Sequential Temporal Anticipation Characterized by Neural Power Modulation and in Recurrent Neural Networks

Authors: Xiangbin Teng [a, b, 1, *], Ru-Yuan Zhang [c, d, 1]

Affiliation:

[a] Neuroscience Department, Max-Planck-Institute for Empirical Aesthetics, 60322 Frankfurt, Germany

[b] Department of Education and Psychology, Freie Universität Berlin, 14195 Berlin, Germany

[c] Institute of Psychology and Behavioral Science, Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai, China. 200030

[d] Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, Shanghai Jiao Tong University, Shanghai, China. 200030

[1] These authors contributed equally

*** Corresponding author**:

Xiangbin Teng
Department of Education and Psychology
Freie Universität Berlin
Habelschwerdter Allee 45
14195 Berlin, Germany
xiangbin.teng@gmail.com

Conflict of Interest: The authors declare no competing financial interests.

Acknowledgement: We thank Johannes Messerschmidt, Anna-Lena Possmayer, and Dominik Thiele for their assistance with data collection. We thank Xing Tian and Nai Ding for their comments on a previous version of the manuscript. We thank David Poeppel for support.

Funding: This work was supported by the Max-Planck-Society. R.Y.Z was supported by Shanghai Pujiang Program (21PJ1407800).

Author contributions:
Conceived and designed the experiments: X.T & R.Y.Z
Performed the experiments: X.T
Analyzed the data: X.T
Contributed reagents/materials/analysis tools: X.T & R.Y.Z
Wrote the first draft of the paper: X.T
Wrote the paper: X.T & R.Y.Z

Abstract

Complex human behaviors involve perceiving continuous stimuli and planning actions at sequential time points, such as in perceiving/producing speech and music. To guide adaptive behavior, the brain needs to internally anticipate a sequence of prospective moments. How does the brain achieve this sequential temporal anticipation without relying on any external timing cues? To answer this question, we designed a 'premembering' task: we tagged three temporal locations in white noise by asking human listeners to detect a tone presented at one of the temporal locations. We selectively probed the anticipating processes guided by memory in trials with only flat noise using novel modulation analyses. A multi-scale anticipating scheme was revealed: the neural power modulation in the delta band encodes noise duration on a supra-second scale; the modulations in the alpha-beta band range mark the tagged temporal locations on a sub-second scale and correlate with tone detection performance. To unveil the functional role of those neural observations, we turned to recurrent neural networks (RNNs) optimized for the behavioral task. The RNN hidden dynamics resembled the neural modulations; further analyses and perturbations on RNNs suggest that the neural power modulations in the alpha-beta band emerged as a result of selectively suppressing irrelevant noise periods and increasing sensitivity to the anticipated temporal locations. Our neural, behavioral, and modelling findings convergingly demonstrate that the sequential temporal anticipation involves a process of dynamic gain control – to anticipate a few meaningful moments is also to actively ignore irrelevant events that happen most of the time.

Keywords: temporal attention, neural dynamics, memory-driven prediction, neural gain modulation, temporal sparseness

Introduction

It is essential in our daily life to navigate in time (Buzsáki and Llinás 2017; Issa et al. 2020). Estimating when a traffic light will change to red and avoiding an incoming car involves anticipating the timing of forthcoming events. It is even more natural for humans to perceive continuous stimuli and to plan a sequence of actions in a series of prospective moments, such as in communicating speech and music (Nobre and van Ede 2018; Rimmele et al. 2018). In those cases, a sequence of prospective moments is required to be anticipated to guide proactive sensory and motor preparation. How does the brain internally construct a temporal template to anticipate a sequence of temporal locations?

It is not enough to consider the temporal anticipation over one timescale in a natural setting where natural stimuli and actions are organized in nested hierarchical structures across multiple timescales (Lashley 1951). On timescales from tens of milliseconds to several seconds, for example, durations of words in a sentence vary with the sentence length; the timing of words is constraint by the sentence length. Communicating speech sequences is based on not only timing word onsets on a local timescale, but also encoding the sentence length on a global timescale (Teng et al. 2017; Ghitza 2012; Ding et al. 2016; Poeppel and Assaneo 2020; Teng, Tian, and Poeppel 2016; Teng et al. 2020; Donhauser and Baillet 2020). This interaction between timescales also exists in music and Morse code comprised of hierarchical structures (Huron 2008; Lerdahl and Jackendoff 1983; Larrouy-Maestri and Pfordresher 2018; Paton and Buonomano 2018; Teng, Larrouy-Maestri, and Poeppel 2021). Consequently, the temporal anticipation of a sequence of prospective moments – the sequential temporal anticipation – requires to consider both the local timing of forthcoming events and the duration of the sequence, the temporal context (Nobre and Stokes 2019; Lee, Aly, and Baldassano 2021). In contrast, in anticipating an event onset, estimating the passage of an interval, or predicting the offset of a lasting event (Nobre, Correa, and Coull 2007; Nobre 2001; Coull and Nobre 1998; Martens and Johnson 2005; Niemi and Näätänen 1981; Samaha et al. 2015), the temporal anticipation is carried out on a single timescale – from a temporal cue to a target time point. Those prospective timing tasks only partially characterize the temporal anticipation.

One candidate experimental paradigm for studying the sequential temporal anticipation is to present a sequence of sensory events (e.g. tones or visual images) and to force human participants to anticipate the timing of the events (Haegens and Zion Golumbic 2018; Rimmele et al. 2018; Jones, Kidd, and Wetzel 1981; Zanto, Snyder, and Large 2006), but there is a problem: the sensory events provide external timing cues. When periodic

sequences are presented, the temporal anticipation can be based on neural oscillatory patterns entrained by the rhythmic sequences (Tal et al. 2017; van Bree et al. 2021; Arnal, Doelling, and Poeppel 2014; Morillon, Schroeder, and Wyart 2014; Morillon et al. 2016; Ten Oever et al. 2017; Herrmann et al. 2016; Henry, Herrmann, and Obleser 2014; Henry and Obleser 2012; Auksztulewicz et al. 2019; Spaak, de Lange, and Jensen 2014; Doelling and Poeppel 2015; Doelling et al. 2019; Assaneo et al. 2019). Even if intervals between events in a sequence are varied, the sequence can still be used as a temporal context for anticipating an incoming event after the temporal contingencies of the sequence are learned (Breska and Ivry 2020; Breska and Deouell 2017; Rimmele et al. 2018; Morillon et al. 2016; Wilsch et al. 2015; Paton and Buonomano 2018). For instance, the timing of an incoming event can be anticipated by referencing to the previous event as a temporal cue and estimating the passage of a fixed interval. Then, it is hard to pin down the sequential temporal anticipation initiated internally by the brain – a new paradigm is needed here.

To investigate how the brain internally anticipates a sequence of prospective moments without relying on any external timing cues, we adapted a simple tone-in-noise paradigm: human listeners were asked to detect a weak tone at one of three fixed temporal locations in a white noise piece; to better detect the weak target tone, human listeners must memorize the temporal locations and anticipate them sequentially in the flat white noise. This can be viewed as the memory-guided temporal anticipation (Janssen and Shadlen 2005; Nobre and van Ede 2018; Nobre and Stokes 2019; Cravo et al. 2017; Mattiesing et al. 2017; Grabenhorst et al. 2019), not for the timing of an individual event, but for three prospective moments. The white noise piece serves as a global temporal context that constrains the temporal locations of the target tone; the tone locations must be determined by referencing to the noise onset or the passage of the noise period. Previous studies have indirectly studied the temporal anticipation by measuring its effects on behavioral and neural responses to sensory events (Grabenhorst et al. 2019; Tavano, Schröger, and Kotz 2019) and by using an encoding framework (Herbst, Fiedler, and Obleser 2018), but we, here, examine directly how the brain carries out the sequential temporal anticipation internally in flat white noise.

It is challenging to identify internal anticipating processes without resorting to any external markers. Nevertheless, innovatively, we constructed a data-driven method – modulation spectrum analyses on group-level neural power and phase measurements. If all participants anticipate the same temporal structure framed in the current task design, their neural dynamics corresponding to the anticipating processes should be similar. Indeed, by identifying the shared neural modulation components across participants, we found a neural

modulation component in the delta band (1 – 4 Hz) encoding the noise duration and – unexpectedly, power modulations in the alpha-band band range (12 – 30 Hz) collectively marking the local temporal locations, which cannot be fully accounted for by expectation violation, trial structure, or false alarms. Lastly, to understand the functional role of the neural power modulation marking the temporal locations, we took advantage of recent developments in artificial intelligence and analyzed 12 recurrent neural networks (RNNs) that were trained on variations of the current behavioral task. By comparing the inner dynamics of the trained RNNs with the neural signatures of the temporal anticipation and by probing RNNs' behavior in different conditions, we revealed a mechanism of dynamic gain control in the sequential temporal anticipation and generated new hypotheses for future investigations.

Results

Tagging prospective temporal locations and behavioral results

To imprint human participants with prospective temporal locations, we presented a white noise piece of 2 s long in each trial and inserted a weak tone, whose magnitude was normalized to each participant's threshold of tone detection (Fig. S1A), at one of three time points: 500 ms, 1000 ms, and 1500 ms (Fig. 1A). This regular spacing between the temporal locations was designed to facilitate the identification of neural correlates of the sequential temporal anticipation in the white noise without resorting to any external markers (See the following and *Methods* for details). After the noise piece was over, the participants made a response to indicate whether they heard a tone. 25 percent of trials only included white noise pieces (i.e., no-tone trials); the target tone appeared at one of the three temporal locations with the same probability in the other 75 percent of trials (i.e., tone trials) (Fig. 1B&S1B). The order of the tone trials and the no-tone trials was randomized, so the participants did not know whether and when the target tone would appear. Consequently, to better detect the weak tone, the participants must memorize the temporal locations in the noise pieces and anticipate each temporal location. If this is the case, even though only flat white noise was presented, certain neural signatures representing the sequential anticipation of the temporal locations can be observed. This task was repeated in three experimental blocks, so that a learning effect on memorizing the temporal locations can be examined across blocks.

