Neurosci. Bull. January, 2021, 37(1):70–80
https://doi.org/10.1007/s12264-020-00527-1

www.neurosci.cn
www.springer.com/12264

ORIGINAL ARTICLE

Dynamic Brain Responses Modulated by Precise Timing Prediction in an Opposing Process

Minpeng Xu[1,2] · Jiayuan Meng[1] · Haiqing Yu[1] · Tzyy-Ping Jung[1,3] · Dong Ming[1,2]

Received: 9 December 2019 / Accepted: 11 February 2020 / Published online: 16 June 2020
© Shanghai Institutes for Biological Sciences, CAS 2020

Abstract The brain function of prediction is fundamental for human beings to shape perceptions efficiently and successively. Through decades of effort, a valuable brain activation map has been obtained for prediction. However, much less is known about how the brain manages the prediction process over time using traditional neuropsychological paradigms. Here, we implemented an innovative paradigm for timing prediction to precisely study the temporal dynamics of neural oscillations. In the experiment recruiting 45 participants, expectation suppression was found for the overall electroencephalographic activity, consistent with previous hemodynamic studies. Notably, we found that N1 was positively associated with predictability while N2 showed a reversed relation to predictability. Furthermore, the matching prediction had a similar profile with no timing prediction, both showing an almost saturated N1 and an absence of N2. The results indicate that the N1 process showed a 'sharpening' effect for predictable inputs, while the N2 process showed a 'dampening' effect. Therefore, these two paradoxical neural effects of prediction, which have provoked wide confusion in accounting for expectation suppression, actually co-exist in the procedure of timing prediction but work in separate time windows. These findings strongly support a recently-proposed opposing process theory.

Keywords Expectation suppression · Predictive coding · Event-related potentials · Timing prediction

Introduction

Human perceptions are shaped by not only sensory inputs, but also prior knowledge stored in the brain [1, 2]. The ability to generate and utilize prediction is fundamental for survival [2, 3], and has been demonstrated to be indispensable in a wide range of mental processes [4–8]. However, the neural effect and underlying mechanism of prediction remains controversial.

Predictive coding, an influential theory of the neural process of prediction, proposes that the brain actively builds up predictive templates to shape perceptions, rather than being driven merely by bottom-up stimuli. It optimizes perceptions using prediction errors, i.e., differences between predictions and sensory inputs [9–11]. In this way, unexpected stimuli would induce more neural activity than expected stimuli, to address larger prediction errors. This is known as the 'dampening' effect, i.e. the neural representations is 'dampened' for predictable stimuli. The phenomenon of expectation suppression has been reported in both the electroencephalographic (EEG) and hemodynamic studies [12–16]. However, some studies have shown that even with lower brain responses, the expected stimuli have more salient neural representation than unexpected stimuli [17–21], which is beyond the 'dampening' account. This

Minpeng Xu and Jiayuan Meng contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12264-020-00527-1) contains supplementary material, which is available to authorized users.

✉ Dong Ming
richardming@tju.edu.cn

[1] College of Precision Instruments and Optoelectronics Engineering, Tianjin University, Tianjin 300072, China

[2] Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin 300072, China

[3] Swartz Center for Computational Neuroscience, University of California, San Diego, CA 92093, USA

unusual phenomenon was explained by Kok *et al.* as prediction actually 'sharpening' and not 'dampening' the neural representation [22]. The 'sharpening' account holds the view that, although the overall neural activity in response to expected information seems to be suppressed, the neurons encoding the expected information are not suppressed—it is the neurons not tuned to the expected information that are suppressed, making the expected features more salient and selective, concurrently resulting in lower neural responses [3]. Although some conflicting phenomena can be explained by the reversal of prediction-related representations by attention [23], we are still far from understanding the mechanism underlying the reported paradoxical effects of prediction.

Recently, increasing numbers of studies have realized that it may be misleading to stress only one effect of prediction during the whole predictive coding process [2, 3]. Clark *et al.* highlighted the duplex architecture of predictive coding theory [2], which assumes that at each prediction processing level, the representational and error signals are encoded by two functionally distinct units [1, 24–26], although how they interact over time remains unclear. Moreover, Friston proposed that prediction processing may dynamically evolve from sensory representation to error correction (or perceptual learning) [10]. Inspired by these assumptions, a recent opposing processing theory posited that perception is initially biased towards the expected information after the sensory representation, and no other process is needed if the inputs are sufficiently in line with current expectation, while error-correction would be conducted if the input is different enough from the expectation [27]. That means, the paradoxical 'sharpening' and 'dampening' effects of prediction may be reconciled by studying the neural signals at fine temporal resolution. Therefore, we set out to investigate the temporal dynamics of predictive timing processes to obtain experimental evidence associated with the 'sharpening' and 'dampening' effects, and to determine whether these conflicting effects are compatible during the dynamic process. To this end, we designed an innovative experimental paradigm to study the time course of neural oscillations under different precise timing prediction states, and found that the 'sharpening' and 'dampening' effects of prediction actually co-exist in the predictive timing process, but in distinct processing stages with an opposing trend.

Materials and Methods

Participants

Forty-five healthy individuals (23 females, 19–28 years old) participated in the current experiment. Forty-two were right-handed and other three were left-handed. All individuals had normal or corrected-to-normal vision, and were free from psychological or neurological diseases. The experimental procedures were approved by the Institutional Review Board at Tianjin University. All possible consequences of the study were explained, and written informed consent was given by all participants.

Stimulus

The double-flashes used in this study were presented by a 15×15 mm^2 LED placed at the eye level and 80 cm from the participant. The LED was driven by a chronometric FPGA platform (Cyclone II: EP2C8T144C8) with 20-ns resolution. The duration of each flash was 120 ms (Fig. 1). Each trial started with an auditory cue that lasted for 1,000 ms. Then there was a blank with a random duration selected from 1,000 ms, 1,500 ms, or 2,000 ms. Next, a double-flash appeared with unpredictable stimulus-onset asynchrony (SOA) of 400 ms, 600 ms, or 900 ms. Finally, there was a blank period between 1,600 ms and 2,600 ms before the next trial. A total of 30 trials were conducted in one session, and 16 sessions were conducted for each participant. Thus, there were 160 trials for each kind of double-flash.

