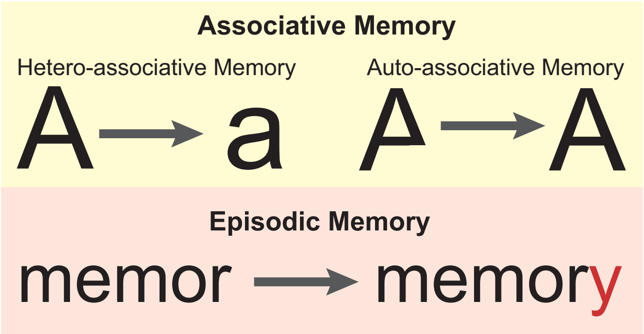

Fig. 1. Retrieval of associative memory and episodic memory. Hetero-associative memory is a mapping from one pattern to another. Auto-associative memory associative the input pattern to itself. Episodic memory is a collection of experiences in time in a serial form.

activities to form memory items. We demonstrate that auto-associative memory is formulated via fast STDP learning (in Layer I and II) and episodic memory produced by slow STDP learning (in Layer II). The main contribution of this model lies in that memory formulation is implemented by temporal population coding and temporal learning, which are missing components in existing memory models [22], [23]. In [24], a hierarchical model based on spiking neurons named *Cortext* was proposed. Although timing of spikes is employed in Cortext, it mainly targets visual recognition, while temporal coding and memory organization are not exploited. The hierarchical temporal memory (HTM) model [25], is successful in emulating the associative and episodic memory functions. However, it is not able to simulate the memory formulation and organization process through complex spatio-temporal dynamics of spiking neurons.

In summary, the main features of the STM model include:

1) It is the first comprehensive computational model that integrates spiking neural learning and memory formulation dynamic process. In contrast to existing memory models focusing only on memory functions, the STM model provides a low-level neural circuits based substrate that is feasible to devise neuromorphic cognitive systems in hardware, e.g., neuromorphic chips.
2) As a multi-layer hierarchical structure composed of spiking neurons, the STM model is capable of analyzing and illustrating various important computational primitives in the complex memory process, from neural assemblies, intra- and inter-assembly connections, to spatio-temporal dynamics of neural activities.
3) STM has successfully emulated important memory functions based on spiking neural dynamics, illustrating that neural assemblies and their spatio-temporal patterns of activities serve as the internal representation of memory, temporal learning steers the network to associate neural activities with input patterns and the learned associations are distributively stored in the connections within and between assemblies.

The remaining of this paper is organized as follows: in Section II, we introduce the general structure of the STM model, neural coding and learning algorithms. In Section III, the performance of the STM model is demonstrated by simulation results. In Section IV, important issues of the proposed model and related works are discussed. The conclusion is drawn in Section V.

## II. The Spatio-temporal Memory Model

This section introduces spiking neuron models, neural oscillations, basic network architecture, neural coding and learning algorithms in detail.

### A. Neuron Model

The spike response model (SRM), which provides a simple description of the spiking neuron, has been widely used in various studies [26]. The state of neuron $i$ is described by its membrane potential $v_i(t)$. A spike is generated when the membrane potential reaches its threshold ($V_{thr} = 1$). The spike response model can be written as

$$v_i(t) = \eta(t - t_i) + \sum_j w_{ij}\varepsilon_{ij}(t - t_j) + h^{ext}(t) \qquad (1)$$

where $t_i$ and $t_j$ denote firing times of the presynaptic neuron $j$ and the post-synaptic neuron $i$, respectively. $w_{ij}$ is the synaptic efficacy from neuron $j$ to neuron $i$. $\eta(t - t_i)$ is the refractory kernel modeling the neural dynamics after firing. $h^{ext}(t)$ is external stimulating input. The kernel $\varepsilon_{ij}(t - t_j)$ models response of neuron $i$ to the presynaptic spike (single spike is assumed for simplicity) from neuron $j$ as

$$\varepsilon_{ij}(s) = V_{norm}(exp(-\frac{s}{\tau}) - exp(-\frac{s}{\tau_s})) \qquad (2)$$

where $s = t - t_i$ is the interval after the firing of presynaptic neuron $i$, $V_{norm}$ is used to normalize the maximal value of $\varepsilon_{ij}(s)$, $\tau = 10\,ms$ and $\tau_s = 2.5\,ms$ are time constants.

Pyramidal cells, which are the most numerous excitatory cell type in mammalian cortical structures, are employed in our model (Layer I) simulating short-term memory. By utilizing the slow build-up ramp of after-depolarizing potential (ADP) of pyramidal cells [27], the status of neurons can be maintained through repetitive firing. We plug ADP kernel into kernel $\eta(t - t_i)$ to describe the dynamics of pyramidal cell $i$ at time $t$ as

$$\eta(t - t_i) = A_{ADP}\frac{t - t_i}{\tau_{ADP}}\exp(1 - \frac{t - t_i}{\tau_{ADP}}) \qquad (3)$$

where $A_{ADP} = 0.88$ is the amplitude of ADP, and $\tau_{ADP} = 200\,ms$ is the time constant affecting the duration of excitatory ramp.

### B. Theta/Gamma Oscillations

Theta and gamma oscillations are two important types of brain wave for synchronizing the neural activity [28]. They are critical for temporal coding/decoding of active neuronal ensembles, learning and memory formation [29], [30]. An external theta oscillatory source $h^{ext}(t)$, which injects current to neurons in Layer I, is modeled as a cosine wave

$$h^{ext}(t) = v_\theta(t) = A_\theta \cos(2\pi f_\theta + \phi_0) \qquad (4)$$

where $A_\theta$ is the amplitude of sub-threshold membrane potential oscillation, $f_\theta = 6$ is the frequency of theta oscillation, and $\phi_0$ is the initial phase. It has been found that memory capacity depends on the theta/gamma cycle length ratio, suggesting that short-term memory is reserved within individual gamma cycles [31]. In the proposed model, each memory item is represented by firings in different gamma cycles. Meanwhile, inhibition from interneurons suppresses other neurons. The theta oscillation is simulated as an external signal source, while gamma oscillation virtually exists as spikes volley at a frequency similar to that of gamma waves.

### C. Network Architecture

The basic STM model is composed of three layers: input layer, Layer I and Layer II as shown in Fig. 2.

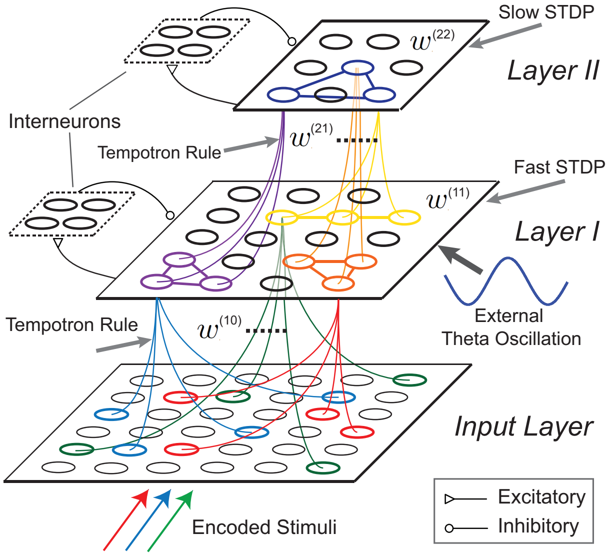

Fig. 2. The architecture of the STM model. Neurons forming neural assemblies and enhanced lateral connections are illustrated in different colors. Fast STDP and slow STDP learning algorithms are employed to adapt the connections within layers (Layer I and Layer II, respectively), while tempotron learning rule is applied to the connections between layers.