We first conducted traditional analyses on behavioral measurements and neural responses to the target tone, which replicated previous findings and showed that the manipulation of the timing of the target event was effective in our task and indeed modulated sensory processing of the target tone (Nobre, Correa, and Coull 2007; Ng, Schroeder, and Kayser 2012; Herbst, Fiedler, and Obleser 2018; Rimmele, Jolsvai, and Sussman 2011; Tavano, Schröger, and Kotz 2019) (Fig. 1C, S1C, S1D&S1E). Next, we specifically focused on the no-tone trials. The three temporal locations of the target tone were designed to be spaced 500 ms apart in the 2-s noise pieces. If the brain both encodes the temporal context, i.e., the elapsed time of noise piece, on a supra-second scale and anticipates the temporal locations on a sub-second scale, we should be able to identify neural signatures in power or phase measurement around 2 Hz and in the relatively low frequency range corresponding to the two timescales, respectively (Fig. 1D). Alternatively, the brain may only encode the elapsed time of noise piece for detecting the target tone, or the brain may simply register the interval length between the temporal locations and estimate one interval after another starting

from the noise onset. In either case, only one neural signature of a single timescale can be observed.

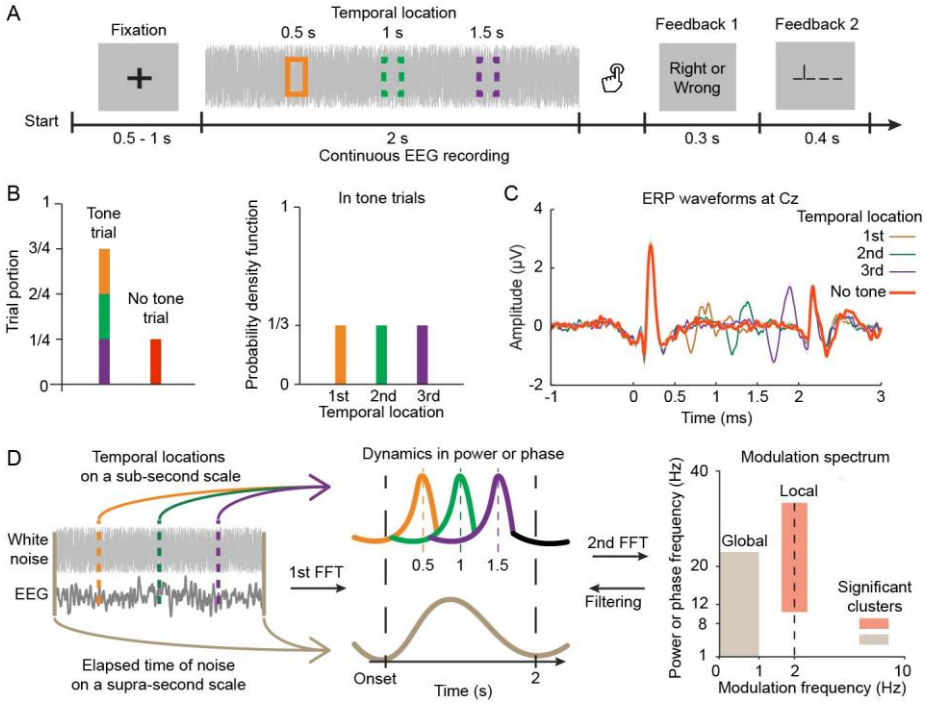

Fig. 1. Experimental paradigm, ERPs to the target tone, and analysis procedures. (*A*) Tone detection task. In each trial, participants were presented with a 2-s white noise piece and were asked to answer whether they heard a tone or not after the white noise was over. The target tone could appear in the noise piece at either 500 ms, 1000 ms, or 1500 ms (temporal locations are indicated by the colored boxes) – or not. Two visual feedbacks were provided: the first feedback indicated whether the responses were correct; the second feedback revealed the tone location. (*B*) Portions of trials and probability of the target tone. The target tone appeared in 75 percent of trials, i.e., tone trials, and the rest of trials only included white noise, i.e., no-tone trials (the left plot); in the tone trials, the target tone appeared at one of three temporal locations with the same probability of 0.33 (the right plot). The color scheme for the temporal locations matches the color scheme in (*A*). (*C*) ERP waveforms of the Cz channel. It can be seen that the target tone evoked larger neural responses at the later temporal locations. See Fig. S1D for more details. The ERP waveform in the no-tone trials did not show dynamics reflecting the temporal locations. (*D*) Hypotheses and analysis procedures. Left plot: in the noise pieces, participants would encode the noise duration and anticipate the temporal locations to better detect the target tone. Middle plot: we expected to observe neural signatures that reflect timing processes on a sub-second scale for the temporal locations and on a supra-second scale for the elapsed time of noise period. Right plot: to identify relevant neural components, we constructed a novel data-driven method. We first derived power or phase measurement of recorded EEG signals in the no-tone trials through a time-frequency decomposition (1st FFT), and then conducted another fast Fourier transformation (FFT) on the time-frequency decomposed neural data to derive a modulation spectrum. After surrogate tests, we expected to observe significant modulation clusters in the power or phase measurement corresponding the timescale of the temporal locations around 2 Hz and the timescale of the elapsed time of noise in low modulation frequency ranges. Then, we filtered the EEG data according to the spectral range of each cluster to derive neural signals in the temporal domain for further analyses.

Robust power modulations identified at 2 Hz and in the low frequency range

We tried to pin down robust neural modulations for the temporal anticipation in the power and phase measurements by employing a novel data-driven method – group-level modulation spectrum analysis (Fig. 1D). We first pooled all no-tone trials over the three blocks and, after the time-frequency decomposition (the 1st Fourier transform), derived induced power and inter-trial phase coherence (ITPC). Then, we applied the 2nd Fourier transform on each frequency of the induced power and the ITPC to derive their modulation spectra. We then conducted surrogate tests to determine significant clusters in the modulation spectra (see *Methods* for details).

The modulation spectrum analyses did reveal a rich profile of neural modulations in the modulation power spectrum (MPS), whereas traditional ERP and power analyses did not help isolate any meaningful neural signatures without post-hoc examination (Fig. 1C, 2A, 2B, S2A&S2B). We only found one cluster in the modulation spectrum of ITPC, which indicates the onset and offset responses to the noise pieces (Fig. S2D&S2E), so we did not further examine the neural phase domain. In the MPS, four significant modulation clusters were identified (Fig. 2B, right plot). As the induced power of 1.5 seconds (from 0.25 s to 1.75 s after the noise onset) was used in the MPS calculation, modulation components below 1.33 Hz did not construct two cycles over 1.5 seconds and probably represented ramping neural activities, whereas the modulation components above 1.33 Hz can be considered as rhythmic components. Hence, the modulation frequency of 1.33 Hz serves as a boundary to separate ramping components from rhythmic components – Trend bound (Fig. 2B, the left plot). Consequently, the two clusters (Cluster 3 and 4) above Trend bound likely represent the temporal anticipation of the local temporal locations. The topography of each cluster is shown in Fig. 2C. We next examined the temporal dynamics of those four clusters.