Training Procedure

According to predictive coding theory, the construction of the perceptual template is a prerequisite for the brain's predictive processing [28–30]. Therefore, it is very important for participants to set up a precise template in their minds before testing its influences on sensory inputs. As the mental clock in the human brain is not very precise, and is relatively insensitive to sub-second time intervals without proper training [31], it is difficult for participants to set up a precise timing template only using recent ambiguous experiences, such as temporarily learning past repetitions [13, 32] or temporal associations [33]. For this reason, training sessions were necessary in this study. The precise timing template built up in the training sessions made the predictive processing in the formal experiment more stable and invariable across trials, and made this study sensitive and reliable.

All participants were trained for about three days before testing (the specific training time in each day differed

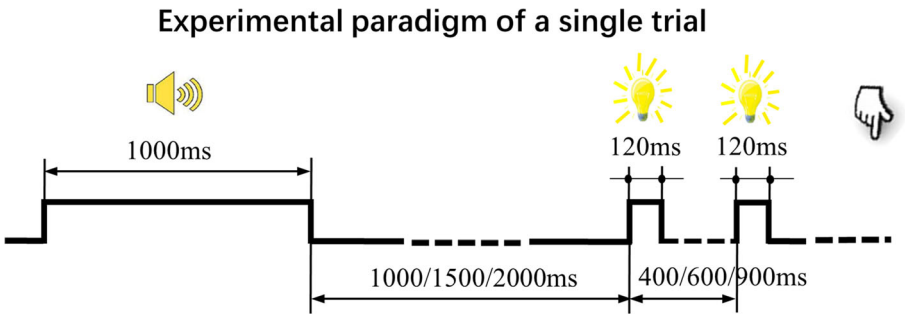

Fig. 1 Illustration of a single trial in the testing experiment. In each trial, a loudspeaker first sounds for 1,000 ms to alert the participant to the upcoming double-flash, followed by a random blank period (1,000 ms, 1,500 ms, or 2,000 ms). Then the LED in front of the participant flashes twice with an unpredictable SOA. Each participant was required to press different buttons to indicate their judgments.

among participants). The averaged task accuracy improved from 62.58% to 92.36%, and the average reaction time shortened from 998.60 ms to 509.32 ms (details in Fig. S1). Only when the participants achieved high enough accuracy (> 85%) for more than five training blocks, which indicated that the timing templates were stably stored in their minds, were they allowed to take part in the formal experiment.

Formal Experimental Procedure

In the formal experiment, each participant was required to conduct four different kinds of mental task. Specifically, they were required to (1) compare the actual double-flash with the SOA400 double-flash in their mind (under this instruction, the participants were only allowed to recall the SOA400 double-flash so that the SOA400 timing template was exclusively deployed as the initial predicative code. Therefore, the actual SOA400 double-flash absolutely matched the prediction, AMP); (2) compare the actual double-flash with the SOA600 timing template (under this instruction, the participants were only allowed to recall the SOA600 double-flash so that the SOA600 timing template was exclusively deployed as the initial predictive code. Therefore, the actual SOA400 double-flash mismatched the prediction, MMP); (3) determine which SOA the actual double-flash was from 400, 600, or 900 ms (under this instruction, the participants needed to concurrently recall SOA400, SOA600, and SOA900 double-flashes so that all three timing templates were deployed for the initial prediction state. Therefore, the actual SOA400 double-flash partially matched the prediction, PMP); and (4) indicate the onset of the second flash (under this instruction, no predictive code, NPC, was initially deployed). After each double-flash, each participant made a judgement by pressing a specific button as soon as possible. This study only focused on the brain responses to stimuli with the SOA400 double-flash, because the longer SOA raised the problem of information disclosure to participants after 400 ms.

There were 4 sessions for each mental task, and all the sixteen sessions were interleaved. In each session, the three kinds of double-flash were presented in a random order with equal probability, and no choice preference was found when participants made decisions (Fig. S2). Moreover, pressing a button with the right or left thumb was balanced across sessions. In particular, for AMP and MMP, each participant was required to use the right or left thumb to press the Match/Mismatch button in two sessions, and have an exchange in the other two sessions. For PMP, they were required to use the right thumb to press the SOA400 button in two sessions, and the left thumb to press the SOA400 button in the other two sessions. For NPC, they were required to use the right thumb to press the button in two sessions, and the left thumb to press the button in the other two sessions.

In addition, two points need to be further explained for a better understanding of the experimental paradigm. First, this study shaped the matching and mismatching processes of timing prediction by comparing the actual with the predicted time lapses, which was adapted from previous studies of feature-based prediction [30, 34]. Second, the current paradigm for shaping the timing prediction avoided the problem caused by using recent experiences of past stimulus repetitions. As the stimulus repetition would entail entrainment of neural oscillations which could influence the following neural process and response [33, 35, 36], it would disturb the natural predictive processing, making it impossible to decouple the top-down influence of timing prediction from the bottom-up influence of neural entrainment.

EEG Recording and Pre-processing

EEG was recorded by a Neuroscan Synamps2 system at a sample rate of 1000 Hz, and filtered by a low-pass filter at 200 Hz and a notch filter at 50 Hz. Sixty-four electrodes

were positioned on the scalp according to the International 10–20 system. We only focused on signals from the parietal and occipital regions. All channels were referenced to the tip of the nose and grounded to the frontal region. Eye-blinks were monitored by signals recorded at FP1 and FP2. The stored EEG data were then filtered by a ChebyshevII low-pass filter cutting at 45 Hz, and down-sampled to 200 Hz. EEG trials were segmented from − 900 ms to 2,100 ms after the second flash onset. Only correct trials with reaction latencies between 100 ms and 800 ms were considered in the subsequent analyses. For the first ERP, the baseline was corrected by subtracting the mean of 200 ms of data before the first flash onset, while for the second ERP, the baseline was corrected by subtracting the mean of 50 ms of data before the second flash onset to avoid the influence of contingent negative variations (CNVs).

ERP Amplitude and Latency Measurements

The predictive coding theory was originally proposed to explain the prediction mechanism in primary visual cortex [9]. Low-level sensory processing is an important issue for the prediction mechanism and has become a research topic of interest [12, 23, 37]. Therefore, we mainly focus on the ERPs of the posterior scalp locations where low-level visual processing predominately takes place. On the basis of grand averages, the time windows of the first N1 and N2 ERPs were defined as − 260 ms to − 225 ms and − 170 ms to − 105 ms, and those of the second N1 and N2 ERPs as 130 ms to 165 ms and 200 ms to 260 ms, respectively. In addition, as the durations of the N2 component varied with conditions, we also applied jack-knife-based scoring methods to measure the N2 components (AMP, 205 ms–225 ms; PMP, 205 ms–250 ms; MMP, 205 ms–275 ms; details in Fig. S3). The amplitude of each ERP component was calculated as the mean within the specified time window. The latency was measured as the time point before which 50% of the total component area occurred in the specified time window.