Neurons in the lower layer are fully connected to the next higher layer, and lateral connections exist in Layer I and Layer II. Interneurons provide feedback inhibition to prevent continuous firing and temporally separate firing events representing different memory items into gamma cycles. In order to distinguish state variables in different layers (e.g., $v_i^{(1)}$ denotes membrane potential of neuron $i$ in Layer I) and stimulating inputs (connections) from different layers, superscripts are used in the following equations (e.g., $v_i^{(10)}$ denotes summed input from the input layer received by neuron $i$ in Layer I. $w_{ij}^{(21)}$ denotes synaptic weight from neuron $j$ in Layer I to neuron $i$ in Layer II). The dynamics of pyramid cells in Layer I is specified by:

$$v_i^{(1)}(t) = \eta(t - t_i) + v_i^{(10)}(t) + v_i^{(11)}(t) + v_{inh}^{(1)}(t) + v_\theta(t) \quad (5)$$

with

$$v_i^{(10)}(t) = \sum_j w_{ij}^{(10)} \varepsilon_{ij}^{(0)}(t - t_j) \quad (6)$$

and

$$v_i^{(11)}(t) = \sum_{j \neq i} w_{ij}^{(11)} \varepsilon_{ij}^{(1)}(t - t_j) \quad (7)$$

where $v^{(10)}(t)$ and $v^{(11)}(t)$ are induced by input currents from neurons in input layer and Layer I, respectively. $v_{inh}^{(1)}(t)$ is the inhibitory feedback from interneurons.

Similarly, each neuron in Layer II receives inputs from other neurons in the same layer and all the neurons in Layer I. The dynamics of neuron $i$ in Layer II is defined by

$$v_i^{(2)}(t) = v_i^{(21)}(t) + v_i^{(22)}(t) + v_{inh}^{(2)}(t) \quad (8)$$

with

$$v_i^{(21)}(t) = \sum_j w_{ij}^{(21)} \varepsilon_{ij}^{(1)}(t - t_j) \quad (9)$$

and

$$v_i^{(22)}(t) = \sum_{j \neq i} w_{ij}^{(22)} \varepsilon_{ij}^{(2)}(t - t_j) \quad (10)$$

Therefore, the network connectivity mainly includes two types of connections: First, lateral connections between neurons in the same layer. Second, inter-layer connections from input layer to Layer I and from Layer I to Layer II.

### D. Temporal Population Coding

The information about stimulation is encoded by the time of spikes generated by a specific population of neurons, and each input pattern is coded by a particular group of neurons. This work employs the temporal population coding to mimic sensory encoding process. Here, we take visual signal as an example to show how real-world stimuli can be encoded into single-spike spatio-temporal patterns as shown in Fig. 3. A grayscale image is fed into Garbor filters [32] and the output are converted into neural firings corresponding to the following equation.

$$t_i = f(s_i) = t_{max} - ln(\alpha \cdot s_i + 1) \quad (11)$$

where $t_i$ is the firing time of neuron $i$, $t_{max}$ is the width of encoding window, $\alpha$ is a scaling factor, and $s_i$ is the intensity of output of Garbor filter. As a result, each spike codes orientation components of the image and the latency denotes the weight of the corresponding component.

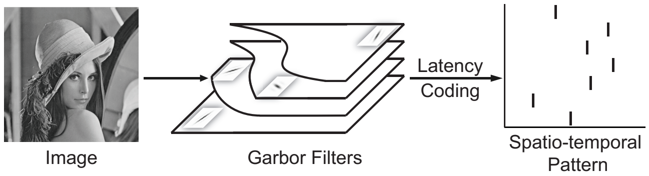

Fig. 3. Encoding scheme. A grayscale image is convolved with Garbor filters to extract orientation related features and then converted into a spike pattern by latency coding method.

### *E. STDP and Tempotron Learning Rule*

As precise spike timing and the interval between pre- and postsynaptic firing were discovered, learning with millisecond precision has intrigued intensive interest. The temporally asymmetric form of Hebbian learning induced by temporal correlations between pre- and postsynaptic spikes is called STDP. Similar to other forms of synaptic plasticity, STDP is believed to be the underlying mechanism for learning and information storage in the brain [33]. It assumes that repeated presynaptic spikes contribute to the closely following postsynaptic action potential and lead to long-term potentiation (LTP) of the synapse, whereas an inverse temporal relation results in long-term depression (LTD) of the same synapse. Therefore, the change of the synapse is defined as a function of the relative timing of pre- and postsynaptic spikes, which is called the STDP function as shown in the following equation:

$$\Delta w_{ij} = \begin{cases} a^+ \cdot exp(\frac{s}{\tau^+}) & \text{if } s < 0 \\ -a^- \cdot exp(\frac{-s}{\tau^-}) & \text{if } s > 0 \end{cases} \tag{12}$$

where $w_{ij}$ is the synaptic weight from neuron $j$ to neuron $i$, $a^+$ and $a^-$ are amplitudes of exponential functions, and $s = t_j - t_i$ denotes the time difference between pre- and postsynaptic spikes. The STDP function (also called learning window) is illustrated in Fig. 4.

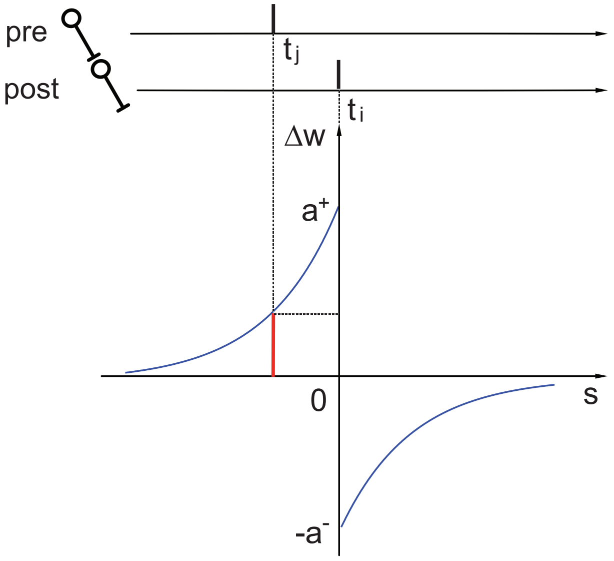

Fig. 4. Spike-timing-dependent plasticity (STDP).

In this model, we apply two different types of STDP processes: fast STDP in Layer I and slow STDP in Layer II. Neurophysiological experiments have found that synaptic modifications varies with different decaying time constants of postsynaptic N-methyl-D-aspartate (NMDA) receptors [34], [35], the predominant molecular device for controlling neural plasticity. In the STM model, STDP learning mediated by fast and slow NMDA receptors (fast: $\tau_{fast} \sim 25\,ms$, slow: $\tau_{slow} \sim 150\,ms$) is referred to as fast and slow STDP, respectively (Fig. 5). Fast STDP in Layer I regulates neurons firing with a temporal distance less than gamma cycles, while slow STDP in Layer II results in synaptic modification between neurons firing with a greater temporal distance. By multiplying a simplified activation function of NMDA channel, the modified STDP can be rewritten as

$$\Delta w_{ij} = \begin{cases} a^+ \cdot exp(\frac{s}{\tau^+}) \cdot G(s) & \text{if } s < 0 \\ -a^- \cdot exp(\frac{-s}{\tau^-}) \cdot G(s) & \text{if } s > 0 \end{cases} \tag{13}$$

with

$$G(s) = \begin{cases} 1 & \text{if } |s| \leq \tau_{NMDA} \\ 0 & \text{if } |s| > \tau_{NMDA} \end{cases} \tag{14}$$

where the decaying constant $\tau_{NMDA} = \tau_{fast}$ is set for fast STDP and $\tau_{NMDA} = \tau_{slow}$ for slow STDP.

