Feature-binding in working memory through neuronal synchronization

Joao Barbosa (palerma@gmail.com)
IDIBAPS, Barcelona, Spain

Kartik Sreenivasan (kartik.sreenivasan@nyu.edu)
NYU Abu Dhabi, United Arab Emirates

Albert Compte (acompte@clinic.cat)
IDIBAPS, Barcelona, Spain

Abstract

Swap-errors occur in working memory (WM) tasks when a wrong response is in fact accurate relative to a non-target stimulus. These errors reflect the failure to bind in memory the conjunction of features that define one object, and the mechanisms implicated remain unknown. Here, we tested the mechanism of synchrony across feature-specific neural assemblies. We built a biophysical neural network model for WM composed of two 1D attractor networks for WM, one representing colors and the other one locations. Within each network, gamma-oscillations were induced during bump-attractor activity through the interplay of fast recurrent excitation and slower feedback inhibition. These two networks are then connected via weak excitation, accomplishing color-location binding through the selective synchronization of pairs of bumps across the networks. Association-encoding was accomplished by stimulating simultaneously the corresponding bumps in each network, and feature-decoding by stimulating the cued location with a .5s pulse, which impacted strongly the corresponding phase-locked bump. In some simulations, "color bumps" abruptly changed their phase relationship with "location bumps" from which we derived a neural prediction: swap-errors are associated with a lower phase consistency of oscillatory activity in the delay period. Finally, we tested this prediction in MEG recorded from n=30 humans.

Keywords: working memory; attractor-networks; binding; network modeling of cognitive tasks

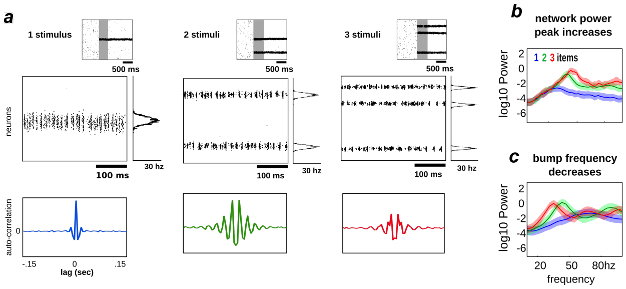

Figure 1: *Multiple bumps are spontaneously anti-correlated.* a) Top row, raster plots of 3 example simulations of load 1, 2 and 3. Middle. Zoomed version of simulations on the top show clear oscillatory activity, confirmed by cross-correlation functions (bottom). For the load 1 case, we computed the auto-correlation. Notably, irregular activity due to external noise coexists with markedly oscillatory dynamics. b, c) *Load-modulation of network and bump oscillatory dynamics.* Power spectrum computed from simulations of increasing load (1-3) using the activity of the whole network b) or of each bump's activity, c).

Working memory load modulates oscillation power and frequency

We built a computational network model of a local neocortical circuit, with excitatory and inhibitory spiking neurons (leaky integrate-and-fire neuron model) connected reciprocally via excitatory AMPAR-mediated and NMDAR-mediated synapses and inhibitory GABAAR synapses. The network model was tuned to support bump attractor dynamics with 3 simultaneous bumps (Edin et al., 2009), and further tuning of the relative weights of AMPAR and NMDAR-mediated currents set active reverberant neurons in the oscillatory regime (Compte, Brunel, Goldman-Rakic, & Wang, 2000). Using this computational model we started by investigating which dynamics were originated within each network. In our model, multiple bumps show anti-correlated oscillatory activity (Figure 1). As we store more bumps in the network, lateral inhibition originating from simultaneous memories establishes anti-phase dynamics during the memory period. Moreover, we found that the anti-phase behavior was robust in a wide range of values for AMPAR recurrent conductances (not shown, but see Figure 3 for the same robustness analyses for the full, connected model). To study load-dependent change in network dynamics, we ran multiple simulations with 3 different loads (presenting 1, 2 and 3 separate bumps during the encoding cue period) and we computed power spectra from either the aggregate activity of the whole network (network power) or from separate populations centered around each presented target (bump power). We found signatures of two different scenarios (Figure 1b,c). As we increase the memory load, the overall network activity oscillates at slightly increasing frequencies (Figure 1b). In contrast, each bump, corresponding to different memories, oscillates at markedly slower frequencies as load increases (Figure 1c). Thus, as shown before (Compte et al., 2000), the interplay between recurrent (fast) excitation and (slower) feedback inhibition acting locally is the basis of the bump oscilla-