Inter-trial Coherence (ITC), Event-Related Spectral Perturbation (ERSP), and Evoked EEG Energy

ITC measures the consistency across trials of the EEG spectral phase at each frequency and time window [38]; it ranges from 0 to 1. The testing trials showed more phase coherence, so the ITC value was closer to 1. ERSP reflects the changes of event-related spectral power at each time-frequency point compared to the baseline of pre-stimulus spectral power in its corresponding frequency band [38]. For the first response, the baseline was the mean of 200 ms of data before the first flash onset, while for the second response, the baseline was the mean of 50 ms of data before the second flash onset. The evoked EEG energy was the sum of the ERSP values from 0 s to 2.4 s after the first flash onset, where the baseline was the mean of 200 ms of data before the first flash onset.

Statistical Tests

One-way repeated-measures ANOVA with Bonferroni correction was used to test the significance of behavioral differences among conditions. Two-way repeated-measures ANOVA (electrode × condition) was used to test the significance of the ERP, ITC, ERSP, and evoked EEG energy differences, and only the effects of condition are shown. Bonferroni correction was used for multiple comparisons of conditions. Paired *t*-tests were used to test the significance of ERP differences between AMP and NPC. All error bars show 95% within-participant confidence intervals of the mean difference between conditions.

Results

Behavioral Analyses

In this study, a participant's behavior could be influenced by two factors – task difficulty and prediction state – which were addressed separately. Since NPC was the easiest and simplest task, it elicited the highest response accuracy (99.89% ± 0.52%) and the shortest reaction time (RT, 268.25 ± 55.50 ms), which were significantly superior to AMP (accuracy, 97.33% ± 2.89%; RT, 398.60 ± 56.42 ms; both $P < 0.001$ after Bonferroni correction), PMP (accuracy, 92.94% ± 5.44%; RT, 494.21 ± 54.85 ms; both $P < 0.001$ after Bonferroni correction), and MMP (accuracy, 89.67% ± 7.66%; RT, 530.22 ± 57.45 ms; both $P < 0.001$ after Bonferroni correction) (Fig. 2A, B). From the number of options, PMP was the most difficult task because participants had to address three options, while there were only two options for AMP and MMP. However, the behavioral performance of MMP was, on the contrary, worse than that of PMP. Specifically, MMP had a longer RT ($P < 0.001$) and lower accuracy than PMP ($P = 0.105$ after Bonferroni correction). This can only be explained by a wrong prediction slowing the reaction, which indicated that the prediction state played a greater role than task difficulty here. In addition, AMP had significantly higher accuracy and a shorter RT than PMP (both $P < 0.001$ after Bonferroni correction) and MMP (both $P < 0.001$ after Bonferroni correction), which further demonstrated that a correct prediction speeds up the reaction [33]. Therefore, the prediction state indeed played a crucial role in responding

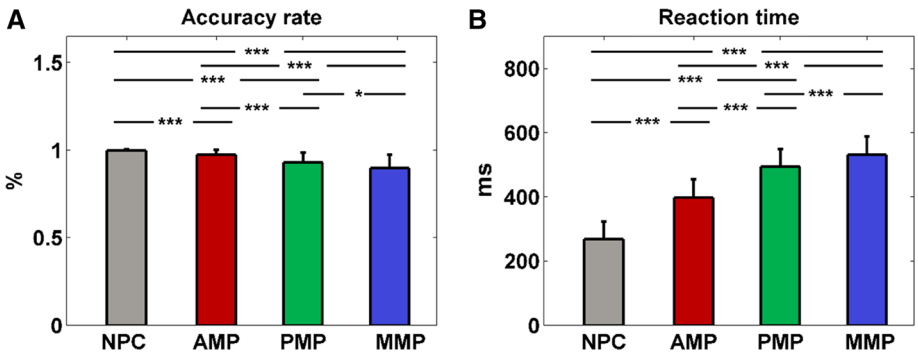

Fig. 2 Accuracy rate (**A**) and reaction time (**B**) for different conditions. Comparison of the evoked EEG energy of the whole predictive response. Evoked energies in the delta and theta bands were summed from 0 s to 2.4 s after the first flash onset. Vertical lines, error bars; $*P < 0.05$, $**P < 0.01$, $***P < 0.001$ after Bonferroni correction.

to the SOA400 double-flash. The setting of three but not two options for PMP was to rule out the risk that participants could mistakenly treat the task in the same way as AMP and MMP, because judging which one the SOA was from two options might be replaced by judging whether the SOA was familiar and expected or not, as the latter was easier, which required participants to remember only one SOA.

Analyses of Evoked EEG Energy

Expectation suppression (neural responses suppressed by correct compared to the incorrect predictions) has been widely demonstrated in blood-oxygen-level-dependent (BOLD) imaging studies [12, 14, 30]. However, electrophysiological evidence is still lacking. Here, we measured the evoked EEG energy of the whole predictive response to the SOA400 double-flash (the background EEG energy estimated by the baseline before the first flash was removed from the total energy of the brain responses after stimulus onset) to assess the degree of neural activation for different conditions. We found that the predictive response under matching prediction induced significantly less energy than that under mismatching prediction (F (2,88) = 5.607, P = 0.005), especially in the low-frequency band (< 13 Hz). Specifically, the evoked energy for AMP was smaller than MMP (AMP *vs* PMP P = 0.012 after Bonferroni correction; Fig. 3). However, there were no significant differences among the three conditions for the other frequency bands (details in supplementary materials). Therefore, the overall neural activation in the early sensory regions was suppressed for the correct compared to the incorrect prediction, in line with the BOLD studies.

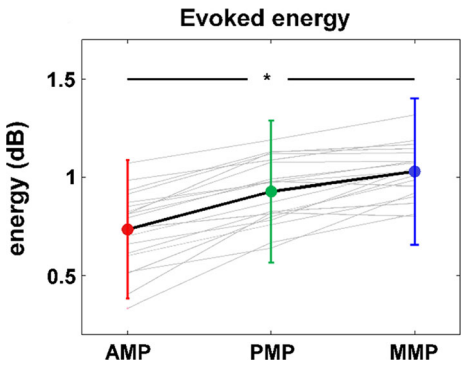

Fig. 3 Analyses on the evoked EEG energy of brain responses.

