Behavioral/Systems/Cognitive

A Neuronal Model of Predictive Coding Accounting for the Mismatch Negativity

Catherine Wacongne,[1,2,3] **Jean-Pierre Changeux,**[4,5] **and Stanislas Dehaene**[1,2,3,5]

[1]Institut National de la Santé et de la Recherche Médicale, Unité 992, Cognitive Neuroimaging Unit, and [2]Commissariat à l'Energie Atomique, DSV/I2BM, NeuroSpin Center, F-91191 Gif/Yvette, France, [3]University Paris 11, F-91405 Orsay, France, [4]Pasteur Institute, Centre National de la Recherche Scientifique Unité de Recherche Associée 2182, F-75015 Paris, France, and [5]Collège de France, F-75005 Paris, France

The mismatch negativity (MMN) is thought to index the activation of specialized neural networks for active prediction and deviance detection. However, a detailed neuronal model of the neurobiological mechanisms underlying the MMN is still lacking, and its computational foundations remain debated. We propose here a detailed neuronal model of auditory cortex, based on predictive coding, that accounts for the critical features of MMN. The model is entirely composed of spiking excitatory and inhibitory neurons interconnected in a layered cortical architecture with distinct input, predictive, and prediction error units. A spike-timing dependent learning rule, relying upon NMDA receptor synaptic transmission, allows the network to adjust its internal predictions and use a memory of the recent past inputs to anticipate on future stimuli based on transition statistics. We demonstrate that this simple architecture can account for the major empirical properties of the MMN. These include a frequency-dependent response to rare deviants, a response to unexpected repeats in alternating sequences (ABABA$\underline{A}$. . .), a lack of consideration of the global sequence context, a response to sound omission, and a sensitivity of the MMN to NMDA receptor antagonists. Novel predictions are presented, and a new magnetoencephalography experiment in healthy human subjects is presented that validates our key hypothesis: the MMN results from active cortical prediction rather than passive synaptic habituation.

Introduction

Since it was first described at the end of 1970s, the mismatch negativity (MMN) has been largely used in theoretical and clinical research (for review, see Näätänen, 2003). It was first recorded by EEG in the context of the oddball paradigm. In the most frequently used version of this paradigm, participants are instructed to listen to repeated occurrences of one sound, called the standard. This monotony is disrupted at rare moments by the presentation of a different sound, called the deviant. The difference in the responses evoked by deviants and standards takes the form of a broadly negative waveform at the top of the scalp, which peaks between 100 and 200 ms after the onset of the sound. MMNs can be elicited by differences in sound frequency, duration (Näätänen et al., 1989), amplitude (Näätänen et al., 1987), or interstimulus interval (ISI) (Ford and Hillyard, 1981). MMN is resistant to manipulations of attention and states of wakefulness (Sculthorpe et al., 2009) even though these parameters can modulate its amplitude. An analog of MMN was described in visual (Tales et al., 1999; Pazo-Alvarez et al., 2003), olfactive (Krauel et al., 1999; Pause and Krauel, 2000), and somatosensory (Kekoni et al., 1997; Shinozaki et al., 1998) modalities, supporting a broad computational significance of MMN as a shared and automatic brain mechanism responsive to stimulus novelty.

MMN is frequently interpreted in terms of predictive coding (Rao and Ballard, 1999; Lee and Mumford, 2003), assuming that the brain does not respond passively to incoming inputs but learns the inputs regularities and uses that knowledge to actively predict what should happen next. The auditory system would acquire an internal model of regularities in auditory inputs, including abstract ones, that are used to generate weighted predictions about the incoming stimuli (Paavilainen et al., 1999; Näätänen et al., 2005; Winkler, 2007). If these predictions differ from the actual stimulus, it results in a mismatch signal.

While mathematical models of predictive coding have been proposed (Garrido et al., 2007; Kiebel et al., 2008, 2009), including some attributing distinct functions to the various cortical layers (Friston, 2005), none of them has yet led to a precise neuronal implementation of the generators of the MMN, in terms of realistic receptors, synapses, and spiking neurons. Nor has there been a systematic comparison of the predictions of the models with actual experimental results. Furthermore, not everyone accepts the predictive interpretation of MMN. May and Tiitinen (2010) argue that synaptic habituation (reduction of the EPSP following repetitive stimulation of the same synapse) is sufficient to explain all of the properties of the MMN and, thus, that there is no need to postulate an elaborate prediction and comparison mechanism.

Here, we propose a neuronal network model, devoid of synaptic habituation but comprising a detailed implementation of predictive coding, accounting for a large amount of data on the MMN. The model leads to the distinction of several processes that contribute to the ob-

Received Sept. 30, 2011; revised Dec. 12, 2011; accepted Jan. 4, 2012.

Author contributions: C.W., J.-P.C., and S.D. designed research; C.W. performed research; C.W. analyzed data; C.W., J.-P.C., and S.D. wrote the paper.

This work was supported by a senior grant of the European Research Council (NeuroConsc program). The NeuroSpin magnetoencephalography facility was sponsored by grants from Institut National de la Santé et de la Recherche Médicale, Commissariat à l'Energie Atomique, Fondation pour la Recherche Médicale, The Bettencourt–Schueller Foundation, and Région île-de-France. We thank Virginie Van Wassenhove, Alain Destexhe, and Karim Benchenane for useful discussions.

Correspondence should be addressed to Catherine Wacongne, Commissariat à l'Energie Atomique/SAC/DSV/DRM/NeuroSpin Center, Bât 145, Point Courier 156, F-91191 Gif-sur-Yvette Cedex, France. E-mail: catherine.wacongne@gmail.com.

DOI:10.1523/JNEUROSCI.5003-11.2012

0270-6474/12/323665-14$15.00/0

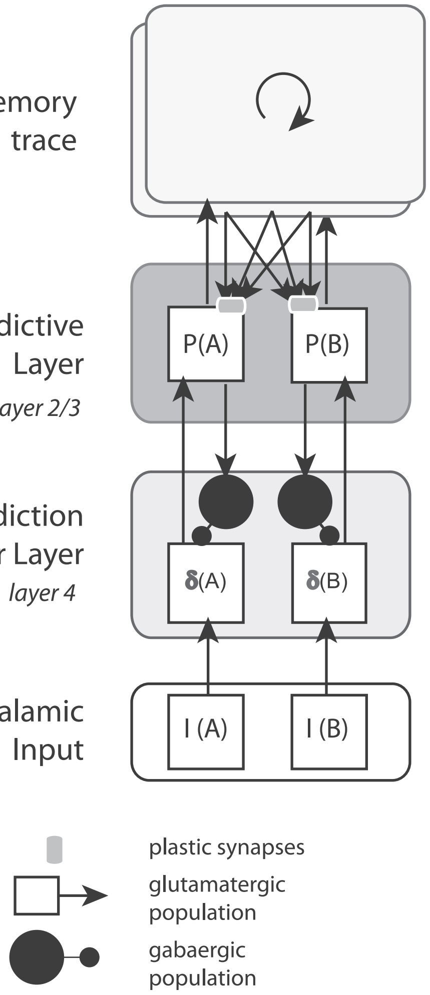

Figure 1. Scheme of the predictive coding model for two sounds. For each layer, two subpopulations are modeled that respond respectively to the frequencies of sounds A and B. Prediction error activity in layer 4 is the result of the difference between thalamic inputs and predictive activity arising from the supragranular layer, whose sign is inverted through inhibitory interneurons (black circles). Prediction error is then fed back to adjust the activity of predictive populations. Dynamic predictions are made possible in the model because predictive units send and receive projections with a recurrent network serving as a short-term memory. NMDA-dependent plasticity adjusts the synaptic weights onto predictive units until their dynamics matches that of the inputs and therefore minimizes the prediction error.

served event-related responses, and makes new predictions, one of which is tested here with magnetoencephalography (MEG).

Materials and Methods

Network architecture

The proposed neuronal network aims at modeling the response of primary auditory cortex to incoming sounds. Figure 1 shows an implementation of the model for an input composed of two pure tones, hereafter called A and B. Each column of the network represents a cortical column with its thalamic input responding maximally to one of the two frequencies of the input. The two frequencies A and B are supposed to be different enough to activate only one of the two columns.

In each column, three populations of neurons are simulated. The essential component of the model is the population of neurons involved in prediction, which we propose to be part of the supragranular layers of the cortex. This population constantly tries to anticipate the upcoming auditory inputs. A prediction of sound A consists in an increase in the population firing rate coding for this stimulus.

At every moment, the continuously variable predictions arising from the predictive populations of neurons are compared with the incoming inputs. This comparison is achieved at the level of a population of neurons called the "prediction error" population, which receives two sets of inputs: excitatory inputs coming from the thalamus and conveying the current sensory stimulus, and inhibitory inputs that reflect the activity of the predictive population. Through this scheme, whenever the thalamic input is not cancelled by predictive signals, the prediction error population fires. The activity of the prediction error population is transmitted to the predictive population as a feedback and this error signal is used to adapt the internal model of this population (see the description of the learning rule further below). We show in Results that this error signal may account for the MMN effect.

The predictive population needs to build an internal model of the regularities of the incoming stimulus to form relevant predictions. We propose that this model is based on learning the statistical temporal dependencies linking the stimuli within the past few hundred milliseconds. A memory of the recent past is needed to achieve such a goal. This memory has to keep the trace of two properties: the identity of the past inputs and the time elapsed since they occurred. We choose to model this function in the simplest manner possible, using a delay line for each frequency, where activation propagates linearly from one neuron to the next as a function of time. The relevance of this model will be discussed later.

Memory neurons are connected to both predictive subpopulations so that predictions of one frequency (A) can be based on the recent occurrence of a sound of the other frequency (B). The internal model of the predictive population is built by adapting the synaptic weights linking the memory neurons and the predictive populations.

Detailed implementation

All subpopulations are composed of 40 neurons, except for delay lines that are composed of 400 excitatory neurons and 100 inhibitory neurons. All populations receive an external input I_{ext} that is Gaussian noise of mean equal to 0 and variance equal to 2.5 for input neurons, 2 for predictive neurons and prediction error neurons, 3.8 for interneurons.

By default, mean synaptic weight between two excitatory neurons is $w_{\text{EE}} = 1.4$, between an excitatory and an inhibitory neuron $w_{\text{EI}} = 4.5$, and between an inhibitory and excitatory neuron $w_{\text{IE}} = 22$. If a presynaptic neuron is excitatory, w_{EI} or w_{EE} is the weight for AMPA-mediated currents. An NMDA receptor (NMDAr)-dependent current is added whose weight w_n is 20% of the AMPA synapse. The synaptic weights are drawn from a Gaussian distribution with a variance of 20% of the mean. These parameters allow a reliable transmission of activity from one population to the other in absence of other inputs, while avoiding unrealistic synchrony of neurons due to excessive homogeneity in the parameters.

The probability of a connection between thalamic inputs and prediction error populations is $p = 0.9$. The probability of a connection between predictive populations and interneurons and between interneurons and prediction error neurons, is $p = 0.55$. Synapses between predictive populations and memory neurons were initialized with weight $w = 0.4$ and variance of 20% with a probability of connection of 0.5. Connectivity between layers is consistent with neocortical local circuitry data (Thomson and Lamy, 2007).

