Achieving Robust Neural Representations: An Account Of Repetition Suppression

Michael C. Mozer
Dept. of Comp. Sci.
U. of Colorado
Boulder, CO 80309

Todd Mytkowicz
Dept. of Comp. Sci.
Colorado State U.
Ft. Collins, CO 80532

Richard S. Zemel
Dept. of Comp. Sci.
U. of Toronto
Toronto, ON M5S 1A4

Abstract

An important source of evidence concerning rapid adaptation and learning in the brain is the robust phenomenon of *repetition suppression*—the long lasting and item-specific decrease in neural activity with repeated exposure to an item, yielding sparser, sharper representations. Existing accounts of repetition suppression are informal and do little more than describe the phenomenon. We explore the hypothesis that repetition suppression arises from an unsupervised learning mechanism that reduces sensitivity to noise by increasing the item-specific gain of neural responses, in conjunction with the assumption that neurons are biased toward infrequent activity. This hypothesis explains key experimental observations concerning changes in neural representation with mere repetition of stimuli, regardless of task relevance. Additionally, this hypothesis explains related data concerning improved discriminability and noise robustness of individual neurons due to practice on a specific task.

A widespread and robust finding in primate electrophysiology is that neural responses decrease over repeated exposure to a stimulus (e.g., [2, 13, 16]). Decreased activation is also observed in human imaging studies (e.g., [11]) and a reduction of waveform amplitude is observed in ERP studies [14]. This *repetition suppression* (or *RS*) effect is often interpreted to reflect an increase in neural efficiency and likely mediates the psychological phenomenon of *repetition priming* [16, 11], in which prior presentation of a stimulus leads to more efficient processing of the same stimulus in the future. Repetition suppression is also referred to as *stimulus-specific adaptation* in the literature.

Key findings in the literature are as follows. (1) RS involves sharpening the neural representation, i.e., reducing overall neural activity and decreasing the number of neurons involved in representing an item [2, 13, 16]. Although fewer neurons are active for an item, a small fraction show an increase in response [13]. (2) RS is item specific, not general habituation [2, 13]. (3) RS is long lasting; it is observed even when the two repetitions are separated by 150 intervening items and delays up to 24 hours [6, 13]. (4) RS is graded, showing a continual reduction in firing rate with each presentation, plateauing at about half of the initial firing rate [6, 13]. (5) RS has been observed in many cortical regions, particularly inferior temporal and prefrontal cortex. However, RS is not omnipresent, e.g., it is not found in V1 [8], and it is found for words but not pseudowords in left posterior fusiform gyrus [3]. (6) RS depends merely on repetition, not behavioral significance [16].

This final characteristic—that RS occurs even with passive viewing, when no response is required and when the stimulus is not associated with a task or a reward—distinguishes RS

in principle from adaptations that occur with skill learning. However, Rainer and Miller [12] have reported intriguing similarities in a study of lateral prefrontal cortex. Using a delayed matching-to-sample paradigm, RS-like sharpening of representation was observed for familiar (i.e., used in the task over many sessions) versus novel stimuli. In addition to sharpening of representation, behavioral and neurophysiological measures showed: (1) familiar stimuli were more resistant to stimulus degradation, in the sense that individual neurons tuned to familiar stimuli showed discriminatory responses at higher levels of degradation than neurons tuned to novel stimuli; and (2) neurons tuned to a familiar stimulus showed a greater selectivity of response—i.e., tendency to respond to only that stimulus—than neurons tuned to novel stimuli. Although the mechanisms of adaptation underlying these effects may be unrelated to those giving rise to RS, a parsimonious account might be able to integrate the two sets of findings. We present such an account.

Existing accounts of repetition suppression are little more than descriptions of the data. Wiggs and Martin [16] and Desimone [2] suggest that cells unnecessary for identifying an item are suppressed, yielding a sparser and more selective representation, which presumably leads to a more efficient or rapid response. Ringo [13] proposes that suppression of familiar items may contribute to automatic orientation to novel items.

Why is repetition suppression important? First, the robustness and ubiquity of RS suggests it may be a fundamental mechanism of adaptation in neocortex. Second, RS has become a key tool for discovering the nature of cortical representations in neuroimaging (e.g., [4, 5, 9]), based on the following argument: If representations in a cortical region are invariant to some dimension of a stimulus, then RS should be observed in that region even when the stimulus repetitions differ along that dimension. For example, in the visual word form area, RS occurs even if the two presentations differ in case, indicating a case-invariant representation; and RS is observed in higher visual areas even if the stimulus varies in retinal size or location, indicating a representation that is somewhat transformation invariant. Because of the key role RS has come to play in cognitive neuroscience, it is important to develop a theoretical perspective that supports the methodology.