ERP Analyses

We compared the posterior ERPs of AMP, PMP, and MMP, and found the variations of N1 and N2 induced by the second flash were relevant to the process of timing prediction. To be specific, the SOA400 double-flash elicited two successive ERPs with a time interval of 400 ms (Fig. 4A). Since the upcoming moment of the first flash was unpredictable for all conditions, their responses would be the same. As expected, the first grand average ERPs following the first flash did not markedly differ across the three conditions, and their topographies were remarkably similar (Fig. 4D, F). Component analyses demonstrated no significant differences among the three conditions for the first N1 (F (2,88) = 0.793, P = 0.456) and N2 (F (2,88) = 0.776, P = 0.463) potentials in the posterior area. However, the grand average ERPs following the second flash showed considerable differences after removal of their different baselines caused by distinct CNVs (Fig. 4B, the reason for removing the CNVs is explained in the discussion). Topographic analyses revealed that the second N1 component was the largest for AMP on the whole posterior scalp, medium for PMP, and smallest for MMP, which were positively related to predictability (Fig. 4E). On the contrary, the second N2 component was largest for MMP, medium for PMP, and smallest for AMP, which were negatively related to the predictability (Fig. 4G). Component analyses further demonstrated that the two potentials had an opposite changing trend, i.e. decreasing N1 but increasing N2 against unpredictability (Fig. 4H, I). Such amplitude changes among the three conditions were significant for both N1 (F (2,88) = 5.367, P = 0.006; AMP *vs* MMP: P = 0.009; AMP *vs* PMP: P = 0.537; PMP *vs* MMP:

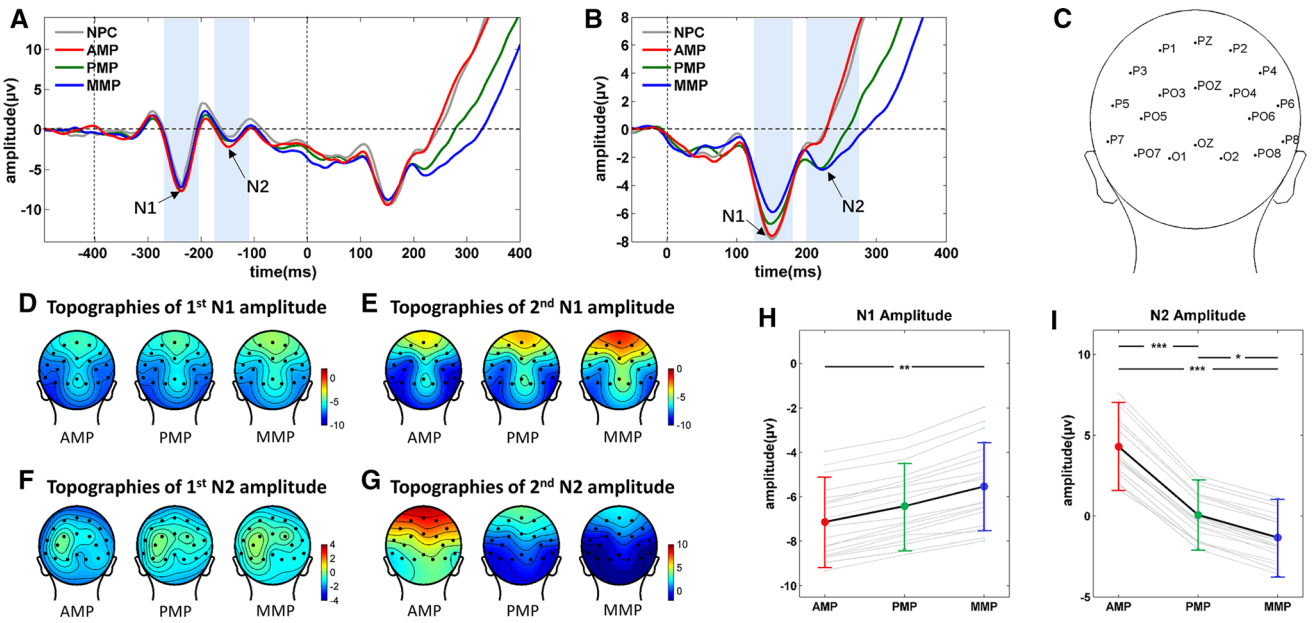

Fig. 4 ERP analyses. **A** Grand average ERPs induced by the SOA400 double-flash, across participants and electrodes on the posterior scalp under NPC, AMP, PMP, and MMP conditions (time zero is defined as the moment when the second flash is triggered). **B** Grand average ERPs induced by the second flash of the SOA400 double-flash (baseline averaged between − 50 ms and 0 ms is removed). **C** Names and locations of electrodes used. **D**–**G** Amplitude topographies of the first N1 (**D**), second N1 (**E**), first N2 (**F**), and second N2 (**G**) under distinct conditions. **H** N1 amplitudes induced by the second flash for AMP, PMP, and MMP (small dots, amplitude at each electrode; large dots, average amplitude). **I** N2 amplitudes induced by the second flash for AMP, PMP, and MMP. Vertical lines, error bars; $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ (*post hoc* tests after Bonferroni correction).

$P = 0.071$ all after Bonferroni correction) and N2 (F (2,88) = 40.907, $P < 0.001$; PMP *vs* MMP: $P = 0.035$; others: $P < 0.001$, all after Bonferroni correction). Then the second posterior scalp ERPs of AMP were compared to those of the NPC. Since the NPC had no timing prediction and needed no comparisons between the predictive code and sensory input, it was the least influenced by the top-down factors, which could act as a baseline for studying the neural effects of timing prediction. The two ERP waveforms were almost identical; no statistical difference was found in both the amplitudes (N1: T (44) = 0.307, $P = 0.760$; N2: T (44) = − 0.620, $P = 0.539$) and latencies (N1: T (44) = 1.366, $P = 0.179$; N2: T (44) = 0.878, $P = 0.385$) (Fig. 5A–E).

The above phenomena indicated that both N1 and N2 were modulated by prediction but in different manners. Specifically, compared to the baseline of NPC, the amplitude decreased for N1 while increasing for N2 with the increase of unpredictability, i.e., concurrent changes of N1 and N2 in opposite trends with the modulation of precise timing prediction.