Spiking neuron model

We used spiking neurons whose membrane potential is computed according to the following Izhikevich (2003) equations:

$$\frac{dv}{dt} = 0.04v^2 + 5v + 140 - u + I_{\text{syn}}$$

$$\frac{du}{dt} = a(bv - u),$$

where v is dimensionless and represents the membrane potential and u is a membrane recovery variable. The neurons fire if their membrane potential reaches 30 mV and is then reset as follows:

$$\text{if } v \geq 30 \text{ mV}, \quad \text{then}\begin{cases} v \leftarrow c \\ u \leftarrow u + d. \end{cases}$$

The parameters for excitatory (respectively, inhibitory) neurons were as follows: $a = 0.02$ (respectively, $0.06 + 0.04 * \text{rand}^2$), $b = 0.2 + 0.04 * \text{rand}^2$ (respectively, 0.2), $c = -65 + 10 * \text{rand}^2$ (respectively, −65), $d = 8 - 2 *$

rand^2 (respectively, 2), where rand is a random number drawn from a uniform distribution between 0 and 1. These parameters correspond respectively to regular spiking neurons for excitatory neurons and fast spiking ones for inhibition (Izhikevich, 2003).

AMPA, NMDA, and GABA synaptic currents are modeled according to Brunel and Wang (2001) as follows:

$$I_{\mathrm{syn}}(t) = I_{\mathrm{AMPA}}(t) + I_{\mathrm{NMDA}}(t) + I_{\mathrm{GABA}}(t) + I_{\mathrm{ext}}(t)$$

with

$$I_{\mathrm{AMPA}}(t) = g_{\mathrm{AMPA}}(v(t) - V_E) \sum_{j=1}^{C_E} w_j^{\mathrm{AMPA}} s_j^{\mathrm{AMPA}}(t)$$

$$I_{\mathrm{NMDA}}(t) = \frac{g_{\mathrm{NMDA}}(v(t) - V_E)}{(1 + [\mathrm{Mg}^{2+}]\exp(-0.062v(t)/3.57))} \times \sum_{j=1}^{C_E} w_j^{\mathrm{NMDA}} s_j^{\mathrm{NMDA}}(t)$$

$$I_{\mathrm{GABA}}(t) = g_{\mathrm{GABA}}(v(t) - V_i) \sum_{j=1}^{C_I} w_j^{\mathrm{GABA}} s_j^{\mathrm{GABA}}(t),$$

where $V_E = 40$ and $V_i = -80$. The dimensionless weights $w_j^{\text{receptor type}}$ represent the strength of synaptic connection associated with each receptor type. The sum over j is the sum over all (C_E) excitatory or (C_i) inhibitory presynaptic neurons. $g_{\text{receptor type}}$ are dimensionless variables that represent the conductances of each receptor type with $g_{\mathrm{AMPA}} = 7.5{*}10^{-3}$, $g_{\mathrm{NMDA}} = 2{*}10^{-3}$, and $g_{\mathrm{GABA}} = 7.5{*}10^{-3}$; $[\mathrm{Mg}^{2+}] = 10^{-3}$. $s_j^{\text{receptor type}}$ is a variable describing the opening dynamic of the receptors: AMPA and GABA receptors have instantaneous opening and close up with time constants $\tau_{\mathrm{AMPA}} = 2$ ms and $\tau_{\mathrm{GABA}} = 10$ ms, as follows:

$$\frac{ds_j^{\mathrm{AMPA}}(t)}{dt} = \frac{s_j^{\mathrm{AMPA}}(t)}{\tau_{\mathrm{AMPA}}} + \sum_k \delta(t - t_j^k)$$

$$\frac{ds_j^{\mathrm{GABA}}(t)}{dt} = \frac{s_j^{\mathrm{GABA}}(t)}{\tau_{\mathrm{GABA}}} + \sum_k \delta(t - t_j^k).$$

where the sum over k represents a sum over spikes emitted by presynaptic neuron j. NMDA receptors have slower dynamics with opening time constant $\tau_{\mathrm{NMDA,rise}} = 2$ ms and closing time constant $\tau_{\mathrm{NMDA,decay}} = 100$ ms, $\alpha = 0.5\ \mathrm{ms}^{-1}$, as follows:

$$\frac{ds_j^{\mathrm{NMDA}}(t)}{dt} = \frac{s_j^{\mathrm{NMDA}}(t)}{\tau_{\mathrm{NMDA,\ decay}}} + ax_j(t)(1 - s_j^{\mathrm{NMDA}}(t))$$

$$\frac{dx_j(t)}{dt} = \frac{x_j(t)}{\tau_{\mathrm{NMDA,\ rise}}} + \sum_k \delta(t - t_j^k).$$

Synaptic plasticity

To internalize the statistical regularities that relate past activity to present stimuli, we implemented synaptic plasticity only between memory neurons and predictive subpopulations. We used a spike timing-dependent plasticity (STDP) rule (Bi and Poo, 1999) producing conditioning association as follows:

If a postsynaptic spike at time t follows a presynaptic spike:

$$\Delta w_{\mathrm{pre,post}} = c_p(I_{\mathrm{ca}^{2+}} - Th)\exp\left(\frac{t - t_{\mathrm{spike\ pre}}}{\tau_p}\right).$$

If a presynaptic spike follows a postsynaptic spike that occurred at time t:

$$\Delta w_{\mathrm{pre,post}} = c_p(I_{\mathrm{ca}^{2+}} - Th)\exp\left(\frac{t - t_{\mathrm{spike\ post}}}{\tau_p}\right).$$

In addition, we used a long-term depression rule, which induces a small depression of synapses whenever the presynaptic neuron spikes. This rule is in agreement with experimental observation that synapses tend to depress when they do not elicit postsynaptic spike (Debanne et al., 1998) as follows:

$$\Delta w_{\mathrm{pre,post}} = -c_d\,\delta(t - t_{\mathrm{spike\ pre}}).$$

The parameters used for the simulations presented in this paper are as follows: $c_p = 60$, $\tau_p = 30$ ms, $c_d = 100$, and $Th = 2.5$.

We verified that our qualitative results were largely independent of the fine tuning of the parameters. $I_{\mathrm{Ca2+}}$ is a calcium current mediated by NMDA receptors. This current is taken equal to I_{NMDA} for each predictive neuron.

Simulations

For each simulation, a new network was generated following the above probabilistic connectivity rules. Each condition was simulated on 5–10 different networks; plotted results are averages over all simulations. Inputs were an additional I_{ext} current with amplitude 1.9, injected in the thalamic subpopulation coding for the sound corresponding to the stimulus presented. The input for each simulation was created by pseudorandomization of a set of trial containing the desired proportions of standard and deviant stimuli. The randomization was made so that two deviants were never consecutive. Standard stimuli immediately following deviant stimuli were removed from analysis.

Various paradigms were simulated by modifying the sequence of A and B inputs in different stimulus blocks. The classical oddball paradigm was simulated as a sequence of 2000 tones, where 5, 10, 20, or 30% of the tones were B tones (deviants) and other tones were A, with a stimulus onset asynchrony (SOA) of 200 ms. The connectivity matrix was saved after each tone, 100 ms after the onset of the tone. The mean connectivity matrix that we report in Figure 4 represents the average connection strength between the memory neurons and the predictive population. It was obtained by averaging these matrices over each subpopulation of predictive neurons and over all tones except the first 200. Alternate sequences were composed of 1500 pairs of alternating tones (ABAB. . . ; ISI = 200 ms). The reproduction of the local-global paradigm (Bekinschtein et al., 2009; Wacongne et al., 2011) was made by starting with 20 standard sequences (100% AAAAB; ISI = 150 ms) followed by 100 sequences comprising 70% standards (AAAAB), 20% deviants (AAAAA), and 10% omissions (AAAA). For the omission effect, a simulation of 1500 pairs of sounds (AA; ISI = 200 ms) was also performed, with 10% of pairs replaced by single tones (A). We compared this with the response to 500 single tones (A).

MEG experiment

Participants. Five healthy volunteers (three males, two females; mean age, 22) with no neurological or psychiatric problems were studied. All participants gave their written informed consent to participate to this study, which was approved by the local ethical committee.

Auditory stimulation. Pairs of 50-ms-duration sounds were presented via headphones with an intensity of 45 dB and 200 ms SOA between sounds. Each sound was a pure sinusoidal tone (either 800 Hz, low; or 1600 Hz, high).

Sounds were organized in two blocks. In each block, the frequent pair, comprising two distinct sounds (AB), was first presented 10 times, with 1 s SOA between pairs. A total of 120 pairs was then presented, with SOA varying between 10 and 20 s, and with 70% of frequent AB pairs, 10% of rare AA pairs, 10% of rare BB pairs, and 10% of rare BA pairs. The identity of the A and B tones was swapped between blocks. The pairs were pseudorandomized so that two rare pairs were never consecutive. Frequent pairs following immediately a rare pair are excluded from the analysis. All stimuli were presented using E-prime software, version 1.1 (Psychology Software Tools).

MEG/EEG recordings. Measurements were performed with the Elekta Neuromag MEG system (Elekta Neuromag Oy) installed at the NeuroSpin center (Saclay, France), which comprises 204 planar gradiometers and 102 magnetometers in a helmet-shaped array. ECG as well as EOG (horizontal and vertical) were simultaneously recorded as auxiliary channels. MEG and auxiliary channels were low-pass filtered at 330 Hz, high-pass filtered at 0.1 Hz, and sampled at 1 kHz. The head position with respect to the sensor array was determined by four head position indicator coils attached to the participant's scalp. The locations of the coils and EEG electrode positions were digitized with respect to three anatomical landmarks (nasion and preauricular points) with a 3D digitizer (Polhemus Isotrak system). Then, head position with respect to the device origin was acquired before each MEG/EEG recording session.

Each participant was recorded for 1 h, 15 min: two sessions of ~33 min duration separated by a short resting period. Participants were asked to keep their eyes open and to avoid eyes movements by staring at a fixation cross. Participants were instructed to pay attention to the auditory stimuli. Importantly, although subjects were attending to the stimuli, which may generate additional attention-dependent components such as N2b, these components typically do not contribute to MEG signals (Alho et al., 1998). At the end of the recording, a question list was submitted to the participant. This list aimed to determine which regularities the participant was able to report after recording.

Postprocessing. Artifacts arising from outside the sensor array, such as those stemming from limb movement or other ambient magnetic disturbances, were greatly reduced by the signal space separation method (SSS) (Taulu et al., 2004). Gradiometers and magnetometers with amplitudes continuously exceeding 3000 fT/cm^2 and 3000 fT, respectively, were set as bad channels and excluded from further analysis. SSS correction, head movement compensation, and bad channels correction were applied using the MaxFilter Software (Elekta Neuromag).

A principal-component analysis (PCA) was used for PCA-based removal of EEG and EOG artifacts. Signal was averaged around artifacts for each channel type (EEG, axial and longitudinal gradiometers, and magnetometers) and a PCA was performed. Main components were saved.

The rest of the preprocessing was performed using Fieldtrip software (http://fieldtrip.fcdonders.nl/). Trials were epoched for each trial type between 200 ms before and 800 ms after the onset of the first sound. A low-pass filter at 40 Hz was applied and PCA correction of cardiac and EOG artifacts was performed using the PCA components previously computed. The trials were baseline corrected using the first 200 ms of the epoch.