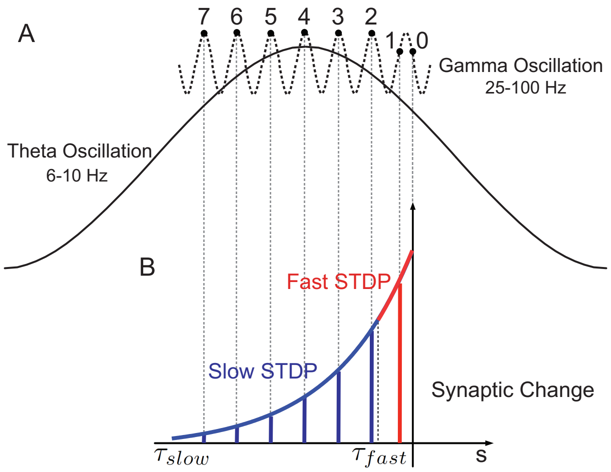

Fig. 5. LTP induced by STDP learning. (A) Firings within each gamma cycle represent memory items 0-7. (B) Synaptic changes depend on the time between firings. Connections within and between neural assemblies are formed via fast STDP ($1 \rightarrow 0$) and slow STDP ($7 \rightarrow 0$), respectively.

Among existing spike-timing based learning approaches, the tempotron rule is a biologically plausible supervised synaptic learning scheme compatible with temporal codes [17]. Tempotron rule is used to train the network to reproduce spatio-temporal patterns by adapting connections between layers. It replaces the post synaptic spike time ($t_j$) in STDP with the time ($t_{max}$) at which the postsynaptic potential reaches its maximal value. As a supervised learning rule, each neuron needs to make a decision on whether the presented stimulus contains features that have been learned before. The connections from neurons that contribute to the integrated postsynaptic membrane potential will be enhanced according to the tempotron learning rule as follows

$$\Delta w_i = \lambda d \sum_{s_i < 0} exp(s_i) \tag{15}$$

where $w_i$ is the synaptic weight from afferent $i$ to the postsynaptic neuron, $\lambda$ is the learning rate, $d$ is the desired output label (either 0 or 1), and $s_i = t_i - t_{max}$ is the delay between presynaptic firing ($S_i$) and the time when postsynaptic membrane potential $V(t)$ reaches its maximal value $V_{max}$. The tempotron learning rule is illustrated in Fig. 6.

## III. Simulation Results

In this section, we demonstrate that the proposed STM model is capable of learning patterns, storing them into individual gamma cycles with population firings, and performing

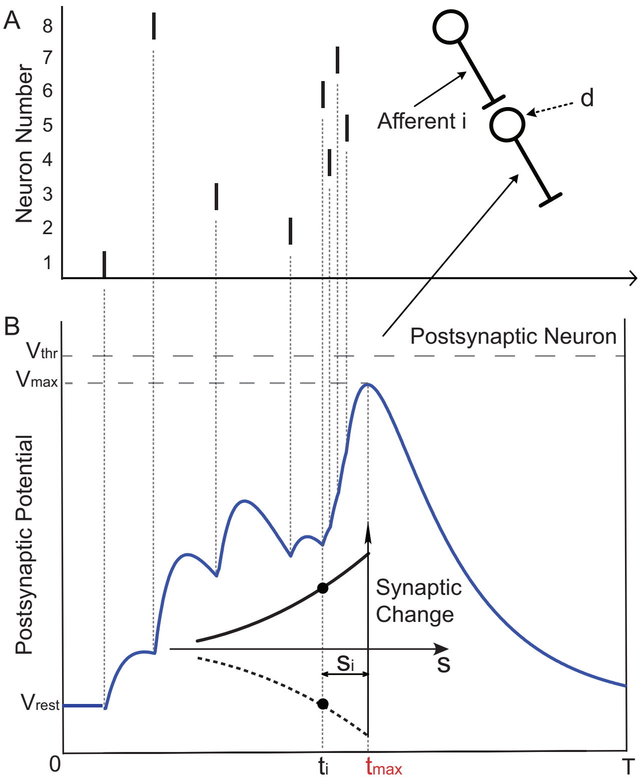

Fig. 6. Illustration of the tempotron rule. (A) Typical spatio-temporal pattern. (B) Membrane potential of the postsynaptic neuron. The maximum value of the membrane potential is reached at $t_{max}$. (inset) The synaptic weight $w_i$ changes accordingly to the time difference between $s$ and the desired signal $d$. If $d = 1$, $\Delta w_i \geq 0$ (solid line), or if $d = -1$, $\Delta w_i < 0$ (dashed line).

sequence learning. The results show how neural populations contribute to the formation of associative memory and episodic memory and how they are organized in a hierarchical network. Several experiments are conducted to illustrate and analyze these processes.

As shown in Fig. 2, the spiking neural network used to implement the STM model is composed of three layers. The synaptic weights are initialized according to the population size of each layer. Each input pattern (e.g., a letter) is represented by tens of spikes using temporal population codes as shown in Fig. 3, and they are introduced to the network during troughs of the theta oscillation. The inter-layer synaptic weights are updated according to the tempotron learning rule during the representation of input patterns (gray strips in Fig. 7), while intra-layer synaptic plasticity is modified by STDP learning.

### *A. Network Performance*

Driven by input synaptic currents, increasing number of pyramidal cells in Layer I start to fire and form different neural assemblies iteration by iteration as shown in Fig. 7. Neural assemblies coding for different input patterns can be identified in Layer I and II, respectively (Fig. 7B and 7C). Within each theta cycle, neural assemblies respond selectively and repetitively to the stimulation in the same order as input patterns get introduced to the network. As can be seen from Fig. 7B, individual letters ('L', 'O', 'V', and 'E') are separately encoded by the volley activities of corresponding neural assembly in Layer I. While neural activities generated by all four neural assemblies in Layer II are coding for the word 'LOVE'. The memory coding principle is explained in detail in section III A and B, respectively.

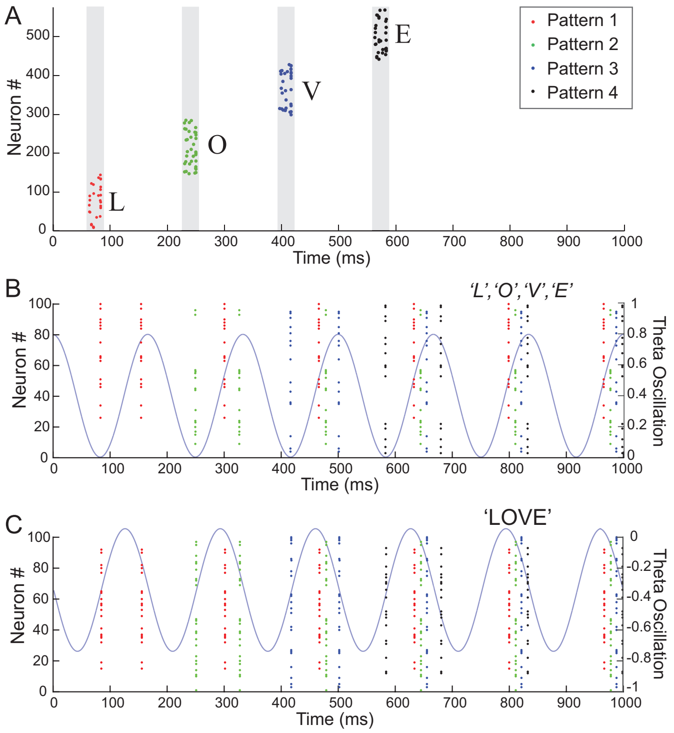

Fig. 7. Neural activity propagates through the system. (A) Each input patterns consists of firings within an encoding window (gray strips). (B) and (C) are the raster plots of the neural activities in Layer I and II, respectively. Colored dots denote spikes generated by neurons coding for different input patterns.