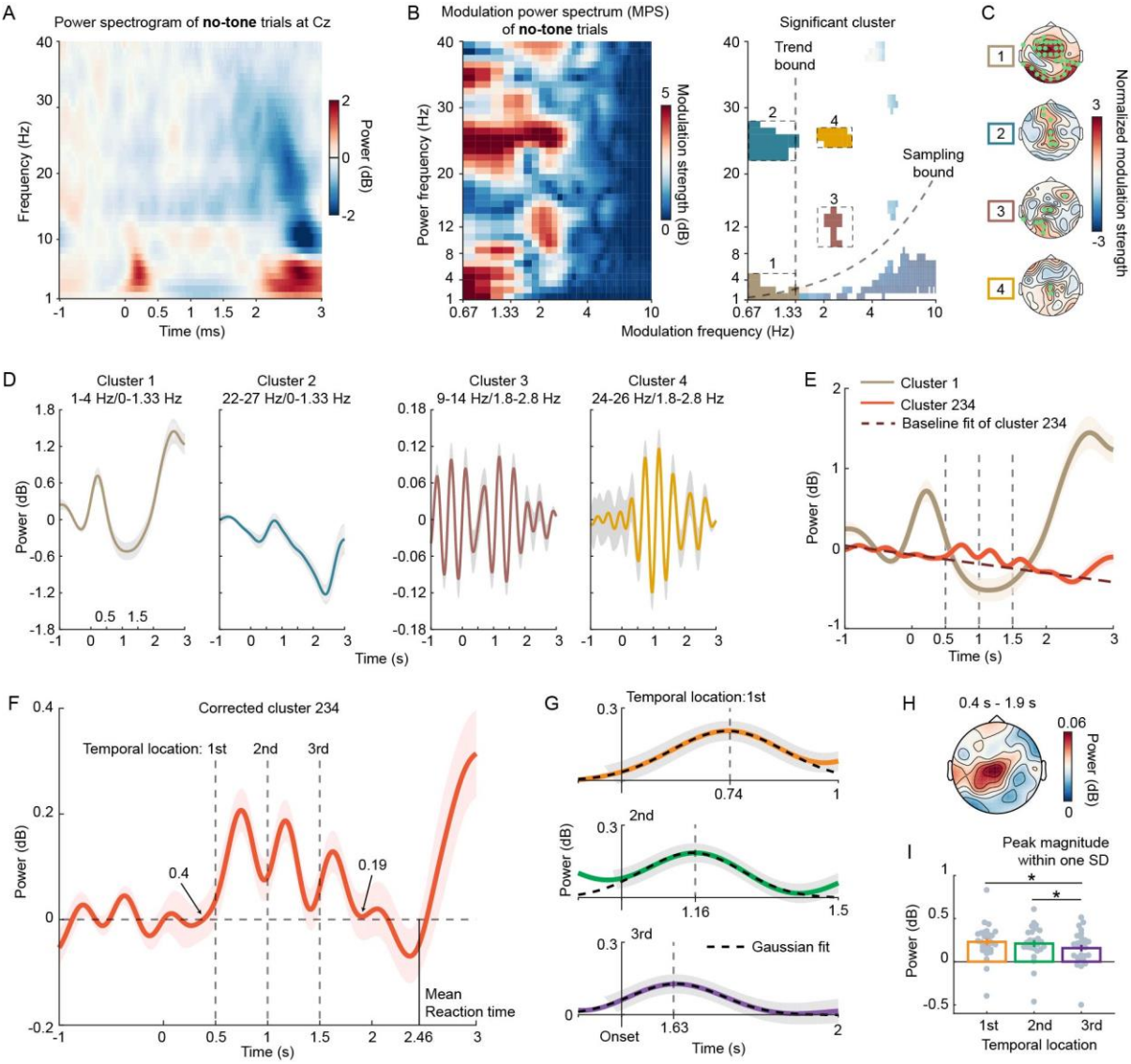

Fig. 2. Modulation power spectrum (MPS) and power dynamics. (*A*) Spectrogram of induced power of all no-tone trials. (*B*) Raw MPS (left plot) and significant clusters (right plot). We first conducted the modulation spectrum analysis on the EEG channel of Cz and then applied the same analysis on other channels. The y-axis denotes the frequency of induced power, i.e., power frequency; the x-axis denotes the modulation frequency. In the right plot, the colored bins refer to significant modulation bins with high modulation strength determined from a surrogate test; the four dashed boxes indicate the power and modulation frequency ranges selected for the clusters. The modulation frequencies larger than Trend bound represent rhythmic neural components and the modulation frequencies smaller than Trend bound ramping neural components. The modulation bins below Sampling bound violate the Nyquist theorem so they are not meaningful. (*C*) Topographies of significant clusters. The green dots denote significant EEG channels. (*D*) Dynamics of significant clusters. We filtered the induced power according to the clusters selected in (*B*) and show the power dynamics of each cluster. (*E*) Merged clusters. According to the power-power coupling between each pair of clusters (Fig. S2) and the power frequency overlaps, we grouped Cluster 2, 3, and 4 into one cluster – Cluster 234. Collectively, Cluster 1 and Cluster 234 captured both the local temporal locations and the global elapsed time of noise piece. (*F*) Corrected cluster 234. We derived Corrected cluster 234 by removing the baseline of Cluster 234 determined from -1 s to

0 s. (*G*) Gaussian fits to each neural peak of Corrected cluster 234. We derived peak latencies (mean of Gaussian fits) and peak widths (standard deviation). (*H*) Topography of modulation strength of Corrected cluster 234 from 0.4 s to 1.9 s after the noise onset. (*I*) Peak magnitude within one standard deviation (SD) of the Gaussian fits. Each gray dot represents each participant's data point. The shaded areas of color and the error bars represent ±1 standard error of the mean over participants. * represents $p < 0.05$.

We filtered the induced power to derive the power dynamics for each cluster according to the selected ranges of power frequency and modulation frequency in Fig. 2B (see *Methods*). The power frequency range of Cluster 1 is within the delta band (1 – 4 Hz) and the temporal modulation trajectory of Cluster 1 captures the duration of the noise pieces, suggesting that the brain registered mentally the length of the noise piece and actively monitored the elapsed time of noise pieces. This is consistent with previous findings that neural signals in the delta band or slow ramping activities correlate with interval (or event duration) estimation (van Wassenhove and Lecoutre 2015; Breska and Deouell 2017), which helps validate our novel method here. Cluster 2, 3, and 4 do not appear to convey any information on the local temporal locations. Since the Fourier transform arbitrarily breaks signals into orthogonal components (sinusoidal waves), we suspected that Cluster 2, 3, and 4 may belong to one complex neural signal, which was broken into parts by the MPS calculation. Hence, we quantified the coupling in power between each pair of clusters to examine how the clusters were related to each other. In addition to replicating a coupling relationship between the delta and the beta bands found in previous literature (Arnal, Doelling, and Poeppel 2014) (Fig. S2C), we found significant coupling relationships after the noise onset between Cluster 2 and 3 and between Cluster 3 and 4. Moreover, Clusters 2 and 4 fell in the same power frequency range. The coupling relationships and the power frequency overlaps between Cluster 2, 3, and 4 gave us a clue that the three clusters collectively represented a common neural process, so we grouped them into Cluster 234. In Fig. 2E, it is evident that Cluster 1 and Cluster 234 concurrently represent the globally elapsed time of noise piece and the local temporal locations.

We next mainly focused on Cluster 234, as it closely captured the temporal anticipation of the local temporal locations and is a neural signature that has not been clearly shown before, to our best knowledge. We removed the baseline of Cluster 234, a downward trend spanning over the entire trial, and plot the dynamics of Corrected cluster 234 in Fig. 2F and the topography of its modulation strength from 0.4 s to 1.9 s in Fig. 2H (see *Methods* for details). Clearly, the neural peaks in Corrected cluster 234 mark the three temporal locations in the noise pieces (Fig. 2F). We fitted a Gaussian curve to each peak and found the peak latencies, referenced to the corresponding temporal locations, to be 0.24 s, 0.16 s, and 0.13 s, respectively (Fig. 2G), suggesting a precession of neural dynamics along the temporal locations. We next averaged the peak magnitudes for each temporal location within one standard deviation (SD) of the corresponding Gaussian fit and conducted a one-way rmANOVA. A significant main effect for Temporal location was found ($F(2,52) = 4.30$, p

= .019, η_p^2 = 0.142). The peak magnitudes of the first temporal location and of the second temporal location are significantly larger than the third temporal location, respectively (Fig. 2I) (paired t-test: $t(26) = 3.38$, $p = .038$; $t(26) = 2.41$, $p = .045$). Adjusted False Discovery Rate (FDR) correction was used here and throughout the whole paper to control false positive rate of multiple comparisons (Yekutieli and Benjamini 1999; Benjamini and Hochberg 1995).

It is reasonable to suspect that Corrected cluster 234 in the no-tone trials reflected expectation violation – as the participants expected the target tone to appear at a temporal location but there was no tone, the resulted prediction error (Friston 2005) or the mismatch (Heilbron and Chait 2018; Näätänen et al. 2007) might induce the neural peaks of Corrected cluster 234. If this is the case, the magnitudes of the neural peaks of Corrected cluster 234 should increase from the first temporal location to the third temporal location as the likelihood for the target tone to appear increases with the elapsed time of noise piece (Nobre, Correa, and Coull 2007; Janssen and Shadlen 2005) (Fig. S1B). However, the data here show that the peak magnitude of Corrected cluster 234 decreased with the temporal location (Fig. 2I), which renders implausible the explanation based on the expectation violation. This decreasing trend of the peak magnitude probably correlates with the estimation of the likelihood of a trial being a tone trial (Grabenhorst et al. 2021) (Fig. S1B), but our following analyses of fitting this likelihood to individual participants' data did not reveal a significant effect ($p > 0.05$). Further studies are needed to specifically test this conjecture. Another possible explanation could be that, since the spacing between the temporal locations was highly regular and the noise length was also fixed, any neural observations could be simply caused by the fixed trial structure but not by the sequential temporal anticipation. But this is not the case, as we conducted another experimental session using the same trial structure with a cued temporal location for the target tone but did not find any robust power modulations around 2 Hz (Fig. S3).

We examined the power dynamics in the frequency bands defined routinely (Fig. S4), from which a vague sign of the induced power marking the temporal locations in the no-tone trials can be seen in the beta2 band (21 – 30 Hz) and in the alpha band (8 – 12 Hz) – after we deliberately divided the beta band (12 to 30 Hz) into the beta1 (13 – 20 Hz) and the beta2 bands (Fig. S4A). Our data-driven method here matched the routine analysis procedures but selectively extracted the key neural features from EEG signals without first plotting power dynamics in pre-defined frequency bands for observation and then conducting post-hoc analyses.

Here, through a novel data-driven analysis on the modulation spectrum of neural power, we replicated previous findings on duration estimation reflected in the delta band and identified a novel neural signature marking the local temporal locations. Timing processes were indeed carried out on both the supra-second scale and the sub-second scale to sequentially anticipate the temporal locations.

Correlation between behavioral measurements and Corrected cluster 234

In the above analyses, we grouped all the no-tone trials across the three experimental blocks to guarantee sufficient statistical power for the MPS analysis. Next, by looking into each block, we explored the correlation between behavioral performance and Corrected cluster 234 as well as the learning effect across blocks.