After visual rejection of jump and pronounced trend artifacts, the data were averaged per condition and per participant. The latitudinal and longitudinal gradiometers were combined by computing the mean square root of signal at each sensor position.

Statistics. Statistics were performed using Fieldtrip cluster-based statistics. To examine differences between experimental conditions, paired t tests were performed with a threshold set at $p = 0.05$. Significant samples were clustered in connected sets on the basis of temporal and spatial proximity. Cluster statistics were calculated by taking the sum of t values in every cluster. To obtain a p value corrected for the size of the search space (time X sensors), a Monte Carlo method was used to evaluate how extreme the cluster statistics of the two conditions were compared with random partitions of the samples. The proportion of random partitions that resulted in larger cluster statistics than the observed one was the p value. The threshold was fixed to corrected $p = 0.05$.

Statistics on the difference between the frequent AB condition and the rare AA condition were computed between 0 and 300 ms after the onset of the second sound.

Response amplitude. The amplitude of the response to each of the two tones was defined as the average response over all magnetometers in the time window of the peak response for each sound (i.e., between 95 and 125 ms after the onset of the first tone and between 135 and 160 ms after the onset of the second tone). The amplitudes were normalized for each subject by the response to the first sound averaged over all conditions.

Results

Oddball paradigm and MMN

We first simulated the response of the network to the classical oddball paradigm. For this simulation, the network received as inputs two stimuli A and B, corresponding to sounds of frequencies distant enough to activate nonoverlapping populations of neurons. The input neurons were supposed to be selective only to the onset of the sound and were thus stimulated by an extra input current on input populations during 10 ms. The first stimulus ("sound A") was presented most of the time (standard tone), and the other one ("sound B") more rarely, with a parametrically variable frequency (deviant tone).

The left panels of Figure 2 show the response to the standard and deviant tones, averaged over all analyzed presentations, in the specific case in which the deviant has a 10% probability of occurrence. One can immediately observe that both the firing rates and the synaptic currents of the prediction and prediction error neurons (but not the sensory neurons) are higher on deviant than on standard trials. The detailed neuronal mechanisms of this mismatch effect are the following. First, note that the predictive population coding for the sound A starts firing shortly before the occurrence of both standard and deviant sounds (top panel, red curve). This activity originates from the EPSCs coming from the memory neurons: the network predicts the forthcoming occurrence of a sound A. This activity inhibits the prediction error layer via an interneuron population. If a sound A is actually presented, it cancels most of the excitation coming from thalamic inputs, resulting in a minimal prediction error response. As seen in Figure 2, only a small proportion of prediction error neurons still fire on standard trials, primarily due to stochastic fluctuations in the onset and strength of delay and predictive neurons, which therefore fails to fully cancel the incoming signal. On the contrary, when a deviant sound B is presented, the prediction of an A sound does not cancel the input for a B sound. This results in a large prediction error response, which is relayed to the predictive subpopulation coding for B to adapt the predictive model. It forces the neurons of the predictive layer to discharge and causes a large NMDAr-dependent current that results in NMDAr-dependent plasticity. This plasticity leads to an adaptation of the internal model of the network, reinforcing the synapses coming from the delay lines that discharged just before the prediction error signal.

The MMN is the result of a subtraction of the event-related potentials (ERPs) to standard and deviant stimuli. The ERPs are believed to be the result of a weighted integration of postsynaptic currents. As a simplified proxy for local field potentials or EEG responses, we calculated the difference in the sum of currents received by each layer for standard or deviant sound. The third column of Figure 2 shows the result of that operation. We can observe that there is indeed a difference in the currents between the two stimuli. For convenience, we will call this analog of the experimental phenomenon the simulated MMN or sMMN.

Behavior of the memory neurons

The memory neurons play an important role in the model. The stimulation of the network results in the activation of the predictive population either because the incoming stimulus is predicted or because of the transmission of prediction error. When the predictive population is active, it triggers the set of delay-line neurons (Fig. 3). The activity propagates linearly in the population, such that there is a direct relationship between the indices of the neurons in the delay line and the temporal information coded by their activity. The precision of timing changes as a function of the interval coded: the jitter in the exact time of activation of the neurons increases with the delay coded (approximating Weber's law). Essentially, the activity of a neuron in a delay line codes for two properties of past inputs: the identity of a past stimulus and the time elapsed since the occurrence of that stimulus. The particular choice we made for the implementation of this double function (delay lines) is not fully physiologically realistic but was

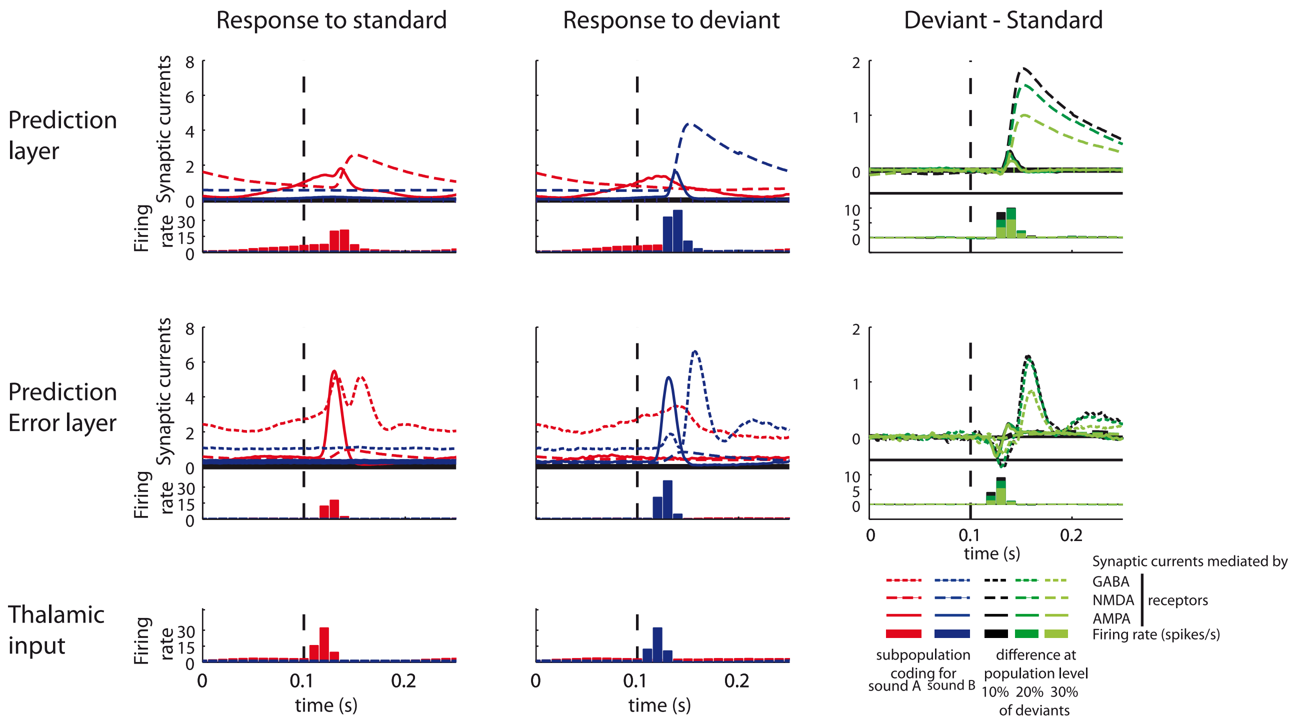

Figure 2. Simulating the MMN in an oddball paradigm: mean synaptic currents and firing rates. The figure shows the mean simulated response to a standard tone (first column), a deviant tone (second column), and their difference (third column) after 200 learning trials in an oddball paradigm. Each line shows the response of a different layer of units in the model (organized as in Fig. 1). For each layer, the top part of the plot represents the synaptic currents received by the subpopulation, separately for the different types of postsynaptic receptors that mediate these currents: AMPA (continuous line), NMDA (dashed line), or GABA (dotted line). The bottom part of each plot displays the mean firing rate of each subpopulation. In the first and second columns, subpopulations responding to the frequent A sound (90% of trials) are represented in red, and those responding to the rare B sound (10%) in blue. The third column shows the results of simulations in which the percentage of deviants was varied (10, 20, or 30%).

made for the sake of clarity and computational economy (see Discussion).

Layer distribution of current sources

We proposed a tentative localization for each functional population within the cortical layers, according to which prediction error populations correspond to granular layer and predictive populations belong to supragranular layer. Javitt et al. (1996) provided relevant intracortical local field potential data on the cortical origins of the MMN in primates. They showed in particular that the MMN mainly originates from supragranular layers of the cortex. The results of our simulations are consistent with these data, as they show that the sMMN primarily originates from synaptic currents impinging upon prediction neurons (and arising from prediction error neurons). Importantly, note that, even though there is a major difference in the firing rate of the prediction error population between the two stimuli, it does not involve a difference in the sum of synaptic inputs received by this layer as a whole, but rather a different distribution of these inputs on neurons coding for sounds A and B.

Studies in mice (Ehrlichman et al., 2008), rats (Tikhonravov et al., 2008, 2010), and monkeys (Javitt et al., 1996) also showed that MMN is strongly affected by NMDAr inhibitors. In our simulations, the sMMN results essentially from NMDAr-dependent currents, which is consistent with this observation.

Effect of deviant probability

The vast literature on the MMN describes a broad set of properties (for review, see Näätänen et al., 2007). To evaluate the range of validity of this model, we next simulated the response of the model in various conditions mimicking classical experimental paradigms. Our first test concerned the effect of the proportion of deviants in the standard oddball paradigm. Sato et al. (2000) described a systematic and parametric dependency of MMN amplitude on the probability of occurrence of a deviant sound. They showed that amplitude of the MMN increases as the frequency of the deviants decreases. We simulated the network for various proportions of deviant in the oddball paradigm (10, 20, and 30%). Results are plotted in the third column of Figure 2. We can see that the amplitude of sMMN indeed increases with the rarity of the deviants. This reduction in sMMN comes from the increased activity of the predictive population coding for B, as a result of its more frequent occurrence after an A, combined with a slightly lower prediction of the A sound that increases the average prediction error to A. This finding closely matches the experimentally recorded ERP data.

The frequency effect shows that MMN is not an all-or-none phenomenon, but a graded response whose amplitude reflects a parametric quantification of the amount of surprise conveyed by the stimulus, given the past stimuli. It is consistent with an internal model that takes into account statistical regularities.

Internal model of the temporal statistics in the input

The simplicity of the population of memory neurons used in our model allows us to visualize the statistical information learned by the network (Fig. 4). The only plasticity in the model occurs at synapses between the memory neurons and the predictive subpopulations. The information coded in these synaptic weights can be directly compared with the actual conditional probabilities in the actual input sequences. Figure 4 shows the mean synaptic weights between the delay lines and the predictive subpopulations as a function of the probability of occurrence of a deviant. They are compared with the actual statistics of transition probabilities in the inputs. Even though the plasticity rule was not specifically designed to converge onto a conditional transition probability, we can observe a close correspondence between the learned synaptic weights and the conditional information contained in the input. The peaks of synaptic strength coincide with the temporal intervals between the stimuli, and their amplitude is proportional to the probability of a transition between two stimuli almost regardless of the probability of occurrence of the first stimulus. Thus, this observation provides a very simple picture of what our model does: it stores, within its synaptic strengths, the conditional probability of observing a second stimulus at a certain latency after the first. Our claim is that the MMN reflects, in a quantitative manner, the degree of violation of such transition probabilities.