Fig. 8 reveals the mechanism underlying repetitive firing of pyramidal cells. After generating the first spike by a particular neuron, its ADP starts to build up. When the slowly ramping up ADP meets near-peak theta current, the pyramidal cell will fire again in the following theta cycle. Meanwhile, inhibitory feedback from interneurons prevents neurons coding for other patterns from firing right after the volley spikes. As a result, spike volleys are temporally separated into individual gamma cycles (Fig. 7A).

In sum, neurons forming the same neural assembly tend to fire in synchrony, and neural assemblies coding for successive patterns are temporally compressed. The synchrony is caused by fast STDP in Layer I, while the compression is resulted from slow STDP in Layer II. As demonstrated in Fig. 7, neurons coding for different memories fire in gamma cycles in Layer I represent the detection of individual patterns, while neural responses within each theta cycle in Layer II represent the recognition of a sequence of patterns. As neural assemblies identified in Layer II can be considered as a whole assembly coding for a particular sequence of patterns, inter-assembly episodic memory binds information about temporally separated patterns (letters) into a compressed pattern (word).

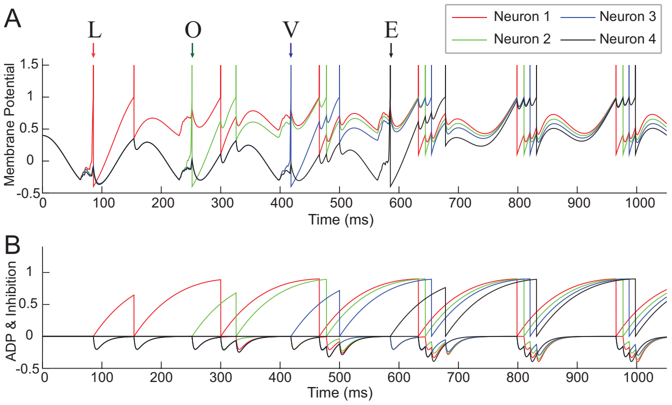

Fig. 8. Typical neural responses of pyramidal cells in Layer I. (A) Membrane potentials of neurons coding for different patterns. (B) ADP of pyramidal cells (positive) and inhibition from interneurons (negative).

## B. Network Connectivity

Since fast and slow STDP processes take place in Layer I and II, respectively, the resulted lateral connectivities are different. Therefore, we examine the synaptic weights, especially intra-assembly connections in Layer I and inter-assembly connections in Layer II as presented in Fig. 9 and Fig. 10.

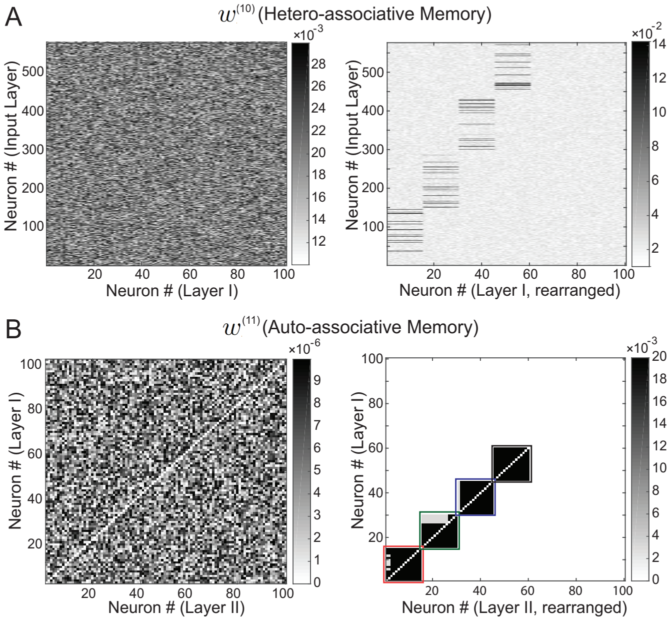

Fig. 9. Neural connectivity from input layer to Layer I (A) and within Layer I (B) before (left) and after learning (right). The activated neurons are picked out and rearranged for clear illustration in the right column. Intra-assembly connections are highlighted by colored boxes.

At the beginning, connections among neurons are randomly initialized (Fig. 9, left column). As learning proceeds, some relative strong synaptic weights (darker dots) are developed (Fig. 9, right column). Similar phenomenon can be observed in Fig. 10. The enhanced connections are caused by the input patterns and reform the structure of the network.

When exposed to external stimuli, neurons in Layer I start to fire due to the enhancement of connections from input layer to Layer I as shown in Fig. 9. Once neurons in Layer II receive enough stimulation from Layer I, they begin to generate spikes. At the same time, activated neurons within the same layer wire together to form neural assemblies as shown in Fig. 9B ($w^{(11)}$) and Fig. 10B ($w^{(22)}$).

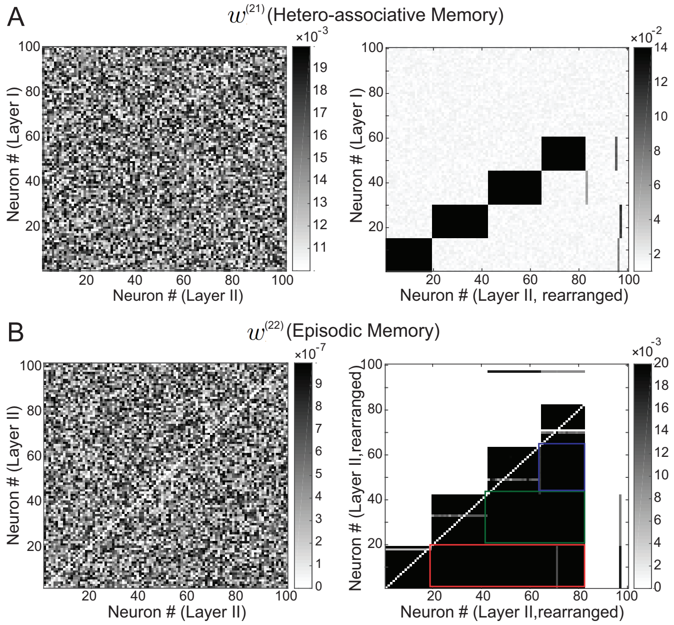

Fig. 10. Evolution of the neural connectivity from Layer I to Layer II (A) and within Layer II (B). Inter-neural assembly connections in Layer II are highlighted by colored boxes.

To further study the resulted neural assemblies and their connectivities, we take a closer look at synaptic connections within Layer I and II. Generally, lateral connections can be divided into intra-assembly, inter-assembly and weak connections. The connectivity developed after learning is illustrated in Fig. 11. As non-activated neurons wire weakly to all the other neurons, only intra-assembly and inter-assembly connections are drawn in Fig. 11.