1 Blind Equalization

We propose an account of repetition suppression based on the innocuous assumption that a subset of cortical neurons encode intrinsically binary hypotheses. These *binary-hypothesis* neurons may have graded firing rates, but the firing rate indicates confidence in or probability of the truth of a hypothesis, not a continuous value (e.g., intensity or frequency). A neuron's firing rate is characterized as the fraction of its maximal rate, yielding a value in [0, 1]. A binary-hypothesis neuron signals "false" or "true" via a value near to 0 or 1.

Now consider a multilayer neural network whose outputs are binary-hypothesis neurons. The analog nature of the neural net readily allows noise to corrupt the firing rates of neurons, causing a low-confidence output (an output near .5) to flip its binary state from false to true or vice versa. For this reason, many digital communication systems that operate on underlying analog representations include an *equalization* process that attempts to undo the effects of noise and other distortions. If the desired output values of the neural net were known, the neural net could be trained via gradient descent to minimize the squared difference between the desired value for neuron i, d_i, and the actual value produced by the neural net, y_i [15]. Such training removes uncertainty in the net's output, making it more resistant to noise perturbations.

Equalization is feasible even in the absence of supervision by inferring the desired value from the actual value: $d_i = 1$ if $y_i > \theta$ or 0 otherwise, where typically $\theta = 0.5$. This scheme for *blind* equalization depends on the assumption that the corruption of the analog firing rate is sufficiently small that its binary counterpart can be recovered, albeit at the cost

of possibly losing information about confidence or probability. Thus, blind equalization trades off the ability to maintain gradations of certainty for noise robustness.

If the brain can be characterized as a noisy system—whether the noise is intrinsic to neural dynamics, due to integration of conflicting cues, or due to the failure of attention to suppress irrelevant inputs—then incorporating blind equalization is a sensible, adaptive strategy. Although we focus on this mechanism of unsupervised learning, we suppose that it serves to supplement, not replace, other supervised and reinforcement learning mechanisms that operate in parallel. Blind equalization conditions representations for noise robustness, whereas supervised and reinforcement learning mechanisms achieve transformations of representations that are useful for specific cognitive activities. Therefore, blind equalization should be broadly applied to all incoming stimuli regardless of their immediate behavioral relevance—a defining characteristic of RS.

Effectively, blind equalization turns up the *gain* of the neural response function. That is, consider a sigmoidal function relating a neuron's summed input to its firing rate, where the gain controls the steepness of the sigmoid. Low and high gain correspond to nearly linear and more step-like response functions. Other theorists have proposed brain mechanisms that dynamically modulate the gain of response functions (e.g., [1]). However, one distinct aspect of the present proposal is that the gain modulation is linked to the specific stimulus that was presented, as well as to highly similar stimuli. Thus, one can conceive of blind equalization as setting the gain on responses to a stimulus that increases with the frequency of recent encounters with the stimulus.

It is easy to get an intuition for why blind equalization leads to RS. Output neurons producing a strong response to the first stimulus presentation will increase their firing rates for a second presentation, whereas those producing a weak initial response will decrease their firing rates. Because neural codes in higher cortical areas such as IT and prefrontal regions appear to be sparse, more neurons will give a weak than strong response initially. Consequently, more neurons will decrease their activity than increase, and the overall effect is a decrease in activity. The noise robustness property described in [12] also naturally emerges from the model: Given a sigmoidal response function, noise input will have little influence on a neuron's output if the output is close to saturation (0 or 1).

2 Simulation Methodology

We model a biological neural net using the simple connectionist abstraction. Each unit in the connectionist net conveys a scalar activation level in [0, 1], interpreted as a mean firing rate relative to the unit's maximum firing rate. We explore fully layered, feedforward nets with logistic activation functions and complete connectivity between layers. In the simulations we report here, we use nets with a 100-unit input layer and a 100-unit output layer. Introducing hidden layers does not qualitatively affect the results. We study changes to the output representations produced by input patterns that are either repeated or not.