Importantly, the present model relies on STDP plasticity to internalize the statistics of the input. Data show that the MMN develops rapidly within few presentations of the standards (Winkler et al, 1996). To account for the MMN with such a mechanism, it is critical that plasticity occurs on a short timescale of a few seconds. To our knowledge, there are no data testing this prediction by trying to induce STDP on short timescales using ecological stimulation, and this hypothesis is therefore a prediction of the model that remains to be tested experimentally.

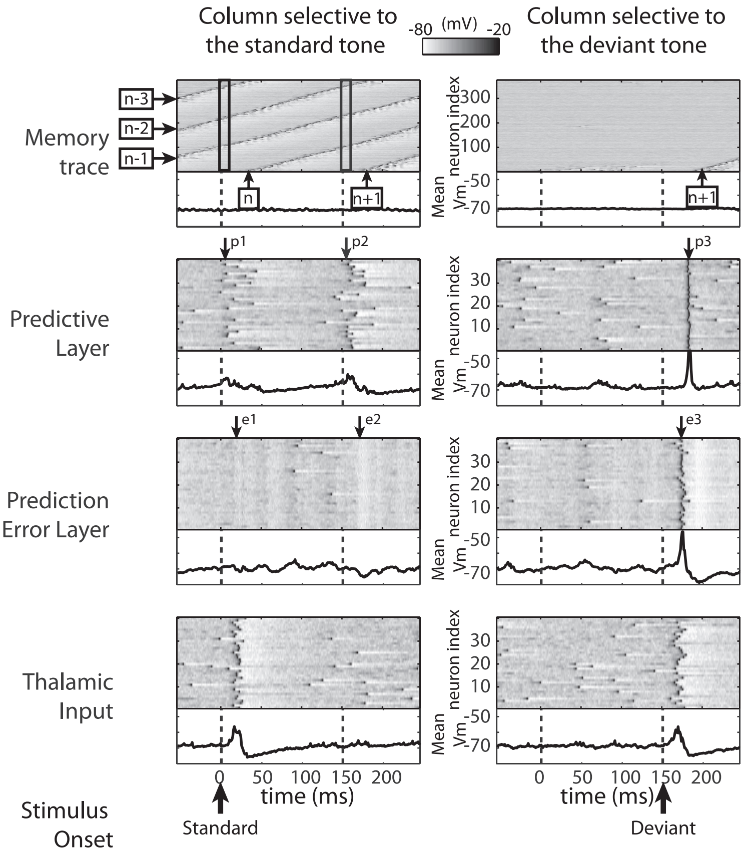

Figure 3. Simulated pattern of neural firing and membrane voltage during a single trial of the oddball paradigm. The figure shows a typical response to a standard tone ($t = 0$ ms) followed by a deviant tone ($t = 150$ ms). Left column, Subpopulations selective to tone A; right column, subpopulations selective to tone B. For each layer, the top part of the panel represents single-unit membrane voltage (one line per simulated neuron); the bottom part is the average voltage over the population. The neurons of the memory trace are reordered so that the propagation of the activity in a synfire chain way is made obvious. "n-1," "n-2," and "n-3" arrowed boxes refer to past stimuli whose activity is propagating in the delay lines initiated. In the left column, "n" and "n+1" arrowed boxes point to the initiation of a new memory trace following synchronous activity of the predictive population corresponding to the prediction of the stimuli n and n+1 ("p1" and "p2" arrows). In the right column, the "n+1" arrowed box shows the initiation of a new memory trace following synchronous activity of the predictive population corresponding to the prediction error signal of the n+1 (deviant) stimulus. After learning (Fig. 4), a reproducible pattern of activation in memory trace produces a depolarization in the predictive layer (black arrows) via a population of interneurons (not displayed here). The activity in predictive layer induces an hyperpolarization in the prediction error layer ("e2" arrow) at the approximate time when an A sound is expected. At $t = 0$, both prediction and input belong to the same column, resulting in a cancellation of excitation and inhibition inside the prediction error layer ("e1" arrow). At $t = 150$ ms, when a deviant stimulus B is presented, a depolarization of the prediction error population selective to the deviant ("e3" arrow) can be observed in parallel to the hyperpolarization of the predictive population selective to the standard ("e2" arrow). This depolarization is transmitted to the predictive ("p3" arrow) and memory (left column "n+1" arrow) populations.

The time span over which the stimulus transitions can be learned is strictly limited by the capacity of the memory. Here, we adopted as a simplifying assumption the hypothesis that the memory trace abruptly vanishes after 400 ms. Despite this artificially abrupt transition, we observe that synaptic weights get progressively weaker for more distant delays, due to the increased jitter in the coding of increasingly longer temporal intervals. In a more realistic memory network, the artificial delay lines that we used could be replaced by more realistic chaotic temporal dynamics, as in "reservoir" or echo state networks models (Maass et al., 2002; Buonomano, 2005; Buonomano and Laje, 2010; Pascanu and Jaeger, 2011). The memory trace would then become increasingly diluted with elapsed time, thus explaining that, in the standard oddball paradigm, a partially preserved but increasingly reduced MMN is observed as the time interval between tones is increased (Pegado et al., 2010).

MMN to repetition in an alternate signal

To further assess the properties of the model, we simulated the response to sequences where two stimuli are presented in an alternate fashion (ABABA . . .). On rare occasions, sound B is replaced by sound A. Horváth and Winkler (2004) showed experimentally that, in this condition, a MMN is now observed to the unexpected repetition of a stimulus B, in a context in which an alternation (ABABA . . .) was expected. This result is counterintuitive for habituation models, but entirely compatible with predictive-coding models. Indeed, we simulated the response of the network for an input constituted by a regular alternation of A and B every 150 ms. Rarely, sound B was replaced by sound A, resulting in the succession of three As in a row. Results are plotted in Figure 5. An sMMN is observed, showing that the unexpected repeated sound behaves as a deviant in the standard oddball paradigm. Indeed, the predictive population coding for B increases its activity 150 ms after an A occurred. In other words, the network learns to predict that after an A comes a B at 150 ms. This internalization of input statistics can also be seen in the synaptic weights.

Blindness to global regularities

Experimentally, the MMN is known to be blind to some global regularities in the stimulus sequence. For example, Bekinschtein et al. (2009) showed that, when participants are presented with

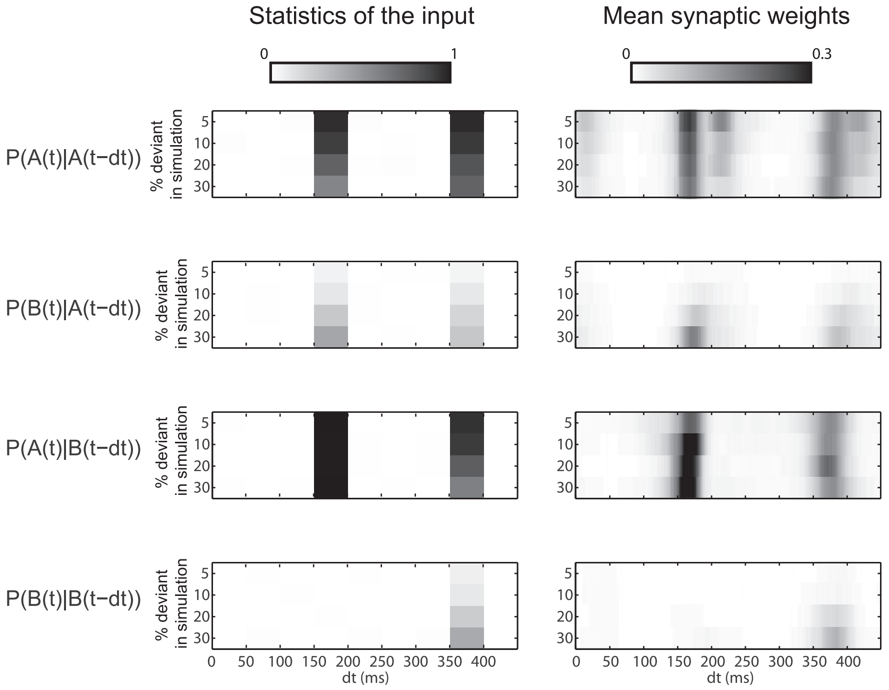

Figure 4. Correspondence between the transition statistics of the inputs (left) and the synaptic weights learned by the model (right). In each panel, the statistics are given for simulations with 5, 10, 20, and 30% of deviant sounds B in an oddball paradigm. Left column, Conditional probabilities of receiving a given sound (A or B) at time *t*, given the recent history of past inputs at times *t-dt* (*dt* ranging from 0 to 400 ms). Right column, Corresponding synaptic weights in the simulation at the end of learning. The gray levels indicate the mean synaptic weights between neurons of the recurrent memory network spiking on average at the time *dt* after the occurrence of an A or B sound, and the predictive neurons coding for the arrival of an A or B sound.

the repetition of a five-tone sequence AAAAB, the final B sound continues to elicit a MMN even though the occurrence of this sound is perfectly predictable based on the prior occurrence of four A sounds. In other words, the MMN seems to be "blind" to the overall sequence, and sensitive primarily to local transition probabilities, which favor the A→A transition over the A→B transition. Figure 6 shows the result of the simulation of our network on this paradigm. A total of 150 sequences of five inputs with ISI of 150 ms was presented. Seventy percent were AAAAB sequences, 20% AAAAA, and 10% AAAA (omission of the last sound, not analyzed here). The SOA between two sequences was 1.2 s. The average response to a frequent sequence is plotted in Figure 6. Note first that the first element of the sequence is not predicted. The time elapsed since the last sound is superior to the span of the delay line. It is consistent with data showing that no MMN exists on the first element of a sequence or for very long ISI (Mäntysalo and Näätänen, 1987; Cowan et al., 1993). Second, the final B sound elicits a stronger prediction error (sMMN) than the previous sounds. This effect arises because (1) the transition probabilities favor the prediction of an A sound following an A sound; and (2) the network cannot use the past occurrence of a B sound to predict a new B sound, because the temporal interval between them (1200 ms) exceeds the time span of the memory neurons. Both the increased response to the first sound and the final MMN tightly reproduced experimental scalp and intracranial recordings (Bekinschtein et al., 2009; Wacongne et al., 2011).

Using a closely related, yet importantly different paradigm, Sussman et al. (1998) showed that the MMN to the deviant sound B in circular sequences AAAABAAAAB. . . actually disappears if the SOA is small (100 ms) and B is presented at regular intervals. This observation is actually consistent with the model we propose. If the time between two B sounds is short enough, the network is able to learn the transition between two consecutive Bs, and the sMMN disappears. Our simulated network predicts that the MMN should reappear as soon as the temporal prediction of B is made impossible, either by spacing the B presentations beyond the capacity of the memory neurons, or by making B appear at irregular time intervals.