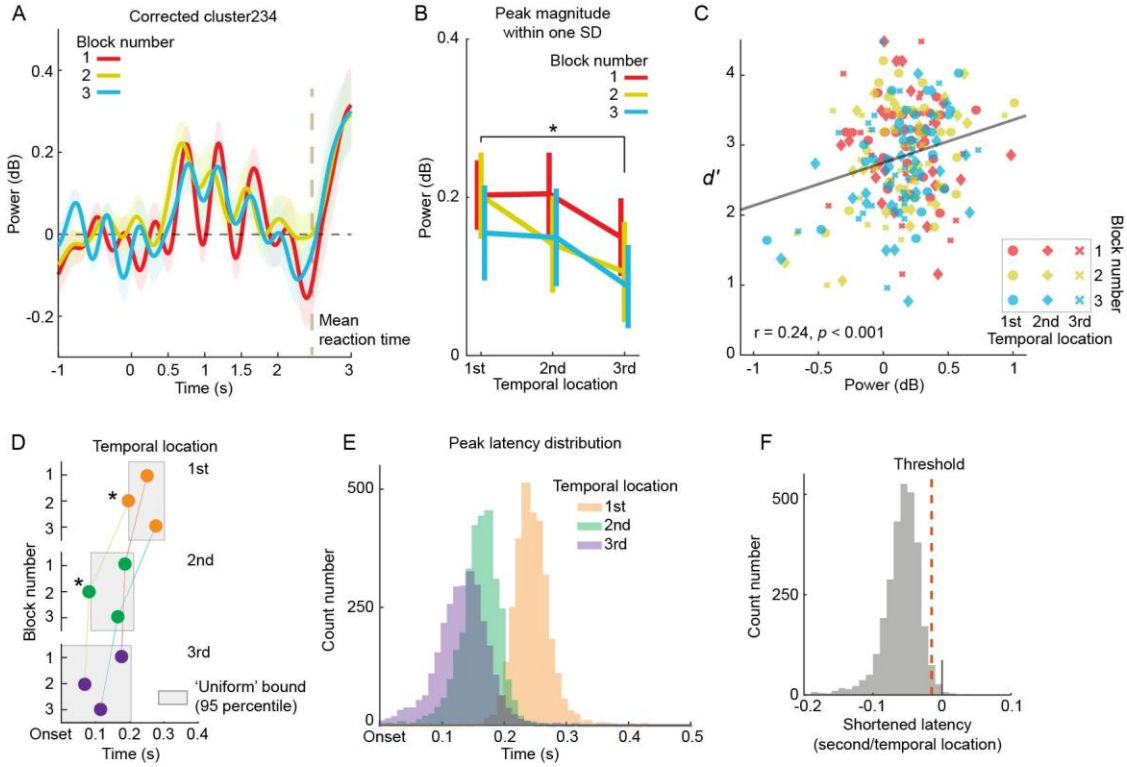

Fig. 3. Correlation of power modulation with tone detection performance, and neural power precession. (*A*) Power dynamics of Corrected cluster 234 in each block. The colored vertical line denotes the group mean of reaction time in each block. (*B*) Peak magnitudes within one SD of the Gaussian fits. The main effect of Temporal location is significant ($p < 0.05$) but the main effect of Block number is not ($p > 0.05$). (*C*) Correlation between d-prime values and neural peak magnitudes. The peak magnitudes of Corrected cluster 234, even though derived from the no-tone trials, positively correlate with tone detection performance. Each dot represents each participant's data point. (*D*) Peak latency differences of Corrected cluster 234 between blocks. The filled circles denote the peak latencies derived from fitting Gaussian curves to Corrected cluster 234; the color scheme of the circles codes for different temporal locations. The gray square areas show the 95% confidence interval of peak latencies derived from permutation tests across blocks, representing a 'Uniform' bound, a null hypothesis that all peak latencies at a temporal location do not differ significantly across blocks. If a peak latency falls out of the 'Uniform' bound, it significantly differs from other peak latencies. (*E*) Peak latency distribution. Three distributions for three temporal locations were derived from the permutation tests. (*F*) Distribution of shortened latency per temporal location. We randomly selected a set of peak latencies from the distributions in (*E*) and fitted a line to the selected peak latencies with the temporal location (1, 2, and 3) as the independent variable. This generated a distribution of slopes representing shortened latency per temporal location. 97.5% of the slope values fall below 0, suggesting that the peak latency significantly decreased with the temporal location. The shaded areas of color in (*A*) and the error bars in (*B*) represent ±1 standard error of the mean over participants. * represents $p < 0.05$.

We show in Fig. 3A Corrected cluster 234 in each block, from which it can be clearly seen that both the peak magnitudes and the peak latencies of Corrected cluster 234 varied with the blocks and the temporal locations. Following the above procedure of fitting Gaussian curves to the power peaks (Fig. 2G), we derived the peak magnitudes in each block (Fig. 3B&S5A) and conducted a two-way rmANOVA with Temporal location and Block number as the main factors. The main effect of Temporal location is significant ($F(2,52) = 5.01$, $p = .010$, $\eta_p^2 = 0.162$), but not the main effect of Block number ($F(2,52) = .51$, $p = .604$, $\eta_p^2 = 0.019$) nor the interaction ($F(4,104) = .25$, $p = .909$, $\eta_p^2 = 0.010$). The peak magnitude decreased with the temporal location, demonstrated by a significant linear trend ($F(1,26) = 9.22$, $p = .005$, $\eta_p^2 = 0.262$). The peak magnitude around the first temporal location is significantly larger than the third location ($t(25) = 3.04$, $p = .016$; adjusted FDR correction was applied). The data here show that the block number did not interact with the neural coding of the local temporal locations, so a learning effect was not observed across blocks.

Contrary to our prediction that learning the timing of prospective moments takes time, the participants learned the temporal locations of the target tone in the noise pieces efficiently and could anticipate the temporal locations even in the first block. One reason is probably that we presented 10 to 15 trials before the formal test to familiarize the participants with the task; the participants might learn the temporal locations during this training session. The visual feedback on the temporal location of the target tone in the tone trials probably also contributed to such an efficient learning process, as the feedback provided clear information on where each temporal location was. Future investigations can investigate this learning process further by removing the visual feedback. Nevertheless, this efficient learning process aided in identifying the neural signatures of the sequential temporal anticipation, which is the main focus here.

One remaining question is: does the neural power modulation marking the local temporal locations in the no-tone trials bear relevance to the participants' tone detection performance in the tone trials. We correlated the peak magnitudes of Corrected cluster 234 with the d-prime values (Fig. S1C&3C) and found a significant positive correlation ($r = 0.24$, $p < 0.001$), which indeed demonstrates that anticipating the local temporal locations was beneficial for the tone detection. We further show in Fig. S5B the correlation between d-prime values and the peak magnitudes in each block as well as in Fig. S5C the correlation at each temporal location. One potential explanation for the emergence of Corrected cluster 234 is that each neural peak after each temporal location is caused by false alarms – the participants may mistakenly think that a tone appears in the white noise, so a neural peak is

induced after a temporal location. If this is the case, the false alarm rate should positively correlate with the peak magnitudes of Corrected cluster 234 in the no-tone trials. However, we found that, mostly, the false alarm rates negatively correlate with the peak magnitudes (Fig. S5D). Hence, Corrected cluster 234 was not caused by the false alarms in the no-tone trials.

Precession of power dynamics suggests active anticipation

We next quantified the observation from Fig. 3A that the peak latencies of Corrected cluster 234 varied with the block number and the temporal location. Here, we conducted statistical tests on the peak latencies on the group level using permutation tests (see *Methods*), as we found that the peak latencies were not reliably derived using the above procedure of the Gaussian fits one each participant's data. The permutation tests included two steps. First, we tested whether the peak latencies changed across blocks. From the permutation tests, we derived a confidence interval (95%) for each temporal location across the three blocks (see *Methods* for details) and found that the peak latencies at the first and the second temporal locations of the second block were significantly shortened (Fig. 3D). Second, we tested whether the peak latencies changed across the temporal locations. We pooled the peak latency samples generated from the permutation tests across blocks to form a peak latency distribution for each temporal location (Fig. 3E) and fitted a line to three peak latencies randomly selected from the three distributions. The distribution of the slopes of the fitted lines shows that the peak latency significantly decreased with the temporal location (Fig. 3F). This finding is probably due to the fact that the temporal locations were anticipated sequentially - the anticipation of the first temporal location facilitated the anticipation of the second and so forth. We further examined this observation in the following RNN perturbation analyses.

Task-optimized RNNs replicate the behavioral and neural findings and suggest a process of dynamic gain control

Up to this point, we have identified a novel neural signature, Corrected cluster 234, marking the local temporal locations in the noise pieces (Fig. 2&3) and found that this neural signature cannot be accounted for by the expectation violation (Fig. S1B), the trial structure (Fig. S3), or the false alarm rate (Fig. S5D). Then, what functional roles did it play in the current task? To answer this question, we trained continuous-time RNNs (CRNNs) to accomplish the same behavioral task. The rational is that the RNNs trained to do the same

task that human participants did may also implement certain computations similarly carried out in the brain (Yamins and DiCarlo 2016; Yamins et al. 2014). If the trained CRNNs can indeed show a behavioral performance comparable to the human participants' performance, we examine internal dynamics of the trained RNNs and investigate how the RNNs accomplish the behavioral task. Moreover, since RNNs are flexible to train, various experimental conditions relevant to the current study can be tested, which helps generate new hypotheses for future investigations.