Weights are drawn from a Gaussian distribution, and then the weights into a given unit are rescaled to have an L1 norm of 2.0, which causes the unit to produce a range of outputs that do not saturate and typically lie in [.1, .9]. The initial weights are meant to reflect prior learning; however, the constraints and task experience which led to these weights is not considered to be relevant to the current simulation.

An independent variable of our simulations is the fraction, α, of the population that is typically highly active (having activation greater than 0.5) for a stimulus at the outset of the simulation. We obtain a specified α over the set of input patterns used in the simulation by selecting the bias weight of each unit such that a fraction α of patterns produce a response greater than 0.5. Because blind equalization pushes activity toward the extremes, the fraction does not change much over the sequence of stimulus presentations. (Learning adjusts

weights, not the representations directly. Consequently, a weight change in response to one input may affect the response to another input and slight changes to α may be observed.) Unless otherwise indicated, results we present are for $\alpha = 0.2$.

Input patterns used in the simulation are random binary vectors. We designate 10 vectors as the *repeated* inputs, and 10 as the *nonrepeated* inputs. The repeated inputs are presented in random order for blind-equalization training. This procedure is repeated a total of ten times, i.e., ten *training epochs*. After each epoch, we freeze the weights and present each of the input patterns—both repeated and nonrepeated—and compute statistics of the output patterns produced by the net. All results reported are based on averaging over 100 replications of the simulation.

During a training epoch, weights are adjusted following each presentation so as to perform gradient descent in $\sum_i (y_i - d_i)^2$, where i is an index over output units, y is the actual output, and d is the desired output, which equals the binary (thresholded) value. Rather than treating the learning rate as a free parameter, we constrained it by the following logic. Larger learning rates are preferable to smaller learning rates because they allow a single training experience to yield a more binary—and hence noise robust—representation. However, if learning rates are too large, weight changes for the current input could alter the responses to other inputs. Corrupting the binary-thresholded output representation of another input destroys prior learning; corrupted representations make subsequent stages of processing less effective. To satisfy the trade-off between maximizing noise robustness and minimizing cross-example interference, we search over learning rates to find the largest value that does not produce interference. For our simulation, we used a value of 0.2, although doubling and halving the value does not change the qualitative pattern of results.

3 Results

Figure 1a shows the mean output-unit activation as a function of training epochs. An activation decrease is evident for repeated stimuli but not for nonrepeated stimuli, consistent with the item-specific changes observed in RS. There is a slight drop for the nonrepeated stimuli, but it is of sufficiently small magnitude that even if it were present in experimental studies, it might not be noticed relative to the large RS drop. The nonrepeated stimulus suppression is due to the fact that occasionally nonrepeated stimuli are similar to repeated stimuli, and transfer of training in the neural network will be based on input similarity. (The expected angle between two random 100-dimensional binary vectors is 60.0°, with a standard deviation of 4.0°. In higher dimensional spaces, the expected angle does not change, but the standard deviation shrinks, for example, with 1000-dimensional vectors, the standard deviation is 1.28°. Repeating our simulation with 1000-dimensional vectors, we find much less suppression of nonrepeated stimuli, suggesting that the small drop for nonrepeated stimuli is an artifact of the low-dimensional input space.)

Figure 1a also provides evidence that RS is long lasting, as one would expect from a mechanisms that affects connectivity in a neural network. Because each training epoch involved presentation of all ten repeated stimuli in random order, two repetitions of the same stimulus were separated by between 0 and 18 intervening items. If RS were not long lasting, the intervening items would have wiped out the effect.

For repeated stimuli, the first repetition leads to the greatest drop in activity, as is observed in RS studies. The asymptotic activity decrease is about 42%. This decrease is not directly comparable to that observed in single-cell recording studies, because the simulated decrease is based on all neurons and the single-cell estimate is based on only the subset of neurons that are initially firing and whose activity drops over repetitions. Similarly, the simulated decrease cannot be compared to the percentage decrease in the BOLD response in imaging studies, because the model assumes that all neurons represent binary hypotheses, whereas they may be only a portion of the ensemble that make up the BOLD response.