MMN to omission

One of the most remarkable properties of the auditory system is that it can generate evoked responses to an absent but expected stimulus (Joutsiniemi and Hari, 1989; Raij et al., 1997; Yabe et al., 1997; Hughes et al., 2001; Todorovic et al., 2011; Wacongne et al., 2011). We similarly tested the response of our network to the omission of an expected sound. We simulated the response of the network to pairs of AB sounds (ISI = 150 ms) separated by 500 ms, and rarely (10% of trials) omitted the second tone of the pair. We compared the response to such omissions to the response to identical single A tones presented every 500 ms in a block in which they were the only stimulus, and therefore no second stimulus was expected. As shown in Figure 7, the predictive currents anticipated the arrival of a second B sound and therefore produced a response to a nonexisting sound, as experimentally observed. Indeed, our results are tightly consistent with MEG and intracranial data obtained on a similar protocol (Hughes et al., 2001; Todorovic et al., 2011).

Interestingly, although this omission response is frequently called an MMN in the literature, our model proposes that it does not have exactly the same computational significance as the classic oddball MMN. In a predictive coding model, the omission response reflects solely a predictive component and not a prediction error per se (i.e., it does not reflect late, NMDA-dependent, prediction error currents, but early predictive currents). In the oddball paradigm, the main origin of the difference is an NMDA-dependent supragranular current, whereas the model predicts that the omission response should be resistant to competitive antagonists of NMDA channels, once the transition probabilities are learned.

MMN to changes in duration

Until now, we only simulated the onset of the input sounds. However, in primary auditory cortex, there are also populations of neurons that respond to sound offset (Volkov and Galazjuk, 1991; Chimoto et al., 2002). In a predictive coding perspective, the mechanism that we describe should capture not only how the onset of one sound can be predicted from the onset of another but also how the offset of one sound can be predicted based on the onset of the same sound. In the present section, we show that this effect can explain the observation of a MMN to a change in sound duration.

We stimulated our network with sounds of 150 ms duration, separated by a 300 ms ISI. We now assumed that the neural population "A" responded to the onset of the stimulus, and the "B" population to the offset. On a rare 10% of trials, the duration

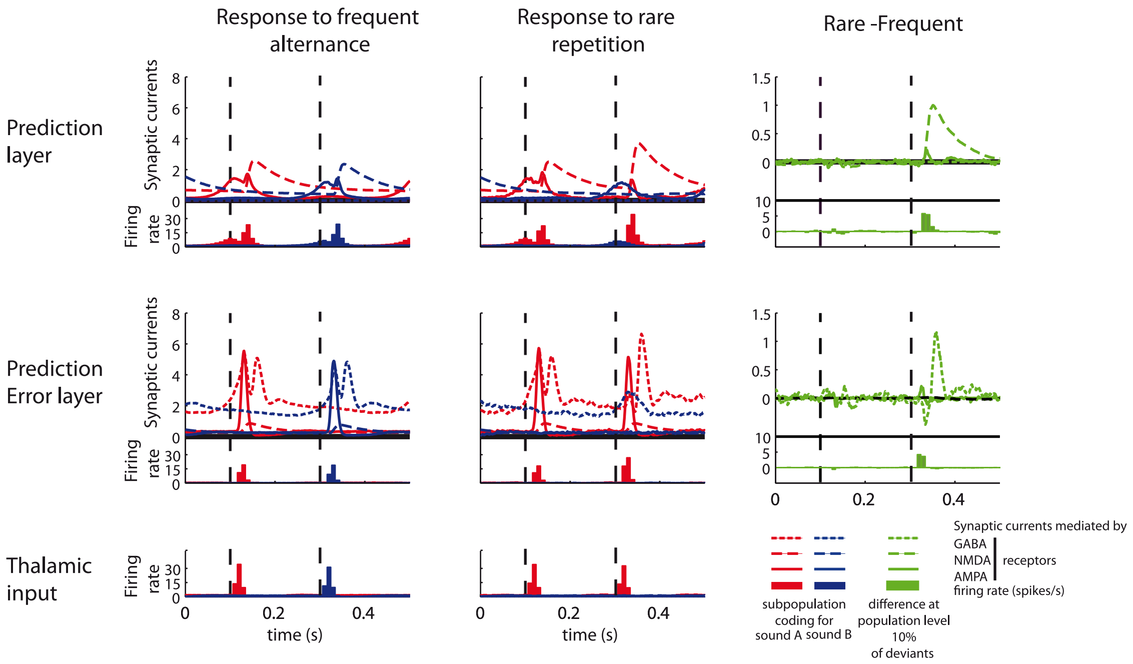

Figure 5. Simulating the MMN in response to an unexpected repetition among alternating stimuli. Left column, Mean response of the model to a frequent AB alternation in a ABABABA. . . stimulus. Middle column, Mean response to the rare AA repetition. Right column, The difference between the rare repetition and the frequent alternation shows a MMN elicited by the repeated sound AA. This prediction distinguishes predictive coding models.

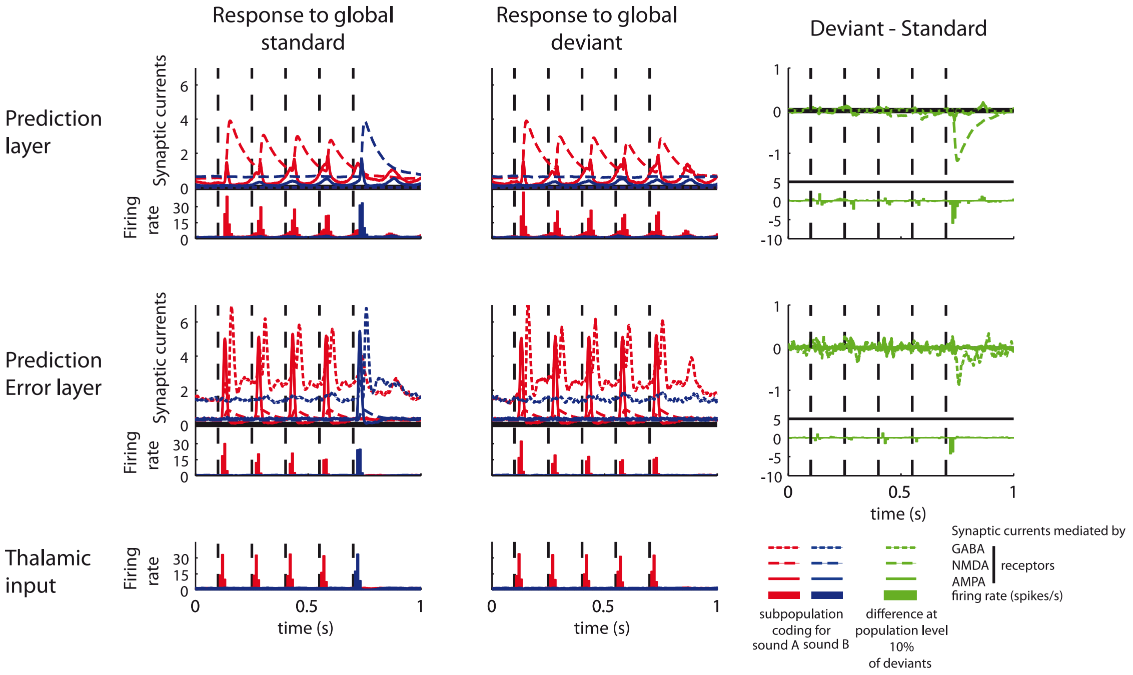

Figure 6. Simulating the lack of sensitivity of the MMN to global regularities that cannot be captured by local transition statistics. Left column, Mean response to a frequent AAAAB stimulus. Middle column, Mean response to the rare AAAAA stimulus. Right column, Difference between rare and frequent sequences. An MMN continues to be elicited by the final B sound of the standard AAAAB stimulus. Although the global sequence AAAAB is frequent and predictable, the MMN effect is driven primarily by the rarity of the local transition A⟶B.

of the sound, that is, the interval between the onset and the offset of the sound, was changed to 200 ms. We also simulated the converse situations in which standard sounds were 200 ms long and deviants, 150 ms long. Results are plotted in Figure 8, in which we compare the response to two physically identical sounds (150 ms duration) that act as standards or as deviants. When the input duration deviates from expectations, the internal model generates a prediction later than the actual arrival of the stimulus. The response to the offset is not cancelled and the prediction error is bigger. This prediction error signal is followed by another component, corresponding to the response to the omission of the later onset. Together, these responses capture the

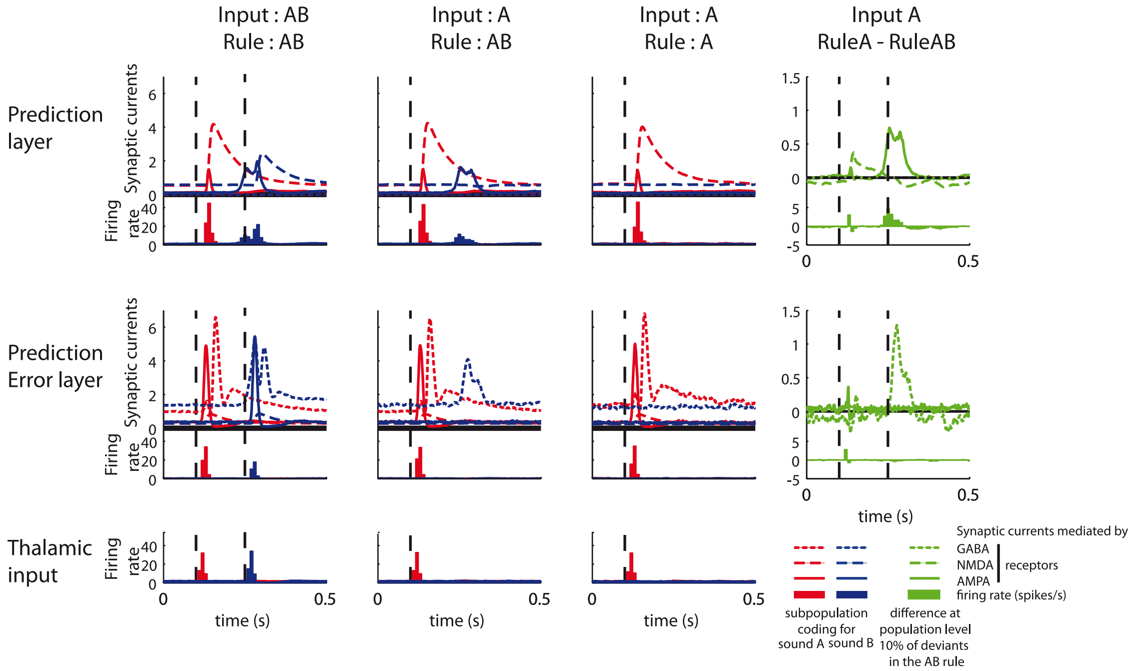

Figure 7. Simulating the MMN to the omission of an expected sound. First column, Mean response to a frequent AB pair. The network learns the predictable local transition A⟶B, which results in a reduced response to the predictable B sound (see arrow). Second column, Mean response to a rare A sound presented in isolation in the same context. The network generates a response to the omission of the expected sound B (arrow). Third column, Response to the same isolated sound A, in a different context where it is the frequent stimulus. Although the stimulus is physically identical to the second column, the predictive response to the omitted B sound is no longer seen. Fourth column, Difference between the second and third columns, isolating the simulated MMN to omission.