It should be noted that a slow positive waveform followed the N2 component. It might be argued that this waveform was caused by the button-press, so the N2 variation recorded here was also due to the button-press rather than by timing prediction. However, this is not the case. First, if the N2 variation was caused by the button-press, the N2 latency would be closely associated with the RT. As the RT of NPC was the shortest of all ($P < 0.001$), accordingly, it should have much shorter latency than the other conditions. However, the N2 latencies of NPC and AMP were almost identical ($P = 0.385$), regardless of the significant differences in RT. Correlation analysis further showed no correlation between the N2 latency and RT ($r < 0.01$, $P = 0.96$; Fig. 5F), indicating that the N2 variations were not caused by the button-press. Second, the component analyzed here occurred in the early sensory processing stage (200 ms–260 ms after target onset), of which the defined temporal window was similar to the typical error-related component reported in a previous study [39], rather than that caused by the button-press or its preparation, which mainly emerged at the late sensory processing stage as a slow waveform [40, 41]. In addition, in order to further confirm that the N2 component measured here was dissociated from the subsequent slow waveform, we also applied the jackknife-based scoring method [42, 43], which defined the N2 time window according to the length of the negative-going component (AMP, 205 ms–225 ms; PMP, 205 ms–250 ms; MMP, 205 ms–275 ms) and obtained similar results (details in Fig. S3).

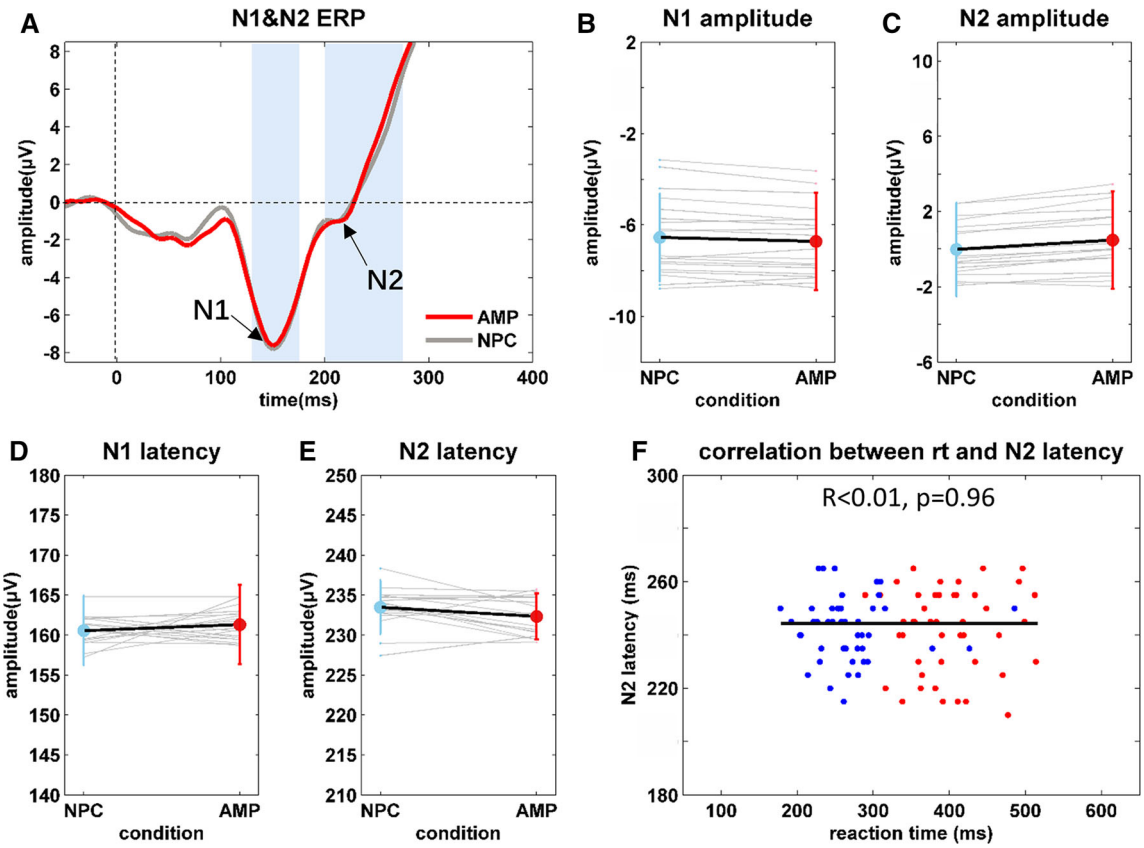

Fig. 5 Second ERPs for NPC and AMP. **A** Grand averages of the second ERP for NPC and AMP. **B**, **C** Comparison of the second N1 (**B**) and N2 (**C**) amplitude under NPC and AMP conditions (small dots, average N1/N2 amplitude at each electrode across all participants; large dots, average across all participants and all electrodes in the posterior area; vertical lines, error bars). **D**, **E** Comparison of the latency of the second N1 (**D**) and N2 (**E**) components. **F** Correlation analysis between RT and N2 latency (blue dots, RT and N2 latency in the NPC condition for each person; red dots, those in the AMP condition).

ITC and ERSP Analyses

Next, we analyzed the second posterior ERP using the ITC and ERSP techniques (Fig. 6A, C). As a result, both ITC and the evoked power of N1 in the theta band decreased from AMP to PMP to MMP, consistent with the above ERP results. Specifically, the theta-band ITC had a significant decreasing trend from AMP to PMP to MMP (Fig. 6B; F (2,88) = 15.827; $P < 0.001$; AMP *vs* PMP: $P = 0.031$; others: $P < 0.001$, all after Bonferroni correction). However, for the first N1, the theta band ITC showed no significant differences among conditions (F (2,88) = 1.021; $P = 0.364$). Furthermore, no significant differences in ITC were found in the other frequency bands for both the first and second N1 components (Fig. S4). The ERSP results were similar to those with ITC. The theta-band ERSP of the second N1 was largest for AMP, medium for PMP, and smallest for MMP (Fig. 6D); the differences were significant (F (2, 88) = 7.550; $P = 0.001$; AMP *vs* PMP: $P = 0.027$; AMP *vs* MMP: $P = 0.002$; PMP *vs* MMP: $P = 0.998$; all after Bonferroni correction). However, there was neither a significant ERSP difference in the theta band for the first N1 (F (2, 88) = 1.183; $P = 0.311$) nor in other frequency bands for both the first and second N1 components (Fig. S5). Here, we could not analyze ITC and ERSP of the second N2 because the small N2 components were smeared by the following large P3 potentials due to the restricted resolution of time-frequency transformation. Therefore, greater theta-band neural activity would be elicited during the sensory matching process for a matching prediction than a mismatching prediction.