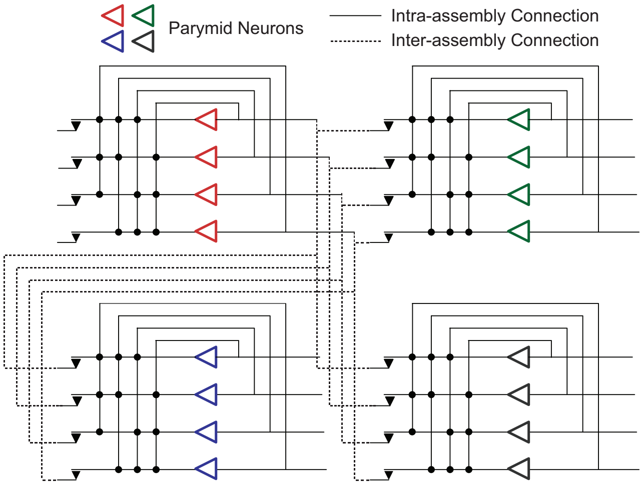

Fig. 11. Schematic diagram of developed lateral connectivity. Lateral connections within the same layer are divided into intra-assembly and inter-assembly connections. Only inter-assembly connections from the first neural assembly to the rest are drawn for clear illustration.

Since fast NMDA receptors stay activated for several milliseconds, only firings of neurons forming the same assembly fall within the time window. Consequently, intra-assembly connections are enhanced via STDP as shown in Fig. 9B. The highlighted weights matrices show that each neural assembly forms a recurrent subnetwork with auto-associative memory coded in the enhanced lateral connections.

Although lateral connections in Layer II were strengthened as those in Layer I, the resulting connectivity is different from that in Layer I. As slow STDP has a wider learning window spanning over several gamma cycles, spikes in different cycles would induce enhancement of inter-assembly connections. Salient weights along the diagonal in Fig. 10B are similar to those in Layer I, where auto-associative memory is stored. While elements in the colored boxes denote connections between neural assemblies, in which episodic memories are encoded.

### C. Auto-Associative Memory

Despite constant changes in real-world environment, our brain has a remarkable ability to associate. Along sensory pathways, information about external stimulation is encoded into reliable neural activities. After training, associative memories (Fig. 1) are stored in the connections between neurons. Input patterns are hetero-associated with neural responses in Layer I via synaptic weights between the input layer and Layer I. Connections from the input layer to Layer I form the mapping from an input pattern to the activation of a particular neural assembly. At the same time, auto-associative memory is represented by intra-assembly connections. These lateral connections form a recurrent subnetwork, which can be activated without enough input stimulation (incomplete pattern).

As neural activities can be observed as an explicit expression of stored memory, pattern completion may refer to the ability that a subset of neurons from a particular neural assembly is able to arouse the rest of that assembly. The trained network is expected to be competent for recalling similar neural activities upon presentation of learned patterns and retaining invariant responses in the presence of noises and even corruption of information. Since temporal population coding scheme is employed, the lost information can be recovered with the aid of other contributing neurons. In order to investigate this capability of reproducing neural activities, time jitter and missing of spikes are considered in the following experiments. A correlation-based measure of spike timing [36] is used to calculate the distance between an output pattern and its corresponding target pattern.

$$C = \frac{\overrightarrow{s_1} \cdot \overrightarrow{s_2}}{|\overrightarrow{s_1}||\overrightarrow{s_2}|} \tag{16}$$

where $C$ is the correlation denoting the closeness between two temporal coded patterns ($s_1$ and $s_2$). They are convolved with a low pass Gaussian filter of a width $\sigma = 2\,ms$.

By shifting input spikes, variability of input patterns is simulated as shown in Fig. 12A. The shifting intervals are randomly drawn from a Gaussian distribution with mean 0 and variance [0, 5] $ms$. The correlation between reproduced neural responses and the desired patterns is presented in Fig. 12. Each experiment has been repeated for 30 times to generate the averaged performance. Fig. 12B shows that the network reproduces reliable neural patterns in the presence of shifted input spikes up to 3 $ms$. However, the performance dramatically drops to around 0.3 as the shifting interval increases to 5 $ms$. Neural response in Layer I is slightly more robust than that in Layer II due to error accumulation during the upwards information propagation.

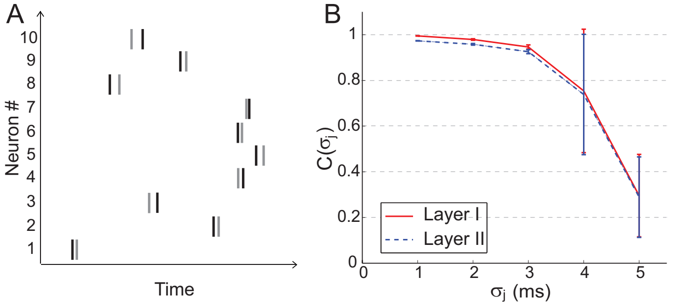

Fig. 12. (A) Illustration of shifted spatio-temporal patterns. Firing times of original input spikes (black bars) are shifted with random jitters (gray bars). (B) Reliability of retrieved neural responses under different noise levels.

Another experiment is conducted to investigate the link between intra-assembly connections and auto-associative memory. All settings are the same as in previous experiments, whereas one out of ten spikes is removed from each input pattern. The experiment has been run for 20 trials and the mean value of the correlation between the actual output and the desired pattern is calculated for each trial.

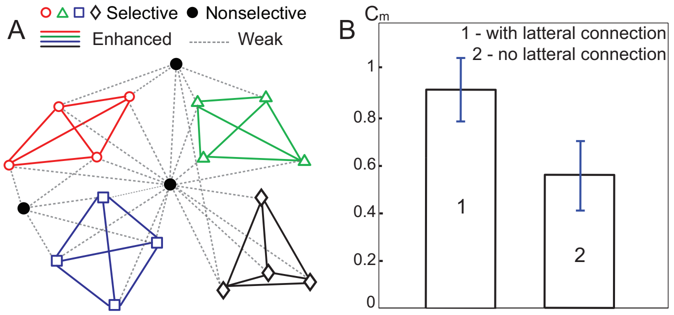

Fig. 13. (A) Illustration of neural assemblies in Layer I coding for different input patterns (letters) after learning. (B) Test results of the associative memory based on the correlation between retrieved and corresponding desired patterns in response to corrupted input patterns.

As illustrated in Fig. 13A, intra-assembly connections are enhanced during learning, while non-selective neurons are weakly connected to all the other neurons. To verify that connections within neural assemblies are responsible for the completion of patterns in Layer I, lateral connections are removed. As shown in Fig. 13B, the experimental results are consistent with our analysis.

Knowledge stored in the synaptic weights from the input layer to Layer I provides the capability of hetero-association (Fig. 9A) by recognizing specific features contained by input patterns. Since corrupted patterns provide insufficient stimulation to neurons in the next layer, some of the trained neurons that should have been activated may not be triggered. Fortunately, lateral inputs from excited neurons can provide supplementary information to recall desired neural responses. Therefore, the ability to retrieve invariant responses with partial information (i.e., associative memory) relies on the distributed knowledge stored in synaptic connections between layers and within layers as well.