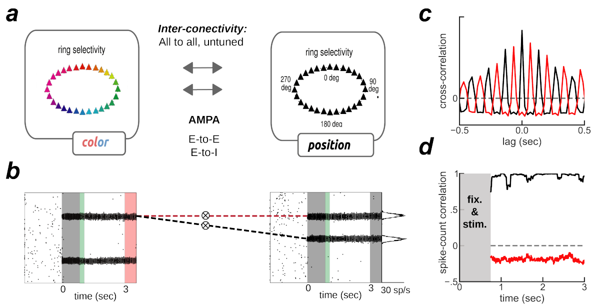

Figure 2: *Feature-binding through weak, uniform coupling.* a) Schematics of the final 2-network architecture, consisting of 2 ring-attractors with all-to-all, uniform connectivity. Each ring is able to store memories from one feature space (e.g. color or location) as activity-bumps. b) One example simulation for the two networks. The red-shaded area marks the period in which we read out the activity of the entire color network, while injecting current at one specific location in the location network (right gray-shaded area in the location rastergram, see main text for details about encoding/decoding). c) Cross-correlation computed between 2 pairs of bumps across networks (as marked with dashed red and black lines in panel b). For the black association, the cross-correlation peak is positive. In contrast, the cross-correlation peak was negative for the red association. d) Spike count correlation (in count bins of 5 ms, windows of 100 ms) of both associations through the memory delay is stable for this simulation.

tory behavior. Moreover, we now show that anti-phase dynamics of simultaneous bumps occurs due to bump competition, accomplished by lateral inhibition. Intuitively, this competition increases with memory load, leading to longer periods of silence during the delay-activity of each bump.

Uniform coupling achieves feature binding

How the conjunctions of different visual features are kept in mind is a long standing question in cognitive neuroscience (Schneegans & Bays, 2018) - the so called binding problem. However, there is consolidating evidence that different features of complex objects are maintained in independent stores (Schneegans & Bays, 2018). This suggests that different ring-attractors could be storing independent features, say 1 ring representing and memorizing colors and another ring storing locations (Ma, Husain, & Bays, 2014). However, how these networks should interact to accomplish feature-binding is unclear. Here, binding between color and location is accomplished through the synchronization of pairs of bumps across two networks connected via weak cortico-cortical excitation (Figure 2). In particular, we connected two ring-attractors in the regime described above with all-to-all, untuned excitatory connectivity. This connectivity was weak and it was mediated exclusively by AMPARs (Figure 2a), acting on all excitatory and inhibitory neurons. Interestingly, anti-phase dynamics within each network (as described above) was maintained robustly for a wide range of connectivity strength values (Figure 3). Across networks, each bump's activity was in phase with one bump in the other network (Figure 2b,c, black) but out of phase with the other (Figure 2b,c, red). On the majority of the simulations, this selective synchronization was maintained through the whole delay period (see Figure 2c,d for an example simulation). This dynamics is therefore interesting as a possible mechanism to maintain bound the information of each presented stimulus.