We designed a CRNN structure following the Dale's rule (Dale 1935; Song, Yang, and Wang 2016) and constructed inputs to CRNNs that captured the essence of the behavioral task (Fig. 1A&4A). The CRNN parameters and the training procedures are detailed in *Methods*. We systemically varied the external noise level of the inputs while keeping the target tone magnitude constant, because, if the timing of the temporal locations was anticipated to aid in detecting the weak target tone, the sequential temporal anticipation should be deployed to a larger extent at a high external noise level (low SNR) but to a less extent at lower external noise levels (high SNR). Then, the hidden activities of the CRNNs trained at the high noise level should show dynamics with 2 Hz modulation components marking the temporal locations, but none or less at the lower noise levels. If this is the case, it suggests that the 2 Hz neural component shown in Fig. 2&3 is related to the brain deploying the sequential temporal anticipation to cope with external sensory noise.

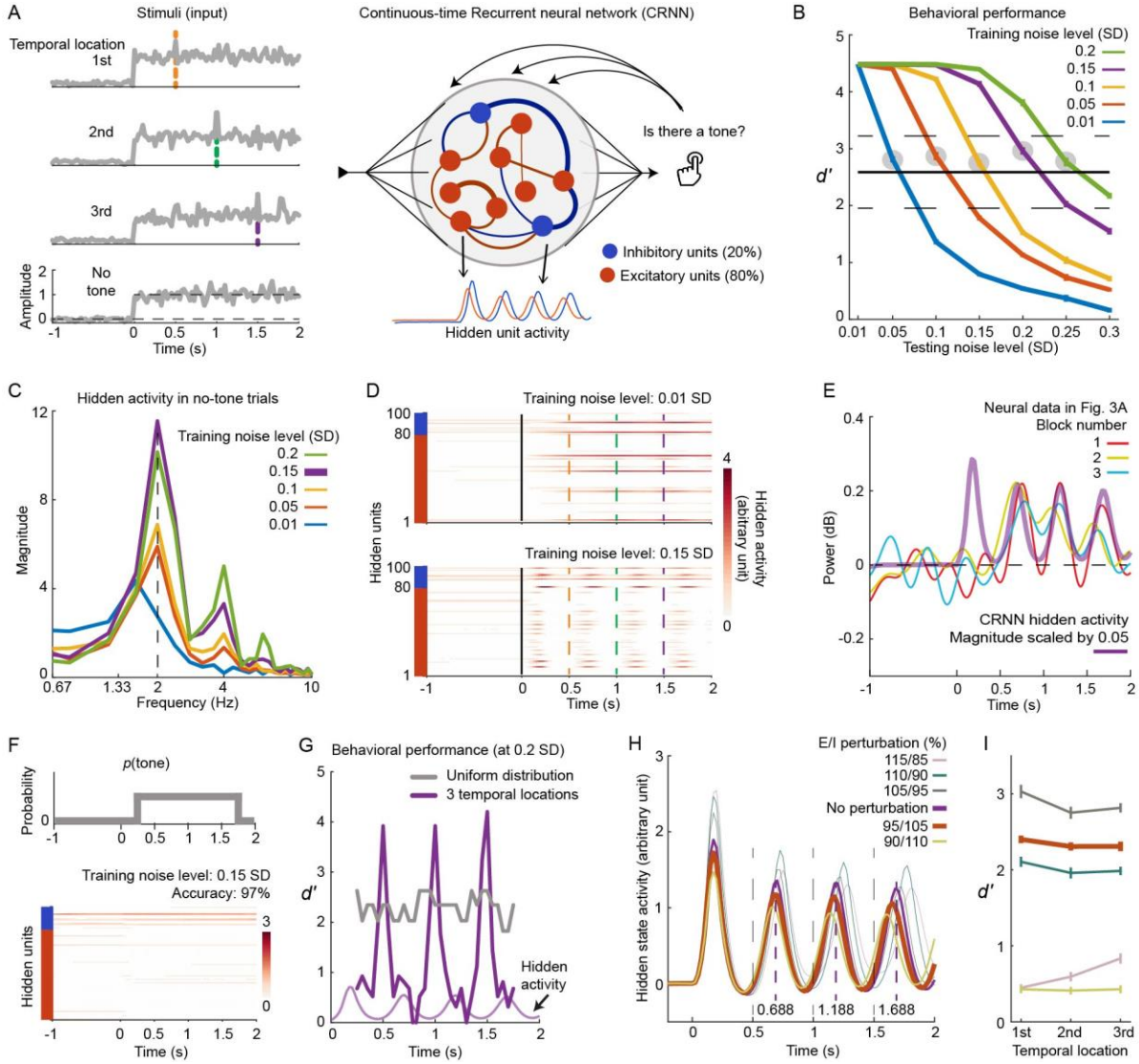

Fig. 4. Task-optimized continuous-time recurrent neural networks (CRNNs) account for neural findings. (*A*) Procedures of training CRNNs. The left plot illustrates the inputs to CRNNs and the right plot the CRNN structure and the training procedures. See *Methods* for details. (*B*) Task performance of trained CRNNs. We trained five CRNNs at five external noise levels, respectively, and tested each of them at seven noise levels. The solid black line denotes the group mean of human behavioral performance (Fig. S1C) and the dashed black lines represent ± 2 standard errors over participants. The filled gray circles indicate the testing noise levels chosen for the corresponding CRNNs whose performance matches the human performance. (*C*) Modulation spectra of hidden activities of the CRNNs at the corresponding testing noise level chosen in (*B*). We derived modulation spectra of CRNN hidden activities following the EEG analysis procedures. (*D*) Hidden activities of two CRNNs derived from 40 no-tone trials. The y axis indicates whether the units are excitatory (red units) or inhibitory (blue units). (*E*) Comparison between CRNN hidden activities and neural data. The neural data come from the group-mean data shown in Fig. 3A. The CRNN was the one trained at the noise level of 0.15 SD and tested at the noise level of 0.2 SD. (*F*) Train CRNN with a uniform probabilistic distribution of tone locations. The upper plot illustrates that, in the trial structure of a uniform distribution of tone locations, a tone can appear any time from 0.25 s to 1.75 s after noise onset with an equal probability. The lower plot shows that the hidden activities of the CRNN trained using the 'uniform' trial structure did not show any rhythmic activities. (*G*) Task

performance in the ‘uniform’ trial structure. At the testing noise level of 0.2 SD, we tested the CRNN trained in (*F*) and the CRNN trained with three temporal locations at the training noise level of 0.15 SD. (*H*) Hidden activities of CRNNs with perturbed excitation-inhibition (EI) weights. The thick colored lines refer to the EI perturbations that caused the peak latencies of CRNN hidden activities to decrease with the temporal location. The violet line represents the CRNN without perturbation selected from (C); the dashed vertical lines represent peak latencies of this unperturbed CRNN. (*I*) Task performance of perturbed CRNNs at the testing nose level of 0.2 SD.

We trained five CRNNs at five noise levels, respectively: 0.01, 0.05, 0.1, 0.15, 0.2 SD of added Gaussian noise, and calculated d-prime values of each trained CRNN at various testing noise levels (Fig. 4B). For each of the five CRNNs, we selected a specific testing noise level at which the corresponding CRNN generated a task performance matching the group mean of the human behavior performance (Fig. 1C&4B). Then, we analyzed the activities of the hidden units of the selected CRNNs and calculated the modulation spectra of their hidden activities when 40 no-tone trials were presented, by following the same procedure of quantifying the power modulations of neural signals (Fig. 4C). Indeed, the modulation strength around 2 Hz increased with the external noise level used to train CRNNs. The CRNN trained at the noise level of 0.15 SD showed the highest modulation strength at 2 Hz whereas the CRNN trained at 0.01 SD showed the least modulation strength (Fig. 4C), which is also evident in temporal dynamics of their hidden activities (Fig. 4D). We averaged the hidden activities of the CRNN trained at 0.15 SD across all hidden units, analogous to EEG recording in which neural signals coming from both excitatory and inhibitory synapses are pooled together, and plotted the averaged hidden activities of the CRNN together with Corrected cluster 234 (Fig. 4E). It can be seen that the dynamics of the CRNN hidden activity closely captured the power dynamics of Corrected cluster 234.

One possibility is that the 2 Hz modulation components in both the brain and the CRNNs carried out an operation of noise suppression – the noise period outside of the tone locations was suppressed to aid in detecting the target tone. If this is the case, a target tone may be suppressed together with noise if it appears at a time point other than the three temporal locations used to train CRNNs. Moreover, is it possible that a CRNN show such 2 Hz hidden dynamics simply because the noise level is high but not because of the specific temporal locations used? We answered those two questions by training a CRNN with a new trial structure in which the training noise level was fixed at 0.15 SD and the target tone could appear at any time points between 0.25 s and 1.75 s after the noise onset (Fig. 4F, upper panel) – a ‘uniform’ trial structure. The newly trained CRNN showed flatten hidden activities without any rhythmic patterns (Fig. 4F, lower panel). Subsequently, in the uniform trial structure at the noise level of 0.2 SD, we tested the newly train CRNN and the CRNN trained with the three fixed temporal locations (both trained at the noise level of 0.15 SD). The tone detection performance of the newly trained CRNN dropped largely but remained at a moderate range; in contrast, the CRNN trained with the three fixed temporal locations showed higher tone detection performance at the temporal locations but much worse

performance at other time points (Fig. 4G). Note that the calculation of d-prime values here was different from Fig. 4B as the trial structures used were different (see *Methods*).