### *D. Episodic Memory*

Since slow NMDA receptors dominate the STDP process in Layer II, it leads to different postsynaptic neural responses and different connectivity. The slow decaying time constant of slow NMDA receptors leads to the accumulation of excitatory postsynaptic potentials (EPSPs) from different neural assemblies. Meanwhile, slow NMDA receptor channel sustains its activation state over several gamma cycles, which enables STDP learning to link sequence of memory items by building up inter-assembly connections. When lateral connections are sufficiently developed, the accumulated EPSPs of occurred memory items would be able to trigger subsequent items without the presentation of expected upcoming input stimulation during memory recall.

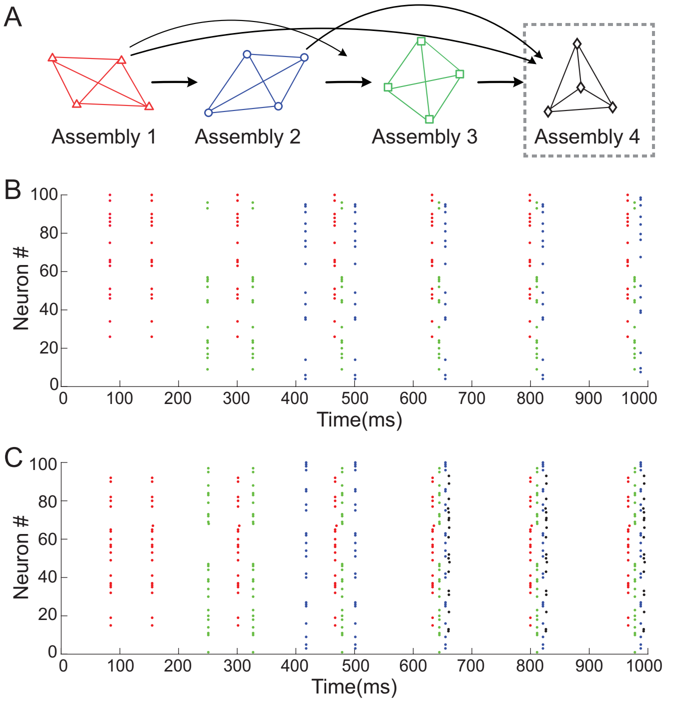

Fig. 14. (A) Illustration of generated connectivity in Layer II. (B) Raster plot of neural activities in Layer I during recall. Neurons coding for letters 'L', 'O', and 'V' detect the presentation of them and trigger firings in Layer II. (C) Episodic memory stored in Layer II helps to recall the missing item ('E').

As demonstrated in Fig. 14, stimulation caused by neural assemblies coding for the first three memory items in the sequence is strong enough to trigger the neural assembly coding for the missing item. The inter-assembly connections may lead to the result that consecutive memory items are temporally compressed as a group of neuron coding for the combination of several patterns in the sequence. This characteristic is crucial for a spike-timing based hierarchical model, which contributes to pattern/information binding process.

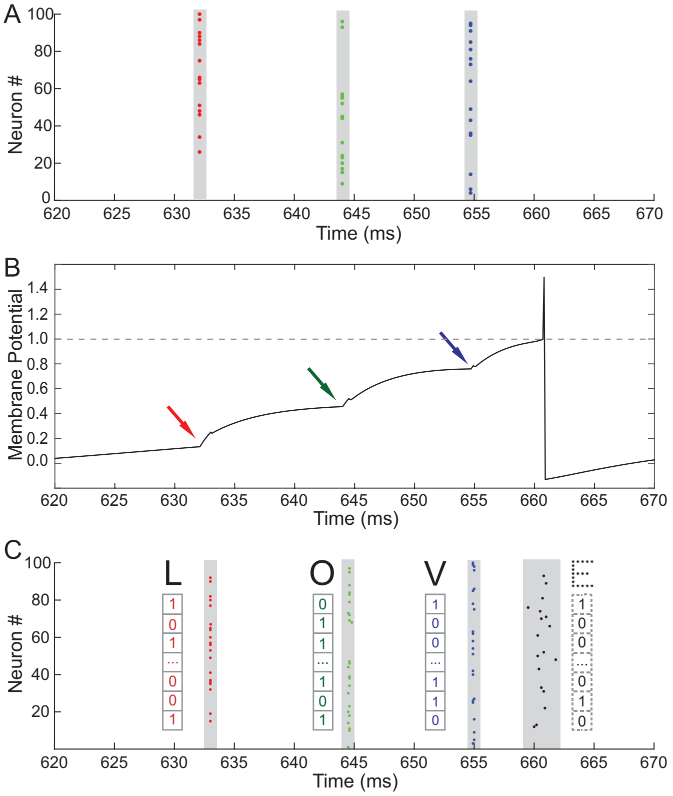

Fig. 15. Recall of neural activities induced by accumulated EPSPs. (A) Response of neural assemblies in Layer I. (B) Membrane potential of an activated neuron coding for the missing pattern ('E') in Layer II. (C) Raster plot of neural activities in Layer II and their corresponding binary codes.

Fig. 15 shows how EPSPs of consecutive memories lead to the activation of neural assemblies coding for the next upcoming pattern. As shown in Fig. 15, neural assemblies coding for them fired in Layer I and Layer II, respectively, while neural assembly coding for the fourth pattern remains silent. Due to the slow STDP and enhanced inter-assembly connections in Layer II, neural assemblies coding for the "missing" pattern were triggered by the accumulation of EPSPs induced by the neural assemblies coding for preceding patterns (Fig. 15B). Neural activities can be converted to binary codes according to their states within a certain coding time window (gray strips in Fig. 15C). Neurons are divided into groups with a constant size and checking the firing state of them. The neural state of a group is "0" if there is no spike generated and vice versa. Therefore, each input pattern (letter) can be coded by such a binary code. By reading out these binary codes, we can identify the presence of individual features/patterns in Layer I and combination of features/patterns (sequence) in Layer II.

## IV. Discussion

### *A. Emergence of Neural Assemblies*

Basically, the information flow in the model is unidirectional from bottom to top. Information between layer travels upwards along the network with a filtering process, while recurrent subnetworks (neural assemblies) exist in Layer I and Layer II. Input stimulation triggers repetitive neural activities in Layer I, these activities further drive neural responses in Layer II.

During learning, neurons compete with each other to respond selectively to specific stimulus. Synaptic weights between layers stop changing when the population sizes of evoked neurons reach the predefined configuration. Although bounded synapses have limited memory storage capacity [37], they are used in our model to ensure a certain number of presynaptic neurons contributing to the generation of postsynaptic spikes. In addition, recent work on neural self-organization can be considered for the generation of neural assemblies to improve the STM model [38], [39].

### *B. Storage, Recall, and Organization of Memory*

Storage and recall of memory are two important issues for a memory model. In the learning phase, hetero-association is achieved by enhanced synaptic weights between layers. Once pyramidal cells are triggered to fire, cooperation of ADP and theta oscillation results in the repetitive firing of neurons coding for memory items as short-term memory. Meanwhile, fast STDP and slow STDP contribute to the enhancement of intra- and inter-assembly connections, respectively.

Theta oscillation has been recorded in hippocampus involved in memory function. The model proposed in [40] suggests that memories might be encoded and recalled during different portions of the theta cycle. Similar scheme is employed in our model, hetero-associative memory storage occurs in troughs of theta oscillation, while stored memories are retrieved in portions near the maximums of theta oscillation. Since the activation of neural activities at troughs requires strong excitation, the resulted synaptic efficacies are stronger than required at the maximums. Redundant excitation and distributed information over neurons improve the robustness of recalling hetero-associative memories. Environmental noises and even information loss will not lead to a severe retrieval failure as demonstrated by the simulation results. Moreover, multiple patterns are encoded and stored into associative and episodic memories following a hierarchical organization principle. Hereto-associative memory is encoded by the connectivity between layers. Along with the development of neural assemblies, lateral connections are enhanced by STDP. Intra-assembly connections represent auto-associative memory, while episodic memory about the sequence of input patterns is encoded in the form of inter-assembly connections.