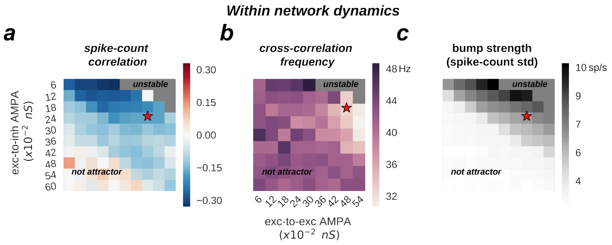

Figure 3: *Anti-correlated oscillatory dynamics as a function of excitatory recurrence (AMPAR conductance) in simulations with load 2.* a) Anti-phase dynamics within each network as measured by spike count correlation between bumps. b) Peak-frequency of power spectrum of the cross-correlation between the two bumps. c) Bump strength measured as spike-count variability at the end of the the delay. e-g) summarizes the dynamics of 22,000 simulations (total) of 100x2 networks.

Encoding/decoding through rate-code

In our simulations, synchronization selection was noise-induced, resulting in across-networks associations between random pairs of bumps for different simulations. To control this association at the time of stimulus encoding, we stimulated strongly and simultaneously 1 bump in each network for a brief period of 50ms (Figure 2b, and Figure 4a, green period), forcing these 2 bumps to engage in correlated activity during the delay period. Nevertheless, this phase-locked dynamics could be broken by noisy fluctuations, leading to possible misbinding of memorized features and swap trials (Figure 4a,b). Finally, our model raised the question of how this binding of information could reasonably be decoded without resorting to complex mechanisms for spike coincidence detection (Shadlen & Movshon, 1999). In our simulated task, the behavioral output, which consisted in answering which color was initially associated with a particular location, should depend on evaluating the pair of bumps in the 2 networks that maintained in-phase synchronization at the end of the delay. This was simulated as follows. For each trial, we probed one location by injecting external current to corresponding neurons in the location network at the end of the delay. This simulated the presentation of a location probe at the end of the delay, as typically done

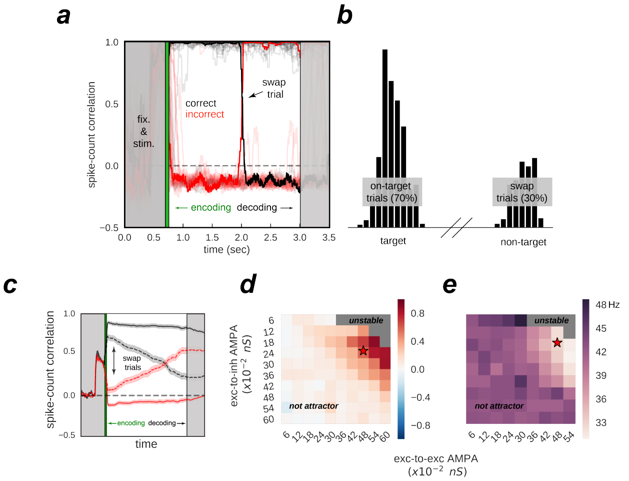

Figure 4: *Encoding and decoding is accomplished without temporal precision.* a) Spike-count correlation during the delay for 20 simulated trials. b) Histogram from 1000 trials. Bumps bound during encoding were more likely to be read-out than unbound bumps (target vs non-target). c) Same as a), averaging separately for swap and on-target trials, as defined by the decoder. d) Spike-count correlation between correct pairs. e) Peak-frequency of power spectrum of the cross-correlation between bound pairs, across networks.

in multi-item working memory tasks (Ma et al., 2014). This external current increased the firing rate in one of the location bumps, and we found that it also resulted in an increase of activity of the associated, in-phase color bump.

Finally, we extracted the behavioral output with a maximum likelihood decoder applied on mean firing rate activity of the color network during the last .5 s, while probing the corresponding location in the location network as described above. This algorithmic read-out could be replaced by a biologically plausible downstream network connected to the color circuit, and tuned to be in a winner-take-all regime - i.e. only able to maintain one bump at a time. Figure 4b shows 1,000 of such simulated trials. Applying our encoding/decoding method to our simulations, results in 30% of trials wrongly associated with the non-target color (swap trials, Figure 4c). We then separated swap trials from on-target trials and computed the spike-count correlation in windows of 5 ms through the whole trial period (Figure 4c), and confirmed that on-target trials were in fact characterized by stable phase-locked activity, while the correlation between bumps in swap trials progressively approached the opposite dynamics (in-phase/anti-phase for the bound/unbound items, Figure 4c). Together, our biologically-constrained simulations demonstrate that feature-binding can be accomplished through selective synchronization. Crucially, encoding/decoding location-color associations was done without a temporally precise code, a long-standing limitation in the binding by synchrony framework (Shadlen & Movshon, 1999).