Discussion

Consistent with previous BOLD studies, in this study we found the phenomenon of expectation suppression, i.e. the evoked EEG energy was lower for the matching than the mismatching prediction. Furthermore, the dynamic process of timing prediction was revealed by ERP analyses. As a result, compared to the control condition of no timing prediction, the N1 potential maintained almost the same

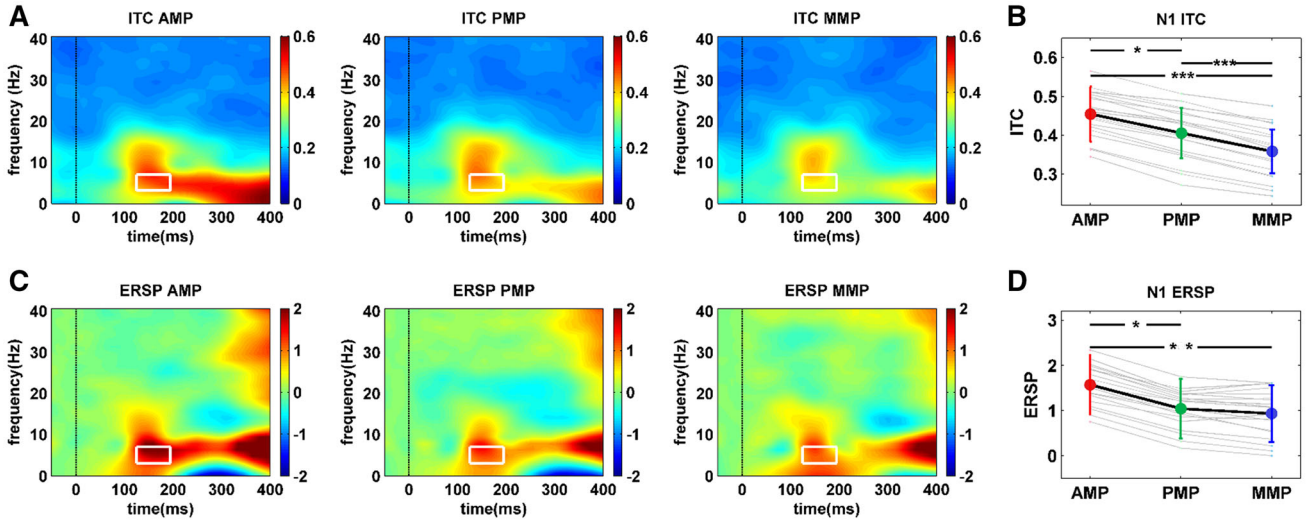

Fig. 6 Time-frequency decomposition analyses. **A** ITC for AMP, PMP, and MMP (white boxes, theta-band N1). **B** Comparison of N1 ITC in the theta band. **C** ERSP for AMP, PMP, and MMP (white boxes, theta-band N1). **D** Comparison of the N1 ERSP in the theta band. Vertical lines, error bars; $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ (*post hoc* tests after Bonferroni correction).

profile for the expected condition but was suppressed for the unexpected condition. This indicates that prediction 'sharpens' the expected sensory input during the N1 period. However, N2 was enhanced for the unexpected compared to the expected condition. This indicates that prediction 'dampens' the expected sensory information and encodes the prediction error. Therefore, underlying the phenomenon of expectation suppression, the 'sharpening' and 'dampening' effects work together but in separate time windows, providing direct neural evidence for the opposing process theory [27].

Attention Cannot Account for Our Results

We found an enhanced N1 in the matching prediction compared to the mismatching prediction. As attention has a facilitatory effect similar to prediction [44] and can be oriented in time [45], it may be argued that the enhanced matching N1 can be alternatively accounted for by temporal attention, rather than solely by timing prediction. It is possible that participants may orient attention to the predicted moment they need to discriminate, so more attentional resources would be allocated to the second flash of SOA400 for AMP than others. However, we argue this is not the case. Actually, participants paid equal attention to the second flash for all conditions, for four reasons. First, as the brain measures sub-second intervals by an automatic timing system which can measure time without attentional modulation [46], the participants would maintain the same attentional level during the whole double-flash, i.e. the second flash attracted as much attention as the first flash for each trial. As the first ERPs of the three conditions had the same profile, we can conclude that the participants were at the same attentional level for the first flash. Therefore, their attentional level would also be the same for the second flash. Second, according to previous studies [47, 48], if more attention was paid to the second flash for AMP, its visual N1 would be larger than that of the NPC. However, the two conditions had almost the same profiles for the second ERPs. Therefore, it was not for AMP to enhance but for PMP and MMP to attenuate the second N1, which showed a sharpening effect of prediction. Third, it has been demonstrated that the alpha band oscillation is closely relevant to the allocation of attention [49–51], while the theta oscillation is relevant to timing prediction [49, 52]. The ITC and ERSP results only showed significant differences in the theta band among conditions (Fig. 5), but not in the alpha band (ITC: F (2,88) = 1.654, P = 0.288; ERSP: F (2,88) = 1.358, P = 0.302), which indicated there were no differences in attention but differences of prediction deployed on the second flash for the three conditions. Fourth, this study not only included the condition of no timing prediction, but also the condition where participants knew there were only three kinds of stimuli and were required to react to the stimuli as fast as possible (PMP condition). As participants had known the content of stimuli and needed to press the button as quickly as possible in PMP, if attention does play a leading role in this process, the SOA400 would be the first moment on which they would allocate more attention. If so, the PMP N1 would be larger than that of other prediction-related conditions. However, the PMP N1 induced by SOA400 was smaller, which goes against the assumption of attentional modulation. In sum, the enhanced matching N1 cannot be accounted for by an attentional effect.

Dynamic Process of Brain Responses Modulated by Precise Timing Prediction

We investigated the dynamic neural responses to the identical double-flash with different timing predictions. Clearly, the CNVs in the midst of a double-flash differed among conditions (Fig. 4A). Previous studies have suggested that the CNV is closely relevant to temporal expectation [53]. However, recent findings have demonstrated that the timing of intervals does not depend on increasing neural activity but is more relevant to evoked potentials [54]. Therefore, we only focused on the evoked potential on which the precise timing template would be imposed. By studying the neural signatures of the evoked responses, we found that the successive N1 and N2 components were modulated by timing prediction but in opposite tendencies. Such modulations reflect the interactive process of the underlying representational and error signals during the predictive brain response. Predictive coding theory states that the external sensory input is matched with the internal perceptual template, resulting in a residual error signal to adjust the initial predictive state [9, 10]. Although previous evidence has demonstrated the existence of representational and error signals [1, 24–26, 55], evidence for the temporal dynamics of the two signals is still lacking. The results of our study show how the two signals wax and wane over time. Compared to the control condition of no timing prediction, a mismatching prediction attenuates the posterior N1 but enhances the following N2, while a matching prediction has no significant effect on the profiles. As N1 is linked to visual discrimination processes [56] and N2 is referred to as a mismatch detector [39], the results indicate that the representational signal during the N1 period is inhibited for unexpected sensory input but maintained for expected sensory input, while an error signal during the N2 period is produced for the unexpected sensory input but absent for the expected sensory input. This finding can help shed light on the neural mechanism of expectation suppression, as discussed below.