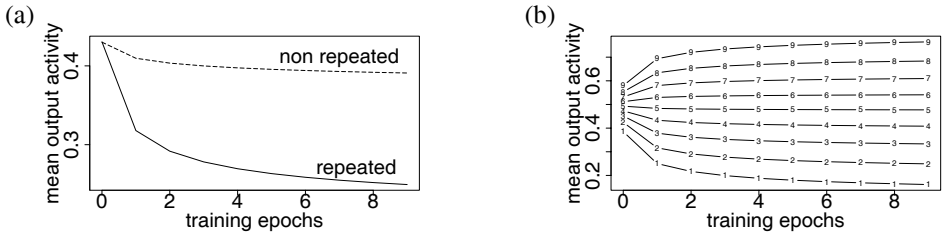

Figure 1: (a) mean activity of model's output layer as a function of training epochs for repeated and nonrepeated stimuli for $\alpha = 0.2$; (b) mean activity of model's output layer as a function of training epochs for repeated stimuli for α varying from 0.1 to 0.9, as indicated by the single digit labeling each line.

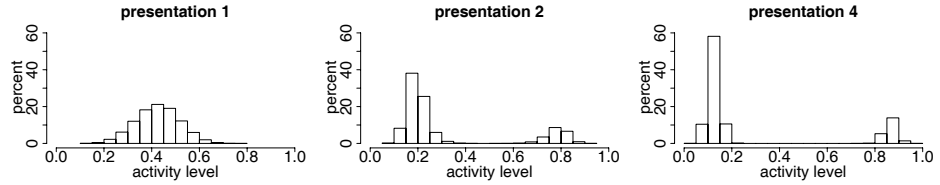

Figure 2: Distribution of output activity for repeated stimuli for the first, second, and fourth presentations of the stimuli.

Figure 2 completes the picture of representation sharpening in the model. The histograms show the distribution of activation for repeated stimuli on the first, second, and fourth presentations. On the first presentation, activation is spread over the [0 1] range in a normal distribution. On subsequent presentations, activation becomes bimodal with a majority of neurons decreasing in activity. Thus, the number of neurons involved in representing a stimulus is reduced. In addition, a minority of neurons show an increase in activity, as is found in the data. Ringo [13] reports that 80% of cells show a decrease and 20% show an increase. Our choice of $\alpha = 0.2$ in the model achieves the same distribution.

The effect of stimulus repetition depends on α, the fraction of output units that typically yield a strong response. Figure 1b shows the mean activity for repeated stimuli as a function of the number of repetitions and α, for α ranging from 0.1 to 0.9; the $\alpha = 0.2$ curve is the same as that in Figure 1a. As one would expect, RS occurs when $\alpha < 0.5$, and repetition *enhancement* occurs when $\alpha > 0.5$. Blind equalization produces an increase in activity for units that are highly active. If this group is the majority, then the net activity increase is larger than the net activity decrease, and the mean activity increases. The model thus identifies a key variable, α, that affects the sign and magnitude of repetition effects. Because neocortex tends toward sparse representations, one would not expect to find many brain regions with $\alpha > .5$. Nonetheless, α provides one handle on understanding why RS is observed in some cortical regions but not others. We return to this topic shortly.

Because blind equalization aims to reduce the influence of noise, we conjectured that our model could account not only for the basic phenomena of RS, but also for the noise-robustness findings of Rainer and Miller [12]. To explore whether repeated stimuli are more resistant to degradation than nonrepeated stimuli, we tested the network at various points in training with independent mean-zero Gaussian noise added to each input unit's activity. Note, however, that here again all training via stimulus repetition and blind equalization is noise free.

If each input were associated with a specific target output, we could use the distance be-

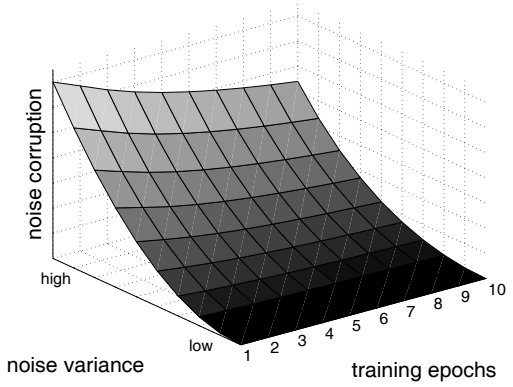

Figure 3: Noise corruption as a function of noise variance and training epochs

tween the actual output and the target as a measure of noise corruption. However, our training paradigm is unsupervised and there are no explicit targets. We might use the binary-thresholded outputs as targets, as we do for blind equalization, but there is an alternative target: the output produced by the noise-free input. This alternative target makes good sense because it yields a distance of zero if the outputs produced by noisy and noise-free inputs are the same, i.e., if input noise has no effect on the output. Clearly, when noise has no effect, subsequent stages of processing will be noise robust. Indeed, it would be ideal if the output produced by noise-free input could be used as target for blind equalization training, but obviously this information is not available. It is available for this test because we are using it to evaluate the model's output, not to train the model.