The results in Fig. 4F&G suggest that the 2 Hz neural modulation shown in Corrected cluster 234 played a role in increasing sensory sensitivity to the target tone at the anticipated temporal locations and selectively suppressing sensory noise at irrelevant noise periods. Coincidently, the power frequencies of Corrected cluster 234 fell in the range of the routinely defined alpha and beta bands (Fig. 2B); it has been demonstrated with mounting evidence that the neural activities in the alpha-beta bands closely relate to suppressing internal/external noise and regulating information flow (Klimesch 2012; Spitzer and Haegens 2017; Haegens, Luther, and Jensen 2012; Wöstmann, Lim, and Obleser 2017; Arnal and Giraud 2012; Wöstmann, Alavash, and Obleser 2019; Zumer et al. 2014; Jensen and Mazaheri 2010). Together, our neural findings and the CRNN analyses suggest that the sequential temporal anticipation was partly carried out as a gain control process in time (Salinas and Sejnowski 2001; Nobre and van Ede 2018; Van Ede et al. 2018; Strauß, Wöstmann, and Obleser 2014). The brain selectively gates information at the future time points in a similar way that, in space, the brain gates in the sensory information at attended locations and ignores background noise (Hillyard, Vogel, and Luck 1998; Reynolds and Chelazzi 2004).

Decreased E/I ratio or increased inhibitory activity explains shortened neural peak latency

It has been known that perturbing excitation-inhibition balance changes the speed of neural oscillatory activities (Brunel and Wang 2003). In the current study, if the temporal locations were fixed in time, it is possible that the shortened neural peak latencies of Corrected cluster 234 (Fig. 3E&F) were facilitated by increased inhibitory (or excitatory) neural activities during the anticipating process. Next, we perturbed the excitation-inhibition balance of the CRNNs and examined how the change of excitation-inhibition balance modulated the peak latencies of CRNN hidden activities.

We changed connectivity strength of excitatory and inhibitory units in the CRNN trained at the noise level of 0.15 SD in a manner that the ratio between excitatory and inhibitory projections (E/I ratio) was systemically manipulated (Fig. 4H). For example, we decreased 5 percent of excitatory unit weights and correspondingly increased 5 percent of inhibitory unit weights in a CRNN, so that excitatory units sent weaker projections to inhibitory units and inhibitory units sent stronger projections to excitatory units. We derived averaged hidden activities of the CRNNs of different E/I ratios (Fig. 4H). When the E/I ratio decreased to less than one (e.g., 95/105) – stronger inhibitory projections but weaker

excitatory ones, the peak latencies decreased with the temporal location, which echoed the neural finding shown in Fig. 3E&F. The peak latencies of the original CRNN (E/I ratio: 100/100) did not decrease with the temporal location and were fixed at 0.188 s (the middle time point between two largest points of each peak, indicated by the vertical dashed lines in Fig. 4H). We then tested the behavioral performance of the CRNNs with perturbed E/I ratios and found that, when the E/I ratio of the CRNN is 95/105, the behavioral performance still fell in the range of the human behavioral performance (Fig. 4I). Overall, this exploration of E/I ratios in CRNNs suggests that the neural power precession observed in Corrected cluster 234 may be related to increased neural activities of inhibitory circuits. The anticipations of the previous temporal locations facilitated the anticipations of the forthcoming ones, which correlates with the increased efficiency of the inhibitory circuits. This is consistent, again, with previous neurophysiological findings and hypotheses that neural signals in the alpha-beta bands correlate with inhibitory activities (Sherman et al. 2016; Lakatos et al. 2016; Klimesch, Sauseng, and Hanslmayr 2007).

Discussion

To study how the brain anticipates prospective moments sequentially, we designed a 'premembering' task by adapting a simple tone-detection paradigm, employed novel modulation spectrum analyses on the electrophysiological data, and trained and analyzed 12 RNN models. We found that the ramping neural activities in the delta band was exploited to encode the global temporal context on a supra-second scale and a neural modulation component on power in the alpha-beta band range was deployed to anticipate locally the prospective temporal locations on a sub-second scale (Fig. 2). The neural modulation signatures for the sequential temporal anticipation positively correlated with the tone detection performance (Fig. 3) and cannot be accounted for by expectation violation, trial structure, or false alarms. By using the task-optimized RNNs (Fig. 4), we further explored the functional role of our neural findings and showed that the neural power modulations underlay a gain control process in time that suppresses irrelevant noise periods and increases sensory sensitivity at the anticipated temporal locations. Our behavioral, neural, and modelling findings collectively demonstrate that the temporal anticipation, a psychological term to describe a general task demand, was implemented in the current context as a process of dynamic gain control in a multi-timescale manner.

One central theme in psychology and neuroscience is how the brain copes with noise and extracts signals (Faisal, Selen, and Wolpert 2008). The brain is constantly bombarded with sensory inputs, but relevant and useful information is sparse in the real world. In the current context, meaningful events (i.e. the target tone) appear potentially at only one of a few time points whereas sensory noise is constantly present. Then, the functional requirement for the temporal anticipation is not only to anticipate the prospective moments but also to ignore the irrelevant noise periods. As demonstrated in our neural and modelling findings, the neural operations carried out in the alpha and beta bands increased sensory sensitivity to the anticipated temporal locations and suppressed the noise outside of the temporal locations. This finding is well aligned with the function of attention in space – in spatial attention studies, the alpha and beta neural bands have been found to correlate with suppression of irrelevant objects or backgrounds (Bauer et al. 2014; Worden et al. 2000; Bauer et al. 2012). Our findings embrace a reinterpretation on the temporal anticipation that has been treated as a unitary construct in various contexts (Nobre and van Ede 2018) – exact underlying mechanisms of the temporal anticipation vary with functional demands.

The sequential temporal anticipation is characterized by two neural correlates, one on the timescale of the global sequence length and one on the timescale of the local temporal

locations. One following deeper question is: how do the timing processes on those two timescales interact? In the spatial domain, it has been suggested that visual patterns are first recognized on a global scale and local details are read out later (Navon 1977; Chen 1982). In our temporal task, the duration of the noise pieces was signaled by the salient noise onset and offset, so it is relatively easy to register mentally the noise length. The brain may first temporally recognize this global temporal frame and then refer to this global frame to locate in time the local events. This may explain why, in our anticipation task, the neural component marking the temporal locations did not show neural peaks before the anticipated temporal locations to indicate upcoming events (Fig. 2). This global-local anticipating scheme suffices to separate the tone locations from noise periods for detecting the target tone but may not be applicable when immediate reactions or decisions are needed after each temporal location, such as in set/go tasks (Correa and Nobre 2008; Tavano, Schröger, and Kotz 2019). Again, detailed task demands matter in timing tasks and potentially mediate anticipating processes.

An alternative interpretation, probably a more intriguing one, for Corrected cluster 234 marking the temporal locations of the target tone (Fig. 2&3) is that this neural signature may represent the probability distribution of event occurrence in time (Grabenhorst et al. 2019, 2021; Janssen and Shadlen 2005) (Fig. 1B). It is worth further investigations to test whether Corrected cluster 234 corresponds to the neural representations of probability distributions shown in animal studies by varying the distribution shapes (Janssen and Shadlen 2005). However, it is not always necessary to represent a probability distribution of event occurrence. If the SNR of tone to noise is low, the tone detection task becomes easy, and it is no longer beneficial to represent the probability distribution of the tone occurrence in time. This is exactly what the RNN modelling results showed: although the probability distribution of the tone occurrence was unchanged, the RNN dynamics marking the tone locations became weaker as the SNR decreased (Fig. 4C). Further investigations can indeed test this hypothesis generated by our RNN models.

Regarding with the brain areas that implement the sequential temporal anticipation, it has been indicated that, on a supra-second scale, anticipating the timing of incoming events and interval estimation are carried out by the parietal and motor areas (Morillon and Baillet 2017; Janssen and Shadlen 2005; Coull, Cotti, and Vidal 2016; Bolger, Coull, and Schön 2014; Merchant and Yarrow 2016; Wittmann 2013). For anticipating the local temporal locations, as only the EEG channels around Cz showed considerable modulation strength (Fig. 2H), we suspect that the corresponding neural modulation component originated from subcortical areas. One candidate locus is the basal ganglia, particularly the striatum that has

been demonstrated to be related to motor sequence generation as well as to several sub-second scale timing tasks (Motanis and Buonomano 2015; Kotz and Schmidt-Kassow 2015; Nozaradan et al. 2017; McClure, Berns, and Montague 2003; Mello, Soares, and Paton 2015). Since to generate sequential actions requires coordinating actions in a sequence of moments (Rimmele et al. 2018; Schubotz 2007), this functional requirement indeed aligns with the cognitive process required in our task – anticipate a sequence of prospective moments.

To conclude, we identified a multi-scale neural processing scheme in the sequential temporal anticipation: the brain exploits neural power modulations on several neural bands to encode duration of global temporal contexts and to anticipate sequentially local temporal structures. The task-optimized artificial neural network models (i.e. the CRNNs) helped reveal a dynamic gain control process underlying the temporal anticipation – to anticipate a few meaningful moments in the future is also to actively ignore irrelevant events that happen most of the time.

Methods

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Xiangbin Teng (xiangbin.teng@gmail.com).

Participants.

Thirty participants took part in the experiment (age 18 – 35, right-handed, sixteen females). All participants had normal hearing and no neurological deficits according to their self-report. Written informed consent was obtained from each participant before the experiment; monetary compensation was provided accordingly after the experiment was finished or terminated. Two participants could not undergo the EEG recording because of the metal fillings or decorations above their neck and chose to terminate the experiment. One participant's data were excluded because triggers for stimulus onsets were not recorded during EEG recording. One participant finished EEG recording in the first session but did not undergo EEG recording during the second session because of battery problems of EEG device. Therefore, the neurophysiological data shown in the main text included twenty-seven participants (age 18 – 35, right-handed, fourteen females); the neurophysiological data from the second session shown in the supplementary information (Fig. S3) included twenty-six participants (age 18 – 35, right-handed, fourteen females). The experimental protocol was approved by the Ethics Council of the Max Planck Society.