Reconciling the 'Sharpening' and 'Dampening' Accounts for Expectation Suppression

The neural effect of prediction is still poorly understood. Although the phenomenon of expectation suppression has been widely reported in previous studies, its underlying neural mechanism is still under debate. Currently, there are two paradoxical accounts for how the brain suppresses the neural response to predictable sensory input. The 'dampening' account proposes that the predictive templates in higher regions are able to dampen or 'explain away' the predictable sensory inputs in lower regions. Therefore, compared to mismatching prediction, matching prediction needs to address fewer inconsistent signals between predictive templates and sensory inputs. Accordingly, the neural activation would be lower for matching prediction than mismatching prediction [12–14, 57]. However, the 'sharpening' account holds the view that the suppressed neural activity of matching prediction is formed mainly by suppressing the neurons encoding the unpredicted, rather than those encoding the predicted information [3, 58], making the predicted information more salient and distinctively represented in a specific area of cortex [22]. We investigated the dynamic process of brain response modulated by the precise timing prediction, and found the phenomenon of expectation suppression by calculating the evoked EEG energy, which was in line with the BOLD studies. By analyzing the dynamic ERP process, we found the representational signal was significantly suppressed for mismatching prediction during the N1 period, which supported the 'sharpening' account. Furthermore, the error signal was almost absent for matching prediction during the N2 period, which supported the 'dampening' account. The results indicate that the neural representation of predicted information is sharpened during the N1 period and dampened during the N2 period. Therefore, the 'sharpening' and 'dampening' effects, which seem to contradict each other, are compatible during the dynamic prediction process. They work together but in separate time windows. These findings fit nicely with a recent proposed notion of opposing perceptual processes [27], which claims that the expected and unexpected events are addressed separately in time by applying Bayesian ('sharpening') and cancellation ('dampening') models, sequentially.

In sum, by using an innovative experimental paradigm, we investigated the fine temporal evolution of the evoked neural responses associated with precise timing prediction. We not only found the phenomenon of expectation suppression as in previous studies, but also found the 'sharpening' and 'dampening' effects of prediction in distinct processing stages with an opposing trend. These results provide direct neural evidence for the opposing processing theory, which is a potential theoretical resolution for the 'sharpening' and 'dampening' effects of prediction.

Acknowledgements This work was supported by the National Key Research and Development Program of China (2017YFB1300302), the National Natural Science Foundation of China (81925020 and 61976152), and the Young Elite Scientist Sponsorship Program of the China Association for Science and Technology (2018QNRC001).

Conflict of interest The authors declare no conflict of interest.