### *C. Temporal Compression and Information Binding*

The discovery of place cells suggests that spatial information can be encoded by the cellular activities of hippocampus. Moreover, dual oscillations have been observed to be involved in memory function. In the STM model, memory items are coded by neuron assemblies firing within different gamma cycles, while past and present events are chunked by the theta oscillation. When presented patterns that have been learned before, neural assemblies coding for each of them will be activated correspondingly. The temporal compression of neural firing volleys contributes to the generation of inter-assembly connections and ability to predict upcoming patterns.

Since neural responses in Layer II are linked with inter-assembly connections, information about different stimuli is binded as shown in Fig. 7C. Temporally compressed neural patterns can be treated as a new spatio-temporal pattern. By duplicating this basic network into a larger network, more powerful ability to organize neural activities representing features with different specificity along the hierarchy can be achieved. Each basic network binds several patterns (features) into a combined pattern (feature) and transmits it to a higher level network as its input stimulus. As a result, neural activities represent more specific and complex patterns along the hierarchical network.

### *D. Related Work*

The STM model shares some similar ideas with several existing studies in the field of neuroscience. The proposed model simulates neural assembly activities reported in [14], and is in agreement with the separation of encoding and retrieval theory suggested by [40], [41]. The mechanism sustaining short-term memory was used in Jensen et al.'s model, while tempotron learning and STDP learning have been employed in contributing long-term memory formation in other models.

Recurrent networks have been an important paradigm to implement auto-associative memory [22]. It has been demonstrated that simultaneous firings of a group of neurons can be stored in a fixed recurrent network modeling hippocampal CA3 area [42]. The idea that dual oscillation interacts with pyramidal cells has been implemented in the model. Although firing times of spikes are considered in the model, the external inputs exciting a specific pyramidal cell are presumed fire in synchrony, which ignores sensory encoding as well as the hetero-association process. In addition, recurrent subnetworks are predefined in the model and input patterns are presented to specific recurrent networks. These assumptions restrict the generalization and adaptability of the STM model.

A sequence learning model based on short-term memory mechanism was proposed in [43], which possesses similar features of recurrent network and hierarchical structure as our model. However, our model implements spiking neurons based computing to achieve a more biologically plausible memory model. Another sequences learning model in a hierarchical structure proposed by [44] employs prediction mechanism and minicolumn structure to realize episodic memory. The prediction mechanism might be considered to improve our model in the future.

The hierarchical temporal memory (HTM) [25] aims to develop a machine learning technology by mimicking the structural and algorithmic properties of the neocortex, featured by a sophisticated columns-based structure. In contrast to fixed structures, our STM model can generate neural assemblies during the learning process. The plasticity of network structure not only provides generalization and scale-up capability, but fully exploits available coding units of the network. Moreover, the HTM assumes that neural information is represented by rate codes and leaves out complexity and processing power of biological neurons. By using spiking neurons and incorporating biologically plausible mechanisms, our model is able to simulate the complex spiking neural dynamics for memory formulation and organization, which is a distinct feature compared to other cognitive learning and memory architectures aiming for developing machine learning alternatives (e.g., the hypernetwork model [45]).

Therefore, the STM model provides a comprehensive approach to build up low-level neural circuits for neuromorphic computing such as neuromorphic chips [46]. As brain-inspired approaches have been applied to solve various real-world problems [47], [48], efficiently implementing the STM model on platforms such as VLSI can utilize the inherent advantage of parallelism of neuromorphic computing.

## V. Conclusion

In this paper, the spatio-temporal memory (STM) model was introduced. The proposed model is able to store and recall both associative and episodic memories with a hierarchical structure. Throughout the STM model, temporal codes and temporal learning were integrated to process external stimuli and formulate memory. The results showed that neural assemblies can serve as the internal representation of memory. They also demonstrated that memories can be stored in the intra- and inter-assembly connections and organized in a hierarchical manner in consistent with neural mechanisms in the brain. Our model provides a comprehensive substrate to elucidate the complex process of memory formulation and organization in virtue of complex spiking neural dynamics. Real-world stimuli such as visual and auditory signals can be employed as the sensory information to investigate potential applications of STM model. Being able to more faithfully implement the dynamic details of memory formulation, our model will provide more insights to the design of neuromorphic cognitive systems.

## References

[1] M. Meister and M. J. B. II, "The neural code of the retina," *Neuron*, vol. 22, no. 3, pp. 435–450, 1999.

[2] P. Heil, "Auditory cortical onset responses revisited. i. first-spike timing," *Journal of Neurophysiology*, vol. 77, no. 5, pp. 2616–2641, 1997.

[3] J. Perez-Orive, O. Mazor, G. C. Turner, S. Cassenaer, R. I. Wilson, and G. Laurent, "Oscillations and sparsening of odor representations in the mushroom body," *Science*, vol. 297, no. 5580, pp. 359–365, 2002.

[4] M. R. Mehta, A. K. Lee, and M. A. Wilson, "Role of experience and oscillations in transforming a rate code into a temporal code," *Nature*, vol. 417, pp. 741–746, 2002.

[5] R. VanRullen, R. Guyonneau, and S. J. Thorpe, "Spike times make sense," *Trends in Neurosciences*, vol. 28, no. 1, pp. 1–4, 2005.

[6] C. Kayser, M. A. Montemurro, N. K. Logothetis, and S. Panzeri, "Spike-phase coding boosts and stabilizes information carried by spatial and temporal spike patterns," *Neuron*, vol. 61, no. 4, pp. 597–608, 2009.

[7] R. Natarajan, Q. J. Huys, P. Dayan, and R. S. Zemel, "Encoding and decoding spikes for dynamic stimuli," *Neural Computation*, vol. 20, no. 9, pp. 2325–2360, 2008.

[8] J. M. Samonds, Z. Zhou, M. R. Bernard, and A. B. Bonds, "Synchronous activity in cat visual cortex encodes collinear and cocircular contours," *Journal of Neurophysiology*, vol. 95, no. 4, pp. 2602–2616, 2006.

[9] A. Mizrahi, A. Shalev, and I. Nelken, "Single neuron and population coding of natural sounds in auditory cortex," *Current Opinion in Neurobiology*, vol. 24, pp. 103–110, 2014.

[10] R. Wyss, P. König, and P. F. M. J. Verschure, "Invariant representations of visual patterns in a temporal population code," *Proceedings of the National Academy of Sciences*, vol. 100, no. 1, pp. 324–329, 2003.

[11] M. Boerlin and S. Denève, "Spike-based population coding and working memory," *PLoS Computational Biology*, vol. 7, no. 2, p. e1001080, 2011.

[12] L. Lin, R. Osan, S. Shoham, W. Jin, W. Zuo, and J. Z. Tsien, "Identification of network-level coding units for real-time representation of episodic experiences in the hippocampus," *Proceedings of the National Academy of Sciences*, vol. 102, no. 17, pp. 6125–6130, 2005.

[13] R. Kiani, H. Esteky, K. Mirpour, and K. Tanaka, "Object category structure in response patterns of neuronal population in monkey inferior temporal cortex," *Journal of Neurophysiology*, vol. 97, no. 6, pp. 4296–4309, 2007.