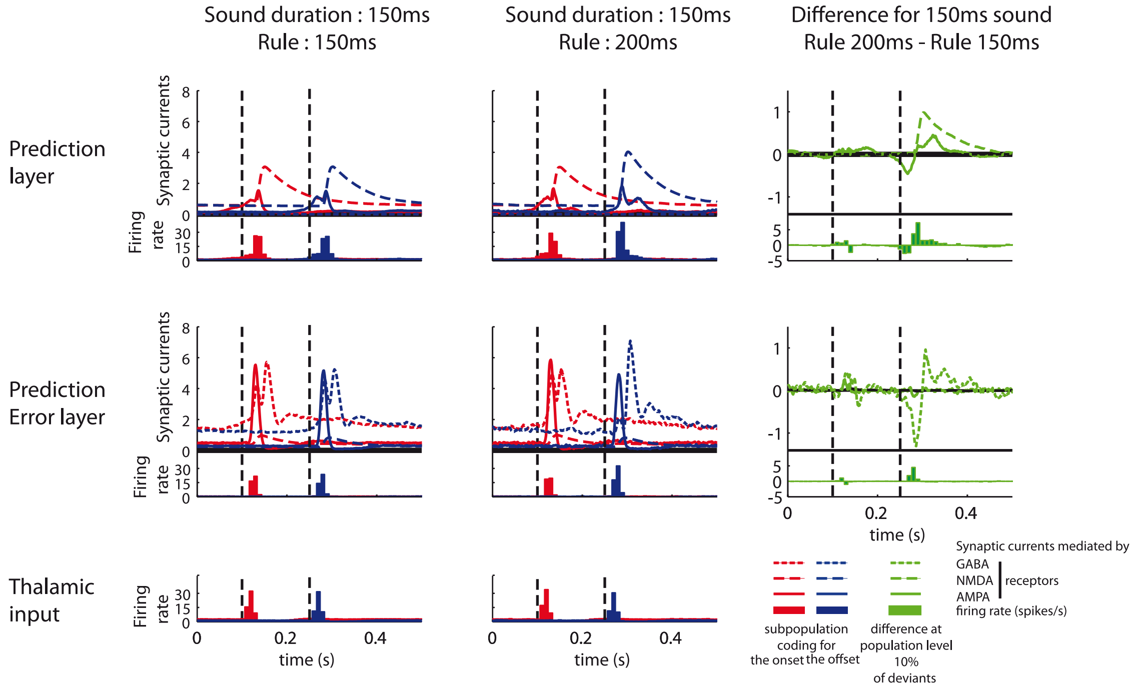

Figure 8. Simulating the MMN to a duration deviant. Blue and red now represent subpopulations selectively responsive, respectively, to sound onset and offset. Left column, Response to a frequent 150-ms-long sound. Middle column, Response to the same physical 150 ms sound when it serves as the rare deviant in an oddball paradigm where the frequent sound is 200 ms long. Right column, Difference between these two responses, isolating the MMN evoked by an unexpected change in duration.

experimentally observed MMN to duration deviants (Jacobsen and Schröger, 2003).

Note that, in our model, the change in duration is formally equivalent to a change in ISI: predictions that are focused in time fail to cancel incoming inputs that are shifted in time. Therefore, the model also reproduces the experimentally observed MMN to ISI deviants (Ford and Hillyard, 1981; Nordby et al., 1988).

Prediction versus habituation: an experimental test of the model

We have shown that a model exclusively based on predictive coding principles can explain, on a parsimonious basis, the major properties of the experimentally observed MMN. However, this is not the only theory proposed in the literature. May and Tiitinen (2010) defend the theory that MMN would only be the result of synaptic habituation, that is to say, the reduction of the amplitude of EPSPs as a result of repeated stimulation of the same synapse. Indeed, synaptic adaptation and short-term plasticity are commonly observed *in vivo* and *in vitro* in cortex (for review, see Calford, 2002), and more specifically in auditory cortex (Condon and Weinberger, 1991; Brosch and Schreiner, 2000), and it is likely that a complete theory of MMN should ultimately take such effects into account. However, is synaptic habituation sufficient to explain all MMN findings? In their review of MMN findings, May and Tiitinen (2010) suggest that all current MMN paradigms remain compatible with a habituation mechanism and argue that there is therefore no decisive evidence in favor of predictive coding models of the MMN. Contrariwise, our model leads us to propose one such critical test separating the predictive coding and habituation interpretations.

To provide a direct test of the two models, we decided to present pairs of closely consecutive sounds AB (200 ms SOA), separated by a broad temporal interval (>10 s). Occasionally, instead of the frequent AB pair (70% of trial), a deviant AA pair is presented in 10% of the trials, in which the same sound is repeated twice. The predictions of our model are straightforward: the first A sound predicts the second B sound in the frequent AB pair, and a mismatch negativity should therefore be generated whenever the unexpected A sound is heard instead (i.e., when the rare AA pair is presented instead of the frequent AB pair). We confirmed this prediction through simulations (the results are essentially identical to the alternation case ABABA… described earlier).

The habituation model, however, makes the opposite prediction: due to synaptic habituation, the second A sound in the AA pair should always elicit a reduced activity compared with the B sound in the AB pair, which solicits nonhabituated synapses. It could be argued that some higher-order neurons might habituate to the presentation of the frequent AB pair as a whole. Indeed, this is how May and Tiitinen (2010) account for the above-described alternation paradigm (ABABA…). However, experimentally, the recovery time of synaptic depression is generally of the order of a few seconds (Varela et al., 1997; Ulanovsky et al., 2004). Thus, by making the temporal interval between pairs as long as 10 s, we should render this putative effect of synaptic habituation at the level of the whole pair quite negligible, especially compared with the short-term adaptation to the individual sounds A in the pair AA, which are only separated by 200 ms. In this case, the habituation model can only predict a reduced brain response to the infrequent AA pair (i.e., the converse of a mismatch negativity).

As a further control, we introduced two additional rare deviants, the BB and BA pairs, which were also presented in 10% of the trials each. These pairs have the same structure as the AA pairs and AB pairs, but are presented with equal probability. In our model, as the transition probabilities B→B and B→A are the same, the predicted evoked responses should be the same. Thus, our model predicts a lack of any difference here, whereas the synaptic habituation model again predicts a reduced response to the repeated pair BB compared with the nonrepeated pair BA.

We recorded MEG signals while five healthy participants were instructed to listen to these stimuli. Each subject listened to two blocks of 120 pairs of sounds. The frequencies of the two sounds were 800 and 1600 Hz, and were counterbalanced between blocks. Figure 9 shows the results. In every subject, the second tone of the rare AA pairs elicited a MMN compared with the frequent AB pairs. The difference between the two conditions was significant for each individual subject and for both types of sensors (subject 1: Grad, 121–206 ms, $p < 1e\text{-}16$; Mag, 131–231 ms; subject 2: Grad, 131–186 ms, $p = 0.028$; Mag, 157–204 ms, $p = 0.044$; subject 3: Grad, 127–226 ms, $p = 0.003$; Mag, 126–264 ms, $p = 0.004$; subject 4: Grad, 109–177 ms, $p = 0.006$; Mag, 110–230 ms, $p = 0.001$; subject 5: Grad, 120–164 ms, $p = 0.04$; Mag, 116–260 ms, $p = 0.01$), as well as at the group level (Grad, 108–232 ms, $p < 1e\text{-}16$; Mag, 145–193 ms, $p < 1e\text{-}16$). The topography of the effect was similar to the classical MEG–MMN topography, with bilateral temporal activations.

Our model predicted that no difference should exist between the two control stimuli BA and BB. Indeed, no significant difference was observed between the two control stimuli (rare BB and rare BA pairs, presented with equal probability). In fact, a nonsignificant trend existed in the direction opposite to the one predicted by the synaptic habituation model (greater brain response to BA). This finding can be explained by the fact that the identity of the sounds serving as A and B was counterbalanced between the two halves of the experiment. As a result, the rare BA pair of the second run was the frequent AB pair of the first run. We reasoned that the transition that was well learned during the first block of trials could have continued to prevail in the second block, especially as the pairs BB and BA were presented for a very small number of times (12 each), thus largely preventing relearning of the actual equiprobability of the B→A and B→B transitions. We confirmed this hunch by separately analyzing the first and second halves of our experiment. When restricted to the first half, the two control stimuli BA and BB did not present any identifiable difference, whereas the same two conditions presented a stronger (yet nonsignificant) difference in the second half. Note again that the latter difference (stronger response to BB) was in the direction opposite to that expected from a habituation mechanism.

The experimental data are therefore consistent with the predictions of our model in great detail and in every single subject. To explain the data with synaptic habituation, one would have to postulate the existence of neurons that (1) respond specifically to the transition between the AB sounds; (2) present significant habituation after 10 s; and (3) whose habituation to AB pairs is strong enough to override the countereffect of habituation to the AA pair for neurons that respond only to frequency A. The latter assumption is particularly implausible because neurons responsive to A alone are likely to be much more numerous than neurons responsive to the AB pair as a whole, and because their habituation would be likely to be much stronger, given that the A–A delay of 200 ms is much shorter than the AB–AB delay of 10 s or more. Furthermore, the responses to BA and BB pairs provide no support for a habituation to individual B sounds. We therefore conclude that any habituation account of our data seems highly implausible.

Discussion

In this study, we developed a spiking neuron model of mismatch negativity, based on a predictive coding approach. We identified key properties of the mismatch effect and simulated the network response to a variety of test sequences. In particular, our model reproduced the known reduction in MMN amplitude when the frequency of the deviants increases, the MMN to repetition in an alternate sequence, and the response to the omission of an expected sound. Without any additional assumption, the model was able to account for the MMN to a change in stimulus duration or in interstimulus interval. We proposed a precise cortical localization of the neuronal populations postulated in the model and showed that our simulated current sources were consistent with actual electrophysiological data. We also showed that the model acquired a quantitative synaptic representation of transition probabilities. An alternative model hypothesizes that MMN arises purely from synaptic habituation. We identified a precise experimental context where the two models lead to opposite predictions and showed that MEG data from human participants fully support our predictions, with no evidence of a synaptic habituation effect.

Predictions versus synaptic habituation

In the present study, we showed that a model based on pure predictive coding, without any synaptic habituation component, could account for a large range of effects. It is important to note that, even though the habituation and predictive/memory accounts of MMN have been often opposed (Näätänen et al., 2005; Winkler, 2007; May and Tiitinen, 2010), the two hypotheses are not logically exclusive. It remains possible that the two processes concur to the final MMN effect, possibly in different proportions according to the paradigm. However, the conclusions of the MEG experimental test of our model are fully consistent with a purely predictive account of MMN and argue against a strong contribution of habituation effects.