Behavioral predictions: swap errors increase with delay (I) and item competition (II)

Swap errors have been described to increase with delay duration (Pertzov, Manohar, & Husain, 2017) and decrease with target to non-target distances (Schneegans & Bays, 2017; Emrich & Ferber, 2012). We therefore validated our feature-binding model against these behavioral findings. Firstly, Figure 5a shows that swap-errors increased with delay duration in the simulations. In our model, swap errors are induced by noisy fluctuations. Therefore, demanding longer delays will increase the probability of a large enough, swap-inducing noisy fluctuation. Secondly, Figure 5b shows how swap errors decrease with target to non-target distances, congruent with previous findings (Pertzov et al., 2017). For very close locations, feedback inhibition is strongest, leading to winner take all dynamics between nearby bumps, explaining an increase of swap errors for such distances. For intermediate distances, similarly to (Almeida, Barbosa, & Compte, 2015), simultaneous bumps interfere (repulsively and through their phase relationship, which is in this case less stable through the delay). Experimentally, these two regimes correspond to different scenarios. In the first case, one color is forgotten, while on the second scenario, there is an actual swap error. This prediction could be tested experimentally by probing subject's memory on all items, instead of just one (Adam, Vogel, & Awh, 2017). In sum, our model is able to describe a previously found dependence of swap errors with delay duration and with target to non-target distance, and it offers mechanistic explanations for such dependencies.

Neural prediction: swap trials show less phase preservation through the delay

Finally, abrupt changes in the phase relationship between oscillating bumps is the central mechanism of swap errors in our model (Figures 4a,b). Therefore, it is worth deriving a testable prediction from this mechanism. Additionally, because these changes in phase relationships are abrupt, they require experiments using high sampling-rate techniques such as MEG or EEG, rather than the slower BOLD signal that would smear out these events. We therefore applied an analysis that has been proposed to test phase consistency in EEG/MEG: the phase-preservation index (PPI, (Mazaheri & Jensen, 2006)). We then calculated the PPI at the end of the delay, relative to the beginning of the delay, and separately for on-target and swap trials. As we expected based on our model simulations (Figure 4), this analysis applied to our simulated field data showed that trials containing swap errors had a lower PPI, compared to on-target trials (Figure 6a). To test this prediction experimentally, we ran n=30 subjects on a multi-item working memory task. In this task, subjects had to report sequentially the location of all colored cues as prompted by successive color probes (Figure 6c). As subjects performed this task, we recorded MEG signals across the brain. When we compared symmetric swaps (reporting location B for color A AND reporting location A for color B) to all other trials, we found that

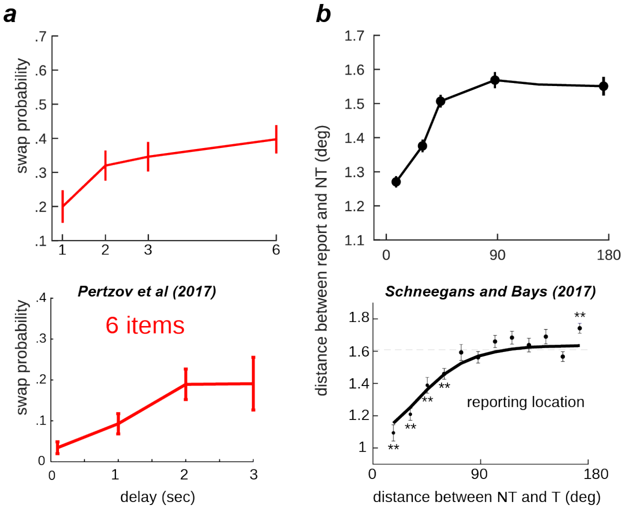

Figure 5: *Swap errors increase with delay duration and decrease with target-to-nontarget distances.* Model simulations (top) explain previous behavioral findings (bottom). a) Swap errors increase with delay duration and b) Simulations where target and non-target bumps are stored close-by increase swap errors, relative to when they are further apart.