Experimental Paradigm and Stimuli

The present study included two sessions. In both sessions, the participants were instructed to detect a 1000 Hz pure tone of 30 ms embedded in a 2-s piece of white noise at one of three temporal locations: 500 ms, 1000 ms, and 1500 ms. The difference between the two sessions was that, in the first session, there was no cue to direct participants' attention to one of the three temporal locations, i.e., *no-cue* session, which was illustrated in Fig. 1A, whereas in the second session there was a cue in every trial that directed participants' attention/prediction to a specific temporal location, i.e., *cued* session (Fig. S3A).

We hypothesize that the participants, not knowing when the target tone would appear in the no-cue session, had to anticipate the three temporal locations of the target tone sequentially, whereas in the cued session the participants knew the target tone location and only had to pay attention to the cued temporal location. Therefore, different cognitive processes could be involved to accomplish the same tone-detection task. Our current study

focused on the no-cue session to study how the participants memorized the temporal locations and anticipated the temporal locations sequentially; the cued session served as a control session or a comparison to the no-cue session to show that the effects found in the no-cue session were mediated by specific cognitive demands but not simply driven by the trial structure – detect a tone in a 2-s noise piece.

The no-cue session was comprised of three blocks and the experimental procedures are illustrated in Fig. 1A. In each trial of each block, the participants were required to keep eyes open and to fix on a white cross in the center of a black screen. The white cross stayed on the screen for a duration randomly chosen between 1 s and 1.5 s. After the white cross disappeared, the 2-s noise piece was presented. The participants pressed one of two buttons to indicated whether they heard a tone after the noise piece was over. A visual feedback after participants' responses was shown on the screen for 0.3 s to inform the participants whether their response was 'correct' or 'wrong'. Then, another feedback was shown on the screen for 0.4 s to indicate whether and when the target tone appeared in the noise piece with one of four symbols: '_ _ _ _' (no tone was presented), '_ | _ _ _' (a tone appeared at 500 ms after the noise onset), '_ _ | _ _' (1000 ms), '_ _ _ | _' (1500 ms). After the two visual feedbacks, the next trial started immediately. In each block, there were four conditions: a tone appeared at either 500 ms, 1000 ms, or 1500 ms, or no tone was presented. There were 40 trials for each condition in each block, so there were 160 trials in each of three blocks. 480 trials in total were presented in the no-cue session. The conditions were randomly chosen in each trial and the trial order was randomized across blocks between participants. After each block, the participants could choose to take a break of one or two minutes or to start the next block immediately.

After the no-cue session and a break of around three minutes, we conducted the cued session. The trial structure was the same as in the no-cue session, but the visual feedback indicating the temporal location of the target tone was presented at the beginning of each trial as a visual cue (Fig. S3A). In each trial, one of the three symbols ('_ | _ _ _', '_ _ | _ _', or '_ _ _ | _') was first presented in the center of the screen for 0.4 s. After the visual cue disappeared, the trial structure was the same as in the no-cue session. A tone appeared at the cued temporal location in half of trials but did not in the other half, so the visual cue could be valid or not. Three cue conditions combined with two conditions of cue validness constructed six conditions in total. The cued session contained two blocks and there were 25 trials in each condition in each block, so there were 50 trials for each condition and 300 trials in total in the

cued session. The conditions were randomly chosen in each trial and the trial order was randomized across blocks between participants.

The white noise piece in each trial of both the no-cue and cued sessions was independently generated using a random number generation function, 'rand', in MATLAB R2016b (The MathWorks, Natick, MA) at a sampling rate of 20,000 Hz. As a new white noise piece was generated in each trial, the participants could not memorize the acoustic details of white noise to help them detect the target tone. The amplitude of white noise pieces was normalized to 70 dB SPL (sound pressure level) by referring the noise pieces to a one-minute white noise piece that was measured beforehand to be 70 dB SPL at the experimental setting. We applied a cosine ramp-up function in a window of 50 ms at the onset of the noise pieces and a sine ramp-down function of 50 ms at the offset.

The 1000 Hz target tone of 30 ms duration was generated using the function of 'sin' in MATLAB with a fixed initial phase of 0 degree. We applied a cosine ramp-up function in a window of 10 ms at the onset of the tone and a sine ramp-down function of 10 ms at the offset. As the threshold for detecting a pure tone in white noise varies across individuals, we set an individualized signal-to-noise ratio (SNR) of the target tone to the white noise for each participant, so that the difficulty level of the task could be comparable between participants and be controlled. We measured each participant' tone detection threshold before the no-cue session (Fig. S1A) and set the SNR of the target tone to be 3 dB above each participant's threshold in both the no-cue and cued sessions.

To measure the tone detection threshold, we employed a two-interval two-alternative forced choice paradigm and the PSI method (Kontsevich and Tyler 1999; Prins 2013) that were implemented in the Palamedes toolbox 1.5.1 (Kingdom and Prins 2016). In each trial, participants were presented with two pieces of white noise of 1 s with an inter-piece interval randomly sampled between 0.5 s and 1 s. The target tone was randomly inserted into one of the two noise pieces from 0.2 to 0.8 s after the noise onset. The white noise and the target tone were generated following the same procedures described above. The participants chose which noise piece included the target tone by pressing one of two buttons after the second noise piece was over. The visual feedback was provided to indicate whether the participants made the correct response in each trial, so that the participants could learn this tone-detection task and show stable performance in the following no-cue and cued sessions. To get robust threshold measurements, we measured each participant's threshold twice. In the first measurement, the starting point of the SNR was set to -5 dB and the threshold was measured in 40 trials. In the second measurement, the starting point of the SNR was the threshold

measured from the first measurement. Every 15 trials, there was a catch trial with a SNR of 5 dB to remind the participants of the target tone. The threshold measured from the second time was selected as each participant's threshold. The threshold here represented 75 percentage correct threshold.

The auditory stimuli were delivered through plastic air tubes connected to foam ear pieces (E-A-R Tone Gold 3A Insert earphones, Aearo Technologies Auditory Systems).

EEG recording

EEG data were recorded using an actiCAP 64-channel, active electrode set (10–20 system, Brain Vision Recorder, Brain Products, brainproducts.com) with a 0.1 Hz online filter (12 dB/octave roll-off). There were 62 scalp electrodes and their impedances were kept below 5 kOhm. The first 18 participants of 27 valid participants were recorded at a sampling rate of 500 Hz with one electrode (originally, Oz) placed on the tip of the nose as the reference channel. The other 9 participants were recorded at a sampling rate of 10000 Hz with one electrode placed at Oz as the reference channel. The impedance of the reference channel was kept below ~10 kΩ. This difference of EEG recording settings was caused by miscommunication between the experimenter and the research assistants who helped conduct EEG recording. To make the EEG data compatible between the two recording settings, in the following data analyses, EEG recordings were off-line referenced to the average of activity at all scalp electrodes. The recording sampling rate difference should not matter here as we later applied an off-line low-pass filter of a cut-off frequency of 30 Hz and downsampled all the data to 100 Hz. Furthermore, the following major analyses included baseline corrections based on pre-stimulus activities in individual participants' data and focused on modulations of EEG power (a second-order statistics), the EEG reference differences between participants should not affect our results.

Behavioral data analysis.

Tone detection performance was analyzed in MATLAB 2016b using the Palamedes toolbox 1.5.1 under the framework of signal detection theory. In each block of the no-cue session, we treated the trials with a tone embedded as the 'target' trials and the trials without a tone as the 'noise' trials. For each temporal location in each block, we derived a hit rate from the 'target' trials ('yes' when there was a tone) and a false alarm rate from the 'noise' trials ('yes' where there wasn't a tone). We then calculated a d-prime value using the hit rate and the false alarm rate assuming an independent observer model and an unbiased observer. To avoid infinite d-

prime values, a half artificial incorrect trial was added when all ‘target’ trials were correct; a half artificial correct trial was added when all ‘noise’ trials were incorrect (Macmillan and Douglas Creelman 2004). Note that the ‘target’ trials with the target tone appeared at one of the three temporal locations shared the same ‘noise’ trials, so the d-prime values and the criteria derived at each of the three temporal locations were not independent and the differences between the temporal locations were dominated by the hit rates.

In the cued session, for each cued temporal location, we calculated a d-prime value using the hit rate from the valid trials and the false alarm rate from the invalid trials.

Raw reaction time data were used in the data analysis. The reason that we did not take a log transformation of reaction time is that previous literature demonstrated that a log transformation of reaction time is not necessarily beneficial (Harald Baayen and Milin 2010). Nevertheless, we conducted statistical tests on both log-transformed and raw reaction time data and found that the conclusions of the statistical tests remained the same.