Other recent studies argue in favor of a negligible role of habituation in the MMN effect. Recent human MEEG recordings indicate that the omission response observed when an expected sound fails to occur conforms to the predictions of hierarchical predictive coding models (Wacongne et al., 2011). In rodents, Farley et al. (2010) showed that stimulus-specific adaptation is indeed observed in auditory cortex but that its properties differ sharply from those of the MMN, in terms of sensitivity to NMDA antagonists or elicitation of a novelty response. Together, these results provide strong evi-

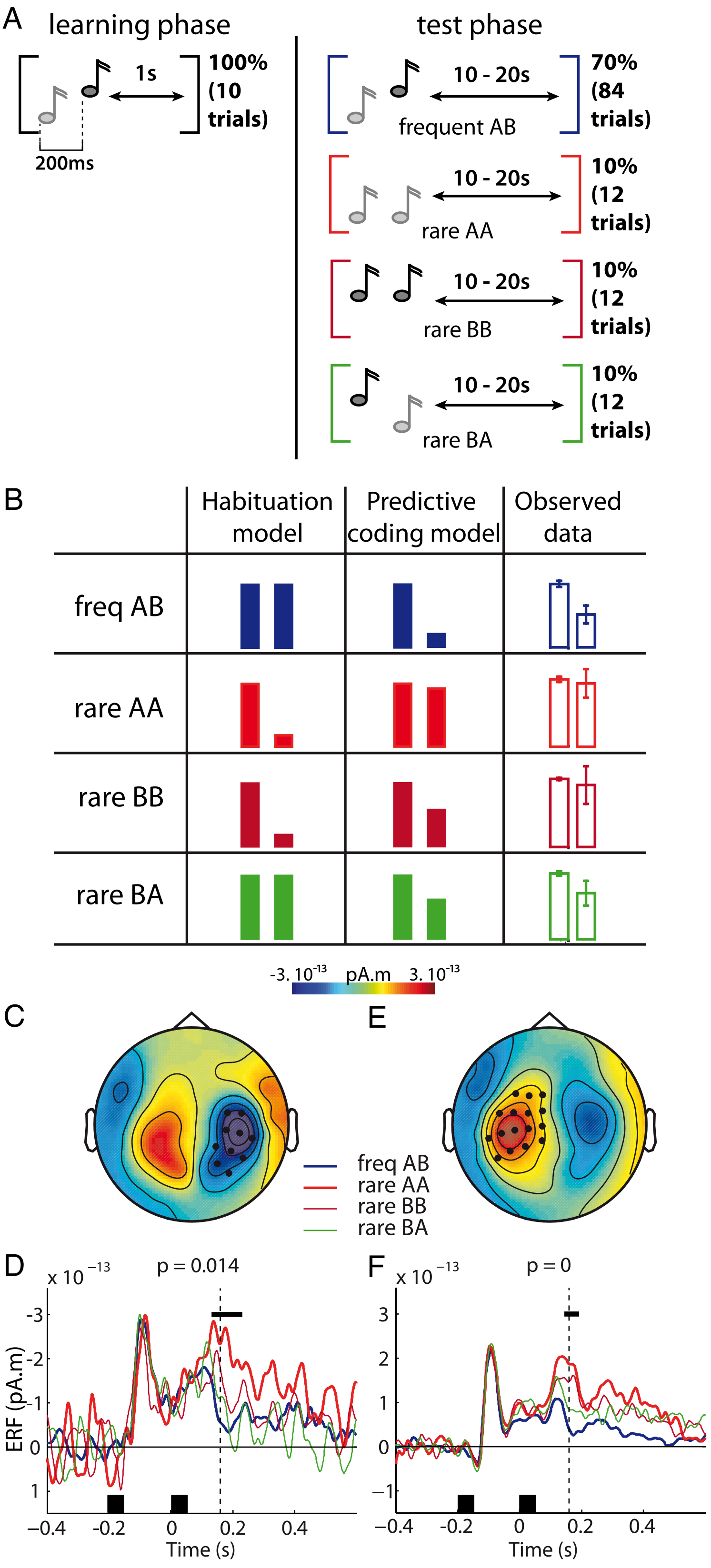

Figure 9. Experimental test of the model using magnetoencephalography. ***A***, Experimental design. Each block of trials begins with 10 identical pairs of tones (A followed by B). A and B are pure tones of 50 ms and frequency 800 and 1600 Hz, counterbalanced between blocks and subjects. The subject then listened to 120 pairs of tones: 70% of frequent AB pairs, and 10% of each of the rare pairs AA, BA, and BB. ***B***, Comparison between the relative response amplitude predicted by the habituation model, the predictive coding model, and the data. In the habituation model (left column), response amplitude is minimal to a repeated tone. In our predictive coding model (middle column), response amplitude depends on transition probabilities between the first and second

←

tone of the pair. The two models generate qualitatively different prediction for the AB and AA pairs. Observed group level responses (right column) to the two tones of each pair fit with predictive-coding predictions (for details, see Materials and Methods). Error bars represent the SEM. ***C–F***, MEG results for magnetometers for one representative subject (left) and for the average over all subjects (right). ***C*** and ***E*** show the sensor-level topography of the average difference in magnetic field between the rare AA and the frequent AB pairs, 170 ms after the onset of the second sound. The most significant cluster of sensors at this time is indicated by dots. ***D*** and ***F*** show the time course of the average response to all conditions within these sensors. The two tones were presented at −0.2 and 0 s (black squares). The line colors correspond to the brackets surrounding the stimuli in ***A***. The black line above the curves indicates the interval where a significant difference was found between AA and AB.

dence against a predominant role of synaptic habituation in the MMN effect and argue for the predictive coding hypothesis. Similar conclusions have been recently reached by other groups (Todorovic et al., 2011).

Extensions and limits of the model

In this study, we limited our simulations to two cortical columns coding for features distinct enough that thalamic inputs did not stimulate both columns at the same time. The model could be easily extended to a more continuous coding of tone frequency, in which each neuronal population codes for one preferred frequency but also responds more weakly to neighboring frequencies. This would give an account of the increase of MMN amplitude with the difference in frequency between standards and deviants (Sams et al., 1985).

Predictive coding requires that a memory of the recent past be used to predict the future. For the sake of simplicity, we adopted here the simplest hypothesis for a neural memory: a delay line. Although this assumption may not seem very realistic, we only argue here that there must be neural populations whose activity contains information about both the identity of recent stimuli and the time elapsed since they occurred. As noted by Buonomano (2005), these neurons need not be ordered in cortical space, but could be intermixed and arise from the partially chaotic temporal dynamics of cortical activation spread. Electrophysiological recordings from auditory cortex slices suggest that such a code might exist within the auditory cortex (Buonomano, 2003): when cortical neurons were stimulated, they triggered other neurons with reliable delays, without any correlation between response delays and the cortical distance from the neuron initially stimulated. Such a code would be ideal to support a memory of the recent past, as required in our model. It would allow the same neuronal populations to code tonotopically for the present and nontonopically for the past.

According to this hypothesis, our entire model would fit within a single cortical column and could constitute a basic building block for sensory predictive learning in various sensory systems. As noted by Friston et al. (2005), the closely similar neuronal architecture of cortical layers throughout the cerebral cortex supports the view that a similar computational principle of predictive coding may apply to the multiple hierarchical levels of the cortical areas of the brain. Thus, our model may be used to account for higher-order instances of mismatch responses, such as the distinct MMNs evoked by a change in phoneme versus speaker (Giard et al., 1995; Dehaene-Lambertz, 1997), or the mismatch responses observed outside the auditory modality, either in visual (Tales et al., 1999; Pazo-Alvarez et al., 2003), olfactive (Krauel et al., 1999; Pause and Krauel, 2000), and somatosensory (Kekoni et al., 1997; Shinozaki et al., 1998) modalities or even in a crossmodal context (Arnal et al., 2011).

Our model makes clear predictions as to the kind of regularities that should be reflected by the MMN. The model is only able to predict incoming stimuli by acquiring an internal representation of the transition probabilities between their onsets and offsets, over a window of a few hundreds of milliseconds. Thus, it fails to detect deviance from a rule that cannot be described at the level of transition probabilities. This statement should help clarify the issue of whether the MMN reflects "rule-based learning," which is often confused in the present literature.

For example, Sussman et al. (1998) showed that when the oddball paradigm was slightly modified so that deviant sounds B occurred regularly at short-enough intervals between the standards (AAAABAAAABAAAAB. . .), the MMN disappeared. Yet in a seemingly contradictory finding, using a minimally different paradigm, Bekinschtein et al. (2009) showed that an AAAAB rule could not be acquired by low-level sensory processing, since the final B sound continued to elicit a MMN even when the entire AAAAB sequence was fully predictable. According to our model, the main difference between the two protocols is the long additional temporal gap between two five-tone sequences that exist in the Bekinschtein paradigm, and which disrupts any recent memory capable of predicting the final B sound. Thus, the apparent inconsistency in the results is easily understandable if we consider the size of the memory delay needed for temporal prediction. This example stresses the importance of carefully assessing the matrix of transition probabilities when trying to design experiments probing rule learning.

An MMN-like response was also recorded for deviance from more abstract kinds of regularities such as tone repetition or ascending/descending tones (Paavilainen et al., 1999; Korzyukov et al., 2003; Endress et al., 2007). Whether or not such rules are learnable by our network depends on the specifics of the experimental design. To make the rule unlearnable by transition probabilities, the design should reserve a broad frequency band never presented during training, or over which the probabilities of ascending and descending tones are equal. Otherwise, given enough training exemplars, our network will learn the "rule" and even generalize to frequencies that are novel but close enough to the training frequencies. These conditions were not fulfilled in many previous papers. If they were, however, and if the MMN resisted to such a control, this would provide definitive evidence that the mechanisms underlying the MMN go beyond our basic transition-probability model. The model might be extended, however, by postulating higher-order neurons sensitive to melodic contours (e.g., any ascending contour). In general, the coding properties of the input neural populations will have a crucial impact on the kind of regularities that can be detected by our model.

Conclusion

The idea that the brain is not a passive input–output device but acts as a predictive system capable of anticipating on the future, has a long history in ethology, psychology, and neuroscience, and has been proven useful in many distinct domains of perception, cognition, and action (Dehaene and Changeux, 1991; Schultz et al., 1997; Sutton and Barto, 1998; Hosoya et al., 2005). Understanding the neural mechanisms by which the brain generates predictions is therefore an important goal for neuroscience. Predictive coding models of the MMN have been previously proposed (Friston, 2005; Friston et al., 2006; Garrido et al., 2009; Spratling, 2010) but only as abstract mathematical descriptions without a precise neurobiological implementation (Marreiros et al., 2009; but see Fiorillo, 2008). The present model resolves the difficulties associated with a neurobiological implementation of predictive coding. We show how the subtraction of observed versus predicted signals can be implemented through a specific architecture of inhibitory interneurons. We also show that a NMDA-dependent STDP plasticity rule is well adapted for learning of stimulus associations, leading to the prediction of a precise and essential contribution of NDMA receptors to predictive coding. The proposed architecture could generalize much beyond the specific domain of the MMN for which it was presently tested.

References

Alho K, Winkler I, Escera C, Huotilainen M, Virtanen J, Jääskeläinen IP, Pekkonen E, Ilmoniemi RJ (1998) Processing of novel sounds and frequency changes in the human auditory cortex: magnetoencephalographic recordings. Psychophysiology 35:211–224.

Arnal LK, Wyart V, Giraud AL (2011) Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. Nat Neurosci 14:797–801.

Bekinschtein TA, Dehaene S, Rohaut B, Tadel F, Cohen L, Naccache L (2009) Neural signature of the conscious processing of auditory regularities. Proc Natl Acad Sci U S A 106:1672–1677.

Bi G, Poo M (1999) Distributed synaptic modification in neural networks induced by patterned stimulation. Nature 401:792–796.

Brosch M, Schreiner CE (2000) Sequence sensitivity of neurons in cat primary auditory cortex. Cereb Cortex 10:1155–1167.

Brunel N, Wang XJ (2001) Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. J Comput Neurosci 11:63–85.

Buonomano DV (2003) Timing of neural responses in cortical organotypic slices. Proc Natl Acad Sci U S A 100:4897–4902.