Figure 3 plots the noise corruption (squared distance between actual and target output) for repeated stimuli as a function of noise variance and number of training epochs. For low-noise stimuli, the output is reliable regardless of the number of repetitions, but for high-noise stimuli, the reliability of the output increases with repetitions. Blind equalization encourages output activities for repeated stimuli near the extreme values of 0 or 1, and due to the sigmoidal response function of a unit, noise on the input has diminishing effect on the output as the output approaches saturation. Thus, consistent with Rainer and Miller [12], our simulation finds that familiar (repeated) stimuli are more resistant to noise degradation than are novel (nonrepeated) stimuli. Our simulation offers the interpretation that familiarity may not depend on task relevance, but is due to mere repetition.

Rainer and Miller also assessed the *selectivity* of individual neurons using a measure that indicates how narrowly tuned a neuron is within a given stimulus set. This measure, denoted S_i for unit i, is $S_i = (n - \sum_j y_i^j / \max_i y_i^j)/(n-1)$, where j is an index over the stimulus set, n is the size of the stimulus set, and y_i^j is the response of a output unit i to stimulus j. The selectivity ranges from 0 to 1, where a value of zero means that the unit gives identical responses to all stimuli, and a value of 1 means that the unit gives a strong response to one stimulus and no response to any other. The selectivity of a neuron within the sets of familiar and novel stimuli is .656 and .402, respectively. This difference is statistically reliable (t(1,19998)=10991, $p < .001$, with replications and neurons as random factors), and shows the same pattern as the neurophysiological data [12].

4 Discussion

We explored the hypothesis that repetition suppression can be characterized in terms of a simple unsupervised learning mechanism, blind equalization, applied in a neural net-

work. Our model can account for the key phenomena of repetition suppression. Although repetition suppression is a widespread and robust phenomenon in neurophysiology and neuroimaging, it has not heretofore been explained in a coherent formal framework. Our model also provides an explanation for a related set of phenomena concerned with stimulus-specific improvement in noise robustness, discriminability, and selectivity following practice on a task. An important contribution of the model is its unification of these two sets of phenomena which might otherwise be considered to arise from distinct mechanisms.

4.1 When Will Repetition Suppression Be Observed?

As we noted earlier, RS is not observed in all cortical regions, and in regions where RS is observed, not all stimuli that activate the region produce RS. Our model offers a framework for understanding the conditions necessary for RS to occur. This framework is most useful for making predictions in neuroimaging studies, because measures of neural activity in these studies aggregate across all neurons in a brain region. It can also be useful in predicting the results of single-cell recording studies, specifically when cells are chosen randomly so as to be representative of the brain region as a whole.

According to the model, two conditions are critical for the occurrence of RS. First, RS should be observed only for binary-hypothesis neurons, because the unsupervised learning procedure is justified only for those neurons. Any cortical region is likely to have a mix of binary-hypothesis and other neurons, and RS effects will be of greater magnitude in regions with a higher proportion of binary-hypothesis neurons. Traditionally, early visual areas are thought to represent continuous quantities (wavelength, orientation, contrast), whereas higher visual areas are thought to encode more symbolic information. Thus, the proportion of binary-hypothesis neurons should increase along the ventral stream, and RS should increase in magnitude. Our account is therefore compatible with the finding that early visual cortical areas, e.g., V1, do not show RS [8], whereas in IT cortex, roughly one quarter to one third of IT neurons produce RS [13].

Second, the model shows RS effects only when fewer than 50% of neurons in a region are highly active (Figure 1b). Because neocortex tends toward sparse representations, few brain regions should violate this condition. Even when fewer than 50% of neurons are highly active, the model predicts that the magnitude of RS effects observed in neuroimaging studies will be inversely related to the fraction of highly-active neurons. This condition applies again to the early visual areas: because these areas encode low-level visual features, one would expect a more distributed representation and therefore, a larger fraction of neurons responsive to any stimulus compared to higher visual areas. This condition on obtaining RS is also compatible with the finding that words but not pseudowords produce RS in left posterior fusiform gyrus [3]. In an area specialized for words, one might argue that word-like but unfamiliar stimuli produce a more diffuse pattern of activation, and therefore a higher fraction of units active. Indeed, repetition of pseudowords produces a slight activation *enhancement*, as predicted by our model for highly distributed neural representations.