References

1. Egner T, Monti JM, Summerfield C. Expectation and surprise determine neural population responses in the ventral visual stream. J Neurosci 2010, 30: 16601–16608.
2. Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci 2013, 36: 181–204.
3. de Lange FP, Heilbron M, Kok P. How do expectations shape perception? Trends Cogn Sci 2018, 22: 764–779.
4. Adams RA, Shipp S, Friston KJ. Predictions not commands: active inference in the motor system. Brain Struct Funct 2013, 218: 611–643.
5. Cope TE, Sohoglu E, Sedley W, Patterson K, Jones PS, Wiggins J, *et al.* Evidence for causal top-down frontal contributions to predictive processes in speech perception. Nat Commun 2017, 8: 2154.
6. Pine A, Sadeh N, Ben-Yakov A, Dudai Y, Mendelsohn A. Knowledge acquisition is governed by striatal prediction errors. Nat Commun 2018, 9: 1673.
7. Summerfield C, de Lange FP. Expectation in perceptual decision making: neural and computational mechanisms. Nat Rev Neurosci 2014, 15: 745–756.
8. Sun H, Ma X, Tang L, Han J, Zhao Y, Xu X, *et al.* Modulation of beta oscillations for implicit motor timing in primate sensorimotor cortex during movement preparation. Neurosci Bull 2019, 35: 826–840.
9. Rao RPN, Ballard DHJNN. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 1999, 2: 79–87.
10. Friston K. A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci 2005, 360: 815–836.
11. Friston K. The free-energy principle: a unified brain theory? Nat Rev Neurosci 2010, 11: 127–138.
12. Alink A, Schwiedrzik CM, Kohler A, Singer W, Muckli L. Stimulus predictability reduces responses in primary visual cortex. J Neurosci 2010, 30: 2960–2966.
13. Barne LC, Claessens PME, Reyes MB, Caetano MS, Cravo AM. Low-frequency cortical oscillations are modulated by temporal prediction and temporal error coding. Neuroimage 2017, 146: 40–46.
14. Blank H, Davis MH. Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception. PLoS Biol 2016, 14: e1002577.
15. Ullsperger M, Fischer AG, Nigbur R, Endrass T. Neural mechanisms and temporal dynamics of performance monitoring. Trends Cogn Sci 2014, 18: 259–267.
16. Richter D, Ekman M, Lange FPDJJoN. Suppressed sensory response to predictable object stimuli throughout the ventral visual stream. J Neurosci 2018, 38:7452–7461.
17. Bueti D, Bahrami B, Walsh V, Rees G. Encoding of temporal probabilities in the human brain. J Neurosci 2010, 30: 4343–4352.
18. Doherty JR, Rao A, Mesulam MM, Nobre AC. Synergistic effect of combined temporal and spatial expectations on visual attention. J Neurosci 2005, 25: 8259–8266.
19. Jaramillo S, Zador AM. The auditory cortex mediates the perceptual effects of acoustic temporal expectation. Nat Neurosci 2011, 14: 246–251.
20. Kok P, Brouwer GJ, van Gerven MA, de Lange FP. Prior expectations bias sensory representations in visual cortex. J Neurosci 2013, 33: 16275–16284.
21. Kouider S, Long B, Le Stanc L, Charron S, Fievet AC, Barbosa LS, *et al.* Neural dynamics of prediction and surprise in infants. Nat Commun 2015, 6: 8537.
22. Kok P, Jehee JF, de Lange FP. Less is more: expectation sharpens representations in the primary visual cortex. Neuron 2012, 75: 265–270.
23. Kok P, Rahnev D, Jehee JF, Lau HC, de Lange FP. Attention reverses the effect of prediction in silencing sensory signals. Cereb Cortex 2012, 22: 2197–2206.
24. de Gardelle V, Waszczuk M, Egner T, Summerfield C. Concurrent repetition enhancement and suppression responses in extrastriate visual cortex. Cereb Cortex 2013, 23: 2235–2244.
25. de Gardelle V, Stokes M, Johnen VM, Wyart V, Summerfield C. Overlapping multivoxel patterns for two levels of visual expectation. Front Hum Neurosci 2013, 7: 158.
26. Miller EK, Desimone R. Parallel neuronal mechanisms for short-term memory. Science 1994, 263: 520–522.
27. Press C, Kok P, Yon D. The perceptual prediction paradox. Trends Cogn Sci 2020, 24: 13–24.
28. Ekman M, Kok P, de Lange FP. Time-compressed preplay of anticipated events in human primary visual cortex. Nat Commun 2017, 8: 15276.
29. Kok P, Mostert P, de Lange FP. Prior expectations induce prestimulus sensory templates. Proc Natl Acad Sci U S A 2017, 114: 10473–10478.
30. Summerfield C, Egner T, Greene M, Koechlin E, Mangels J, Hirsch J. Predictive codes for forthcoming perception in the frontal cortex. Science 2006, 314: 1311–1314.
31. Bueti D, Lasaponara S, Cercignani M, Macaluso E. Learning about time: plastic changes and interindividual brain differences. Neuron 2012, 75: 725–737.
32. Arnal LH, Doelling KB, Poeppel D. Delta-beta coupled oscillations underlie temporal prediction accuracy. Cereb Cortex 2015, 25: 3077–3085.
33. Stefanics G, Hangya B, Hernadi I, Winkler I, Lakatos P, Ulbert I. Phase entrainment of human delta oscillations can mediate the effects of expectation on reaction speed. J Neurosci 2010, 30: 13578–13585.
34. Summerfield C, Koechlin E. A neural representation of prior information during perceptual inference. Neuron 2008, 59: 336–347.
35. Calderone DJ, Lakatos P, Butler PD, Castellanos FX. Entrainment of neural oscillations as a modifiable substrate of attention. Trends Cogn Sci 2014, 18: 300–309.
36. Lakatos P, Karmos G, Mehta AD, Ulbert I, Schroeder CE. Entrainment of neuronal oscillations as a mechanism of attentional selection. Science 2008, 320: 110–113.
37. Melloni L, Schwiedrzik CM, Muller N, Rodriguez E, Singer W. Expectations change the signatures and timing of electrophysiological correlates of perceptual awareness. J Neurosci 2011, 31: 1386–1396.
38. Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods 2004, 134: 9–21.
39. Folstein JR, Van PCJP. Influence of cognitive control and mismatch on the N2 component of the ERP: a review. Psychophysiology 2008, 45: 152–170.
40. Brunia CHM, Damen EJP. Distribution of slow brain potentials related to motor preparation and stimulus anticipation in a time estimation task. Electroencephalogr Clin Neurophysiol 1988, 69: 234–243.
41. Luck SJ, Kappenman ES (Eds). The Oxford Handbook of Event-Related Potential Components. Oxford Library of Psychology 2008.
42. Stahl J, Gibbons HJP. The application of jackknife-based onset detection of lateralized readiness potential in correlative approaches. Psychophysiology 2004, 41: 845–860.

43. Miller J, Patterson AT, Ulrichb R. Jackknife-based method for measuring LRP onset latency differences. Psychophysiology 1998, 35: 99–115.
44. Summerfield C, Egner T. Expectation (and attention) in visual cognition. Trends Cogn Sci 2009, 13: 403–409.
45. Nobre AC, van Ede F. Anticipated moments: temporal structure in attention. Nat Rev Neurosci 2018, 19: 34–48.
46. Buhusi CV, Meck WH. What makes us tick? Functional and neural mechanisms of interval timing. Nat Rev Neurosci 2005, 6: 755–765.
47. Hillyard SA, Anllo-Vento L. Event-related brain potentials in the study of visual selective attention. Proc Natl Acad Sci U S A 1998, 95: 781–787.
48. Stormer VS, McDonald JJ, Hillyard SA. Cross-modal cueing of attention alters appearance and early cortical processing of visual stimuli. Proc Natl Acad Sci U S A 2009, 106: 22456–22461.
49. Arnal LH, Giraud AL. Cortical oscillations and sensory predictions. Trends Cogn Sci 2012, 16: 390–398.
50. Capilla A, Schoffelen JM, Paterson G, Thut G, Gross J. Dissociated alpha-band modulations in the dorsal and ventral visual pathways in visuospatial attention and perception. Cereb Cortex 2014, 24: 550–561.
51. Klimesch W. alpha-band oscillations, attention, and controlled access to stored information. Trends Cogn Sci 2012, 16: 606–617.
52. Cravo AM, Rohenkohl G, Wyart V, Nobre AC. Endogenous modulation of low frequency oscillations by temporal expectations. J Neurophysiol 2011, 106: 2964–2972.
53. van Rijn H, Kononowicz TW, Meck WH, Ng KK, Penney TB. Contingent negative variation and its relation to time estimation: a theoretical evaluation. Front Integr Neurosci 2011, 5: 91.
54. Kononowicz TW, van Rijn H. Decoupling interval timing and climbing neural activity: a dissociation between CNV and N1P2 amplitudes. J Neurosci 2014, 34: 2931–2939.
55. Ni B, Wu R, Yu T, Zhu H, Li Y, Liu Z. Role of the hippocampus in distinct memory traces: timing of match and mismatch enhancement revealed by intracranial recording. Neurosci Bull 2017, 33: 664–674.
56. Mangun GR, Hillyard SA. Modulations of sensory-evoked brain potentials indicate changes in perceptual processing during visual-spatial priming. J Exp Psychol Hum Percept Perform 1991, 17: 1057–1074.
57. Summerfield C, Trittschuh EH, Monti JM, Mesulam MM, Egner T. Neural repetition suppression reflects fulfilled perceptual expectations. Nat Neurosci 2008, 11: 1004–1006.
58. Chalk M, Marre O, Tkacik G. Toward a unified theory of efficient, predictive, and sparse coding. Proc Natl Acad Sci U S A 2018, 115: 186–191.