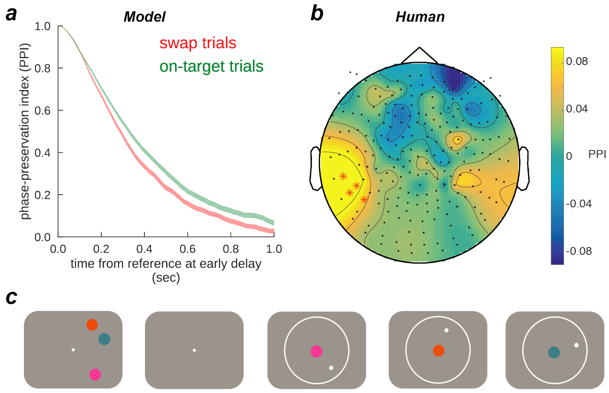

Figure 6: *Neural prediction validated in a human experiment* a) Swap-error trials (red), compared with on-target trials (green) in the model are associated with a lower phase consistency of oscillatory activity in the delay period, as measured with phase-preservation index (PPI) using early delay as the reference time point. b) In humans (n=30) performing the task depicted in c), PPI was significantly lower for swap trials compared to all other trials. This was specific to contro-lateral temporal sensors (marked with red stars, $p < 0.05$ cluster corrected).

swap-trials showed significant phase inconsistencies in the alpha band (10-13hz), as predicted by the model, and this was specific of contro-lateral temporal sensors (Figure 6c).

Conclusions

Aiming to account for swap-errors, a prominent source of multi-item working memory interference, we extended the classical bump-attractor model. Our biologically-constrained model offers a plausible mechanism for feature-binding through selective synchronization. Importantly, it explains when this feature-binding fails, including how it depends on delay duration and inter-item distances. Critically, we validated a strong prediction from its central underlying mechanism - phase-locked oscillatory activity during the memory periods - in humans performing a WM task.

References

Adam, K. C. S., Vogel, E. K., & Awh, E. (2017). Clear evidence for item limits in visual working memory. *Cognitive Psychology*, *97*, 79-97.

Almeida, R., Barbosa, J., & Compte, A. (2015). Neural circuit basis of visuo-spatial working memory precision: a computational and behavioral study. *Journal of Neurophysiology*, *114*(3), 1806-1818.

Compte, A., Brunel, N., Goldman-Rakic, P. S., & Wang, X. J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, *10*(9), 910-923.

Edin, F., Klingberg, T., Johansson, P., McNab, F., Tegnér, J., & Compte, A. (2009). Mechanism for top-down control of working memory capacity. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(16), 6802-6807.

Emrich, S. M., & Ferber, S. (2012). Competition increases binding errors in visual working memory. *Journal of Vision*, *12*(4).

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347-356.

Mazaheri, A., & Jensen, O. (2006). Posterior alpha activity is not phase-reset by visual stimuli. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(8), 2948-2952.

Pertzov, Y., Manohar, S., & Husain, M. (2017). Rapid forgetting results from competition over time between items in visual working memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *43*(4), 528-536.

Schneegans, S., & Bays, P. (2018). New perspectives on binding in visual working memory.

Schneegans, S., & Bays, P. M. (2017). Neural architecture for feature binding in visual working memory. *The Journal of Neuroscience*, *37*(14), 3913-3925.

Shadlen, M. N., & Movshon, J. A. (1999). Synchrony unbound: a critical evaluation of the temporal binding hypothesis. *Neuron*, *24*(1), 67-77, 111.