[14] L. Lin, R. Osan, and J. Z. Tsien, "Organizing principles of real-time memory encoding: Neural clique assemblies and universal neural codes," *Trends in Neurosciences*, vol. 29, no. 1, pp. 48–57, 2006.

[15] M. Tsodyks, "Spike-timing-dependent synaptic plasticity–the long road towards understanding neuronal mechanisms of learning and memory," *Trends in Neurosciences*, vol. 25, no. 12, pp. 599–600, 2002.

[16] B. Szatmáry and E. M. Izhikevich, "Spike-timing theory of working memory," *PLoS Computational Biology*, vol. 6, no. 8, p. e1000879, 2010.

[17] R. Gütig and H. Sompolinsky, "The tempotron: A neuron that learns spike timing-based decisions," *Nature Neuroscience*, vol. 9, no. 3, pp. 420–428, 2006.

[18] F. Ponulak and A. Kasinski, "Supervised learning in spiking neural networks with resume: Sequence learning, classification, and spike shifting," *Neural Computation*, vol. 22, no. 2, pp. 467–510, 2010.

[19] R. V. Florian, "The chronotron: A neuron that learns to fire temporally precise spike patterns," *PLoS ONE*, vol. 7, no. 8, p. e40233, 2012.

[20] J. Hu, H. Tang, K. C. Tan, H. Li, and L. Shi, "A spike-timing-based integrated model for pattern recognition," *Neural Computation*, vol. 25, no. 2, pp. 450–472, 2013.

[21] Q. Yu, H. Tang, K. C. Tan, and H. Li, "Precise-spike-driven synaptic plasticity: Learning hetero-association of spatiotemporal spike patterns," *PLoS ONE*, vol. 8, no. 11, p. e78318, 2013.

[22] H. Tang, H. Li, and R. Yan, "Memory dynamics in attractor networks with saliency weights," *Neural Computation*, vol. 22, no. 7, pp. 1899–1926, 2010.

[23] E. Y. Cheu, J. Yu, C. H. Tan, and H. Tang, "Synaptic conditions for auto-associative memory storage and pattern completion in Jensen et al.'s model of hippocampal area CA3," *Journal of Computational Neuroscience*, vol. 33, no. 3, pp. 435–447, 2012.

[24] S. Schrader, M.-O. Gewaltig, U. Körner, and E. Körner, "Cortext: A columnar model of bottom-up and top-down processing in the neocortex," *Neural Networks*, vol. 22, no. 8, pp. 1055–1070, 2009.

[25] D. George and J. Hawkins, "Towards a mathematical theory of cortical micro-circuits," *PLoS Computational Biology*, vol. 5, no. 10, p. e1000532, 2009.

[26] W. Maass and C. M. Bishop, *Pulsed Neural Networks*. Cambridge, MA: MIT Press, 1998.

[27] M. S. Jensen, R. Azouz, and Y. Yaari, "Spike after-depolarization and burst generation in adult rat hippocampal CA1 pyramidal cells," *The Journal of Physiology*, vol. 492, pp. 199–210, 1996.

[28] J. O'Keefe and M. L. Recce, "Phase relationship between hippocampal place units and the EEG theta rhythm," *Hippocampus*, vol. 3, no. 3, pp. 317–330, 1993.

[29] B. C. Lega, J. Jacobs, and M. Kahana, "Human hippocampal theta oscillations and the formation of episodic memories," *Hippocampus*, vol. 22, pp. 748–761, 2012.

[30] N. Axmacher, F. Mormann, G. Fernández, C. E. Elger, and J. Fell, "Memory formation by neuronal synchronization," *Brain Research Reviews*, vol. 52, no. 1, pp. 170–182, 2006.

[31] J. Kamiński, A. Brzezicka, and A. Wróbel, "Short-term memory capacity (7+/-2) predicted by theta to gamma cycle length ratio," *Neurobiology of Learning and Memory*, vol. 95, no. 1, pp. 19–23, 2011.

[32] Q. Yu, H. Tang, K. C. Tan, and H. Li, "Rapid feedforward computation by temporal encoding and learning with spiking neurons," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 10, pp. 1539–1552, 2013.

[33] G. Q. Bi and M. M. Poo, "Synaptic modification by correlated activity: Hebb's postulate revisited," *Annual Review of Neuroscience*, vol. 24, pp. 139–166, 2001.

[34] R. C. Malenka and M. F. Bear, "LTP and LTD: An embarrassment of riches," *Neuron*, vol. 44, no. 1, pp. 5–21, 2004.

[35] J. W. Newcomer and J. H. Krystal, "NMDA receptor regulation of memory and behavior in humans," *Hippocampus*, vol. 11, no. 5, pp. 529–542, 2001.

[36] S. Schreiber, J. Fellous, D. Whitmer, P. Tiesinga, and T. Sejnowski, "A new correlation-based measure of spike timing reliability," *Neurocomputing*, vol. 52-54, pp. 925–931, 2003.

[37] S. Fusi and L. F. Abbott, “Limits on the memory storage capacity of bounded synapses,” *Nature Neuroscience*, vol. 10, no. 4, pp. 485–493, 2007.

[38] J. Chrol-Cannon and Y. Jin, “Computational modeling of neural plasticity for self-organization of neural networks,” *BioSystems*, vol. 125, pp. 43–54, 2014.

[39] ——, “Learning structure of sensory inputs with synaptic plasticity leads to interference,” *Frontiers in Computational Neuroscience*, vol. 9, no. 103 2015.

[40] M. E. Hasselmo, C. Bodelón, and B. P. Wyble, “A proposed function for hippocampal theta rhythm: Separate phases of encoding and retrieval enhance reversal of prior learning,” *Neural Computation*, vol. 14, no. 4, pp. 793–817, 2002.

[41] S. Kunec, M. E. Hasselmo, and N. Kopell, “Encoding and retrieval in the CA3 region of the hippocampus: A model of theta-phase separation,” *Journal of Neurophysiology*, vol. 94, no. 1, pp. 70–82, 2005.

[42] O. Jensen, M. Idiart, and J. E. Lisman, “Physiologically realistic formation of autoassociative memory in networks with theta/gamma oscillations: Role of fast NMDA channels.” *Learning & Memory*, vol. 3, no. 2-3, pp. 243–256, 1996.

[43] D. Wang and M. Arbib, “Complex temporal sequence learning based on short-term memory,” *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1536–1543, 1990.

[44] J. Starzyk and H. He, “Anticipation-based temporal sequences learning in hierarchical structure,” *IEEE Transactions on Neural Networks*, vol. 18, no. 2, pp. 344–358, 2007.

[45] B.-T. Zhang, “Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory,” *Computational Intelligence Magazine, IEEE*, vol. 3, no. 3, pp. 49–63, 2008.

[46] E. Neftci, J. Binas, U. Rutishauser, E. Chicca, G. Indiveri, and R. Douglas, “Synthesizing cognition in neuromorphic electronic systems,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 37, pp. E3468–E3476, 2013.

[47] Q. Ren, J. Xu, L. Fan, and X. Niu, “A gim-based biomimetic learning approach for motion generation of a multi-joint robotic fish,” *Journal of Bionic Engineering*, vol. 10, no. 4, pp. 423–433, 2013.

[48] B. Zhao, R. Ding, S. Chen, B. Linares-Barranco, and H. Tang, “Feedforward categorization on AER motion events using cortex-like features in a spiking neural network,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 9, pp. 1963–1978, 2015.