4.2 Future Work

We close with three extensions to the model that we are pursuing.

• The unsupervised learning algorithm we explored is but one of many possible algorithms for blind equalization. For example, one algorithm proposes generating graded desired response values that are the expectation of the binary values under a Gaussian noise distribution [10]. It may turn out that different variants of blind equalization yield different predictions, and that experiments can help us to further refine the model.

• Repetition priming is often viewed as the behavioral correlate of RS. Consequently, a coherent account of RS should also explain repetition priming. We believe our model can do so in the following way. Consider a variant of the artificial neural net in which each unit acts as a leaky integrator, resulting in gradual accumulation of information in the output units over time. If repetition leads to a faster convergence of output representations, then one would expect a priming effect—a more rapid response. We would expect our model to show this behavior because blind equalization tends to increase the weight magnitudes which would transmit information more rapidly to the leaky integrator neurons.

• We are examining the relationship between blind equalization and previously proposed neural mechanisms for increasing memory capacity. For example, the hippocampus may have a role in associative learning by projecting patterns into a high-dimensional space, making them more orthogonal and noise robust, and less subject to interference [7]. Although blind equalization does not explicitly attempt to push patterns apart, we predict that it will have achieve a similar effect. This prediction can be assessed by decoding the equalized representations.

Acknowledgements

This research was supported by NIH/IFOPAL R01 MH61549–01A1, and a CIHR NET Grant.

References

[1] Cohen, J. D., & Servan-Schreiber, D. (1992). Context, cortex and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review*, *99*, 45–77.

[2] Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. *Proceedings of the National Academy of Science*, *93*, 13494–13499.

[3] Devlin, J.T., Jamison, H., Gonnerman, L.M., & Matthews, P.M. (2004). The role of the left posterior fusiform gyrus in reading. Poster presentation at the *Cognitive Neuroscience Society* Conference, April 2004.

[4] Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., & Malac, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron*, *24*, 187–203.

[5] Kourtzi, Z., & Kanwisher, N. (2001). Representation of perceived object shape by the human lateral occipital cortex. *Science*, *293*, 1506–1509.

[6] Li, L., Miller, E. K., & Desimone, R. (1993). The representation of stimulus familiarity in anterior inferior temporal cortex. *Journal of Neurophysiology*, *69*, 1918–1929.

[7] Marr, D. (1971). Simple memory: a theory for archicortex. *Phil. Trans. Royal Soc. London*, *262*, 23–81.

[8] Murray, S.O., & Wojciulik, E. (2004). Attention increases neural selectivity in the human lateral occipital complex. *Nature Neuroscience*, *7*, 70–74.

[9] Naccache, Dehaene (2002). Naccache, L., & Dehaene, S. (2001). The priming method: Imaging unconscious repetition priming reveals an abstract representation of number in the parietal lobes. *Cerebral Cortex*, *11*, 966–974.

[10] Nowlan, S.J., & Hinton, G.E. (1993). A soft decision-directed LMS algorithm for blind equalization. *IEEE Trans. Comm.*, *41*, 275–279.

[11] Poldrack, R.A., & Gabrieli, J.D. (2001). Characterizing the neural mechanisms of skill learning and repetition priming. Evidence from mirror reading. *Brain*, *124*, 67–82.

[12] Rainer, G., & Miller, E.K. (2000). Effects of visual experience on the representation of objects in the prefrontal cortex. *Neuron*, *27*, 179–189.

[13] Ringo, J. L. (1996). Stimulus specific adaptation in inferior temporal and medial temporal cortex of the monkey. *Behavioral Brain Research*, *76*, 191–197.

[14] Rugg, M.D., Soardi, M., & Doyle, M.C. (1995). Modulation of event–related potentials by the repetition of drawings of novel objects. *Cognitive Brain Research*, *3*, 17–24.

[15] Widrow, B., & Hoff Jr., M.E. (1960). Adaptive switching circuits. In *IRE WESCON Convention Record, part 4*, 96–104.

[16] Wiggs, C.L., & Martin, A. (1998). Properties and mechanisms of perceptual priming. *Current Opinion in Neurobiology*, *8*, 227–233.