Buonomano DV (2005) A learning rule for the emergence of stable dynamics and timing in recurrent. J Neurophysiol 94:2275–2283.

Buonomano DV, Laje R (2010) Population clocks: motor timing with neural dynamics. Trends Cogn Sci 14:520–527.

Calford MB (2002) Dynamic representational plasticity in sensory cortex. Neuroscience 111:709–738.

Chimoto S, Kitama T, Qin L, Sakayori S, Sato Y (2002) Tonal response patterns of primary auditory cortex neurons in alert cats. Brain Res 934:34–42.

Condon CD, Weinberger NM (1991) Habituation produces frequency-specific plasticity of receptive fields in the auditory cortex. Behav Neurosci 105:416–430.

Cowan N, Winkler I, Teder W, Näätänen R (1993) Memory prerequisites of mismatch negativity in the auditory event-related potential (ERP). J Exp Psychol Learn Mem Cogn 19:909–921.

Debanne D, Shulz DE, Frégnac Y (1998) Activity-dependent regulation of "on" and "off" responses in cat visual cortical receptive fields. J Physiol 508:523–548.

Dehaene S, Changeux JP (1991) The Wisconsin Card Sorting Test: theoretical analysis and modeling in a neuronal network. Cereb Cortex 1:62–79.

Dehaene-Lambertz G (1997) Electrophysiological correlates of categorical phoneme perception in adults. Neuroreport 8:919–924.

Ehrlichman RS, Maxwell CR, Majumdar S, Siegel SJ (2008) Deviance-elicited changes in event-related potentials are attenuated by ketamine in mice. J Cogn Neurosci 20:1403–1414.

Endress AD, Dehaene-Lambertz G, Mehler J (2007) Perceptual constraints and the learnability of simple grammars. Cognition 105:577–614.

Farley BJ, Quirk MC, Doherty JJ, Christian EP (2010) Stimulus-specific adaptation in auditory cortex is an NMDA-independent process distinct from the sensory novelty encoded by the mismatch negativity. J Neurosci 30:16475–16484.

Fiorillo CD (2008) Towards a general theory of neural computation based on prediction by single neurons. PLoS One 3:e3298.

Ford JM, Hillyard SA (1981) Event-related potentials (ERPs) to interruptions of a steady rhythm. Psychophysiology 18:322–330.

Friston K (2005) A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci 360:815–836.

Friston K, Kilner JM, Harrison L (2006) A free energy principle for the brain. J Physiol 100:70–87.

Garrido MI, Kilner JM, Kiebel SJ, Friston KJ (2007) Evoked brain responses are generated by feedback loops. Proc Natl Acad Sci U S A 104:20961–20966.

Garrido MI, Kilner JM, Kiebel SJ, Friston KJ (2009) Dynamic causal modeling of the response to frequency deviants. J Neurophysiol 101:2620–2631.

Giard MH, Lavikahen J, Reinikainen K, Perrin F, Bertrand O, Pernier J, Näätänen R (1995) Separate representation of stimulus frequency, intensity, and duration in auditory sensory memory: an event-related potential and dipole-model analysis. J Cogn Neurosci 7:133–143.

Horváth J, Winkler I (2004) How the human auditory system treats repetition amongst change. Neurosci Lett 368:157–161.

Hosoya T, Baccus SA, Meister M (2005) Dynamic predictive coding by the retina. Nature 436:71–77.

Hughes HC, Darcey TM, Barkan HI, Williamson PD, Roberts DW, Aslin CH (2001) Responses of human auditory association cortex to the omission of an expected acoustic event. Neuroimage 13:1073–1089.

Izhikevich EM (2003) Simple model of spiking neurons. IEEE Trans Neural Netw 14:1569–1572.

Jacobsen T, Schröger E (2003) Measuring duration mismatch negativity. Clin Neurophysiol 114:1133–1143.

Javitt DC, Steinschneider M, Schroeder CE, Arezzo JC (1996) Role of cortical *N*-methyl-D-aspartate receptors in auditory sensory memory and mismatch negativity generation: implications for schizophrenia. Proc Natl Acad Sci U S A 93:11962–11967.

Joutsiniemi SL, Hari R (1989) Omissions of auditory stimuli may activate frontal cortex. Eur J Neurosci 1:524–528.

Kekoni J, Hämäläinen H, Saarinen M, Gröhn J, Reinikainen K, Lehtokoski A, Näätänen R (1997) Rate effect and mismatch responses in the somatosensory system: ERP-recordings in humans. Biol Psychol 46:125–142.

Kiebel SJ, Daunizeau J, Friston KJ (2008) A hierarchy of time-scales and the brain. PLoS Comput Biol 4:e1000209.

Kiebel SJ, Daunizeau J, Friston KJ (2009) Perception and hierarchical dynamics. Front Neuroinform 3:20.

Korzyukov OA, Winkler I, Gumenyuk VI, Alho K (2003) Processing abstract auditory features in the human auditory cortex. Neuroimage 20:2245–2258.

Krauel K, Schott P, Sojka B, Pause BM, Ferstl R (1999) Is there a mismatch negativity analogue in the olfactory event-related potential? J Psychophysiol 13:49–55.

Lee TS, Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. J Opt Soc Am A Opt Image Sci Vis 20:1434–1448.

Maass W, Natschläger T, Markram H (2002) Real-time computing without stable states: a new framework for neural computation based on perturbations. Neural Comput 14:2531–2560.

Mäntysalo S, Näätänen R (1987) The duration of a neuronal trace of an auditory stimulus as indicated by event-related potentials. Biol Psychol 24:183–195.

Marreiros AC, Kiebel SJ, Daunizeau J, Harrison LM, Friston KJ (2009) Population dynamics under the Laplace assumption. Neuroimage 44:701–714.

May PJ, Tiitinen H (2010) Mismatch negativity (MMN), the deviance-elicited auditory deflection, explained. Psychophysiology 47:66–122.

Näätänen R (2003) Mismatch negativity: clinical research and possible applications. Int J Psychophysiol 48:179–188.

Näätänen R, Paavilainen P, Alho K, Reinikainen K, Sams M (1987) The mismatch negativity to intensity changes in an auditory stimulus sequence. Electroencephalogr Clin Neurophysiol Suppl 40:125–131.

Näätänen R, Paavilainen P, Reinikainen K (1989) Do event-related potentials to infrequent decrements in duration of auditory stimuli demonstrate a memory trace in man? Neurosci Lett 107:347–352.

Näätänen R, Jacobsen T, Winkler I (2005) Memory-based or afferent processes in mismatch negativity (MMN): a review of the evidence. Psychophysiology 42:25–32.

Näätänen R, Paavilainen P, Rinne T, Alho K (2007) The mismatch negativity (MMN) in basic research of central auditory processing: a review. Clin Neurophysiol 118:2544–2590.

Nordby H, Roth WT, Pfefferbaum A (1988) Event-related potentials to breaks in sequences of alternating pitches or interstimulus intervals. Psychophysiology 25:262–268.

Paavilainen P, Jaramillo M, Näätänen R, Winkler I (1999) Neuronal populations in the human brain extracting invariant relationships from acoustic variance. Neurosci Lett 265:179–182.

Pascanu R, Jaeger H (2011) A neurodynamical model for working memory. Neural Netw 24:199–207.

Pause BM, Krauel K (2000) Chemosensory event-related potentials (CSERP) as a key to the psychology of odors. Int J Psychophysiol 36:105–122.

Pazo-Alvarez P, Cadaveira F, Amenedo E (2003) MMN in the visual modality: a review. Biol Psychol 63:199–236.

Pegado F, Bekinschtein T, Chausson N, Dehaene S, Cohen L, Naccache L (2010) Probing the lifetimes of auditory novelty detection processes. Neuropsychologia 48:3145–3154.

Raij T, McEvoy L, Mäkelä JP, Hari R (1997) Human auditory cortex is activated by omissions of auditory stimuli. Brain Res 745:134–143.

Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive field effects. Nat Neurosci 2:79–87.

Sams M, Paavilainen P, Alho K, Näätänen R (1985) Auditory frequency discrimination and event-related potentials. Electroencephalogr Clin Neurophysiol 62:437–448.

Sato Y, Yabe H, Hiruma T, Sutoh T, Shinozaki N, Nashida T, Kaneko S (2000) The effect of deviant stimulus probability on the human mismatch process. Neuroreport 11:3703–3708.

Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593–1599.

Sculthorpe LD, Ouellet DR, Campbell KB (2009) MMN elicitation during natural sleep to violations of an auditory pattern. Brain Res 1290:52–62.

Shinozaki N, Yabe H, Sutoh T, Hiruma T, Kaneko S (1998) Somatosensory automatic responses to deviant stimuli. Brain Res Cogn Brain Res 7:165–171.

Spratling MW (2010) Predictive coding as a model of response properties in cortical area V1. J Neurosci 30:3531–3543.

Sussman E, Ritter W, Vaughan HG Jr (1998) Predictability of stimulus deviance and the mismatch negativity. Neuroreport 9:4167–4170.

Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. Cambridge, UK: Cambridge UP.

Tales A, Newton P, Troscianko T, Butler S (1999) Mismatch negativity in the visual modality. Neuroreport 10:3363–3367.

Taulu S, Kajola M, Simola J (2004) Suppression of interference and artifacts by the Signal Space Separation Method. Brain Topogr 16:269–275.

Thomson AM, Lamy C (2007) Functional maps of neocortical local circuitry. Front Neurosci 1:19–42.

Tikhonravov D, Neuvonen T, Pertovaara A, Savioja K, Ruusuvirta T, Näätänen R, Carlson S (2008) Effects of an NMDA-receptor antagonist MK-801 on an MMN-like response recorded in anesthetized rats. Brain Res 1203:97–102.

Tikhonravov D, Neuvonen T, Pertovaara A, Savioja K, Ruusuvirta T, Näätänen R, Carlson S (2010) Dose-related effects of memantine on a mismatch negativity-like response in anesthetized rats. Neuroscience 167:1175–1182.

Todorovic A, van Ede F, Maris E, de Lange FP (2011) Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: an MEG Study. J Neurosci 31:9118–9123.

Ulanovsky N, Las L, Farkas D, Nelken I (2004) Multiple time scales of adaptation in auditory cortex neurons. J Neurosci 24:10440–10453.

Varela JA, Sen K, Gibson J, Fost J, Abbott LF, Nelson SB (1997) A quantitative description of short-term plasticity at excitatory synapses in layer 2/3 of rat primary visual cortex. J Neurosci 17:7926–7940.

Volkov IO, Galazjuk AV (1991) Formation of spike response to sound tones in cat auditory cortex neurons: interaction of excitatory and inhibitory effects. Neuroscience 43:307–321.

Wacongne C, Labyt E, van Wassenhove V, Bekinschtein T, Naccache L, Dehaene S (2011) Evidence for a hierarchy of predictions and prediction errors in human cortex. Proc Natl Acad Sci U S A 108:20754–20759.

Winkler I (2007) Interpreting the mismatch negativity (MMN). J Psychophysiol 21:147–163.

Winkler I, Cowan N, Csépe V, Czigler I, Näätänen R (1996) Interactions between transient and long-term auditory memory as reflected by the mismatch negativity. J Cogn Neurosci 8:403–415.

Yabe H, Tervaniemi M, Reinikainen K, Näätänen R (1997) Temporal window of integration revealed by MMN to sound omission. Neuroreport 8:1971–1974.