EEG preprocessing

EEG data analysis was conducted in MATLAB 2016b using the Fieldtrip toolbox 20181024 (Oostenveld et al. 2011) and the wavelet toolbox in MATLAB. EEG recordings were off-line referenced to the average of activity at all scalp electrodes. Raw EEG data were first filtered through a band-pass filter from 1 Hz to 30 Hz embedded in the Fieldtrip toolbox (a ‘fir’ two-pass filter with an order of 4). The cutoff frequencies of 1 Hz and 30 Hz were selected here because previous literature showed that effects of temporal anticipation in the auditory modality were primarily observed in the delta band (1 – 4 Hz) and the beta band (12 – 30 Hz) (Arnal, Doelling, and Poeppel 2014; Morillon and Baillet 2017). Nonetheless, although the upper cutoff frequency of the bandpass filter was 30 Hz, we could still check the experimental effects on EEG signals above 30 Hz. The reason for this is that the ‘fir’ filter has a shallow stopping band and the baseline correction of EEG power analyses cancelled out the negative filter gain (see *Time-frequency decomposition of EEG signals*).

Continuous EEG data were first divided into 6-second epochs including a 2-s pre-stimulus period, a 2-s stimulus period, and a 2-s post-stimulus period for each trial. By preserving long periods before and after the stimulus onset and offset, we had enough data to calculate low-frequency EEG power (e.g. at 1 Hz) to conduct the following baseline corrections and to remove filter artefacts in the beginning and the end of the EEG signals of each trial. The epoched trials were then visually inspected, and those with artifacts such as channel jumps and large fluctuations were discarded. An independent component analysis

was applied over all the trials pooling over all the blocks in each session and was used to correct for artifacts caused by eye blinks and eye movements. After preprocessing, up to 10 trials were removed for each of the four conditions in the no-cue session and up to 6 trials were removed for each of the six conditions in the cued session. Baseline in each trial was corrected by subtracting out the mean of -1 s to 0 s.

We first conducted various following analyses on the Cz channel. After we found certain effects at Cz, we conducted the same analyses at other channels.

Event-related potential (ERP) to tones

We first calculated evoked responses to the target tone at each temporal location. The baseline of each trial was corrected as described in *EEG preprocessing.* We chose not to conduct baseline correction again for ERPs to the target tone at different temporal locations, because the target tone at 500 ms was affected by the noise onset response whereas the tones at 1000 ms and 1500 ms were not. Re-correcting baselines for ERP to the target tone at each temporal location may introduce artefacts. We focused on the period of ERPs from 0 ms to 500 ms after the tone onset in the following statistical analyses.

In the no-cue session, we first pooled trials across all the three blocks and calculated ERPs. This pooling procedure guaranteed that we had enough trials to detect differences of ERPs between three temporal locations with sufficient statistic power. After we identified critical temporal regions showing effects of the temporal location, we then tested how the learning effect across blocks affected the neural responses within those temporal regions.

We conducted a permutation-based cluster repeated measure ANOVA (rmANOVA) analysis over the three temporal locations (Maris and Oostenveld 2007), in which we shuffled the labels on the temporal locations in each participant's data. By doing this, we constructed a null hypothesis, which was that the ERPs to the target tone at three temporal locations do not differ from each other. During each time of shuffling, we constructed a new dataset and conducted a one-way rmANOVA test at each time point with Temporal location as the main factor. We identified clusters of connected significant time points (> 2 time points) at a two-sided alpha level of 0.05 and calculated a cluster-level F value by adding up F values within each cluster across time points. The biggest cluster F value was selected as the cluster-level F value for each time of shuffling. We conducted this procedure 1000 times and derived a distribution of cluster-level F values, from which a threshold of cluster-level F values at a one-sided alpha level of 0.05 was derived (the 950th cluster F value). We then derived significant clusters from our empirical data and determined which cluster was significant by

comparing their cluster-level F values with the threshold of cluster-level F values. Then, we conducted paired t-tests within the temporal regions of the significant clusters to compare ERPs between each pair of temporal locations. Adjusted FDR was used for multiple comparison correction.

We next investigated the effect of block order in the no-cue session within the significant temporal regions identified above. We calculated ERPs to the target tone in each block and averaged ERP values within each significant temporal region. A two-way rmANOVA was conducted with Block number and Temporal location as the main factors.

In the cued session, the same analyses on ERPs were conducted, except that we did not investigate the effect of block order.

Time-frequency decomposition of EEG signals

To extract time-frequency information, in both the no-cue session and the cued session, single-trial data in each EEG channel were transformed using functions of the Morlet wavelets embedded in the Fieldtrip toolbox, with a frequency ranging from 1 to 40 Hz in steps of 1 Hz. To balance spectral and temporal resolution of the time-frequency transformation, from 1 to 40 Hz, the window length increased linearly from 2 circles to 20 circles. Phase and power responses (squared absolute value) were extracted from the wavelet transform output at each time-frequency point. The temporal resolution was set as 100 Hz and the power responses from -1 s to 3 s were selected for the following analyses.

In each condition, we first averaged power values over all the trials and then calculated a baseline by averaging the mean power values from -.75 s to -.25 s. The averaged power after the noise onset (0 s) was normalized by dividing the baseline and taking logarithms with base 10, and then was converted into values with unit of decibel by multiplying 10. As the power was first calculated on the single-trial level and was then averaged across trials, the resulted power was routinely called induced power. The induced power is different from the evoked power in that the evoked power is calculated by first averaging trials in the temporal domain and then conducting the time-frequency decomposition.

Note that, as the low-pass filtering in the preprocessing procedure was conducted on the entire EEG recording, the baseline correction of EEG power here re-corrected the power gain of different EEG frequencies in each trial according to its pre-stimulus period. This cancelled out the low-pass filtering effect, which is why considerable power magnitude can still be seen in the frequency range of 30 Hz to 40 Hz.

Modulation power spectrum (MPS) of induced power in the no-cue session

We next quantified how the induced power was modulated by temporal anticipation for the temporal locations over the entire trials by calculating modulation power spectrum (MPS). The hypothesis we had in the no-cue session is twofold. First, if the participants could register the three temporal locations of the target tone in their memory and actively anticipate the tagged temporal locations every 500 ms, we would find a neural signature around 2 Hz reflecting such sequential anticipating processes even when only the white noise was presented in each trial. Secondly, if the duration of the noise on a supra-second scale is encoded in the brain to constrain the temporal locations of the sub-second scale, we would find a neural signature around or below 1 Hz that matches the timescale of the noise duration.

We focused on the condition in which only the white noise was presented and pooled all the no-tone trials over three blocks. We selected the neural data between 0.25 s and 1.75 after the noise onset to avoid effects of onset and offset responses. We did not define any EEG frequency bands but used a data-driven method to find significant modulation effects on the induced power from 1 Hz to 40 Hz. We first applied a high pass filter at 0.5 Hz on the induced power at each frequency over entire trials (-1 s to 3 s) so that we removed neural components over a duration larger than the length of the white noise pieces (2 s). We next applied a hamming window of 1.5 s (151 points) on the selected time range, 0.25 s to 1.75 s after noise onset, at each frequency of induced power and conducted the Fourier transform with a length of 500 points (151 data points padded with 349 zero points). The spectral resolution of the resulted spectrum of the induced power was 0.2 Hz. The MPS derived have two spectral dimensions: one dimension was the frequency of induced power (‘power frequency’) and the other dimension was the ‘modulation frequency’ of modulations of induced power. As the induced power was normalized to have a unit of dB, the unit of MPS was also dB since the Fourier transform only includes linear operations.

Here, we laid out two further constraints due to the parameter selections in the above analyses used to derive the MPS. First, the duration of induced power selected for the MPS analyses was 1.5 s long (0.25 s to 1.75 s after noise onset), which meant that the lowest possible modulation frequency representing more than two cycles of induced power fluctuations was 1.33 Hz. At the modulation frequencies less than 1.33 Hz, one could not determine whether the modulation components represented by the frequencies are rhythmic or belong to a duty cycle of an even lower frequency, or simply indicate a trending signal. Therefore, there is Trend bound at 1.33 Hz along the axis of modulation frequency - neural components above 1.33 Hz can be considered to have rhythmic modulation components;

neural components below 1.33 Hz likely represent trending responses from 0.25 s to 1.75 s. Secondly, the induced power at one power frequency can only fluctuate at a modulation frequency less than one half of the power frequency according to the Nyquist-Shannon sampling theorem. The induced power at one power frequency can be viewed as a signal carrier; the power modulation envelope of this carrier can only have a modulation frequency as high as half of the carrier frequency. For example, the modulation frequency derived from the induced power at the power frequency of 1 Hz can only be as high as 0.5 Hz in theory. If we examine the modulation frequency of 2 Hz on EEG power, we have to look at the EEG power at least above 4 Hz of the power frequency. Hence, there is Sampling bound. We indicated those two constraints, Trend bound and Sampling bound, in Fig. 2B.

Note that the MPS was not affected by different baseline correction procedures used in *Time-frequency decomposition of EEG signals*. The reason is that we took a log transformation of EEG power after the division between the power of the stimulus period and the baseline power. This is equal to taking a log of the power of the signals and a log of the baseline power separately and then subtracting the two values linearly. The baseline power did not change in time and, after the Fourier transform was applied, was separated out from the EEG power of the signals as a DC component in the modulation spectra (the component at 0 Hz).

Surrogate test on MPS

To determine the spectral regions of MPS showing significant modulation strength, we conducted a surrogate test in the spectral domain at the group level. The rationale is that, as all the participants were doing the same task, neural dynamics reflecting the underlying cognitive processes should be similar across individual participants. At the group level, we should be able to identify neural dynamics that are consistent across the participants. Hence, the null hypothesis here is that no consistent temporal modulation of induced power at any power frequency exists across the participants.

To construct the null hypothesis, after we applied the Fourier transform on the EEG power at each power frequency and obtained the Fourier coefficients (complex numbers), we multiplied a complex number of a random phase and of a norm of 1 with the Fourier coefficient at each modulation frequency. By doing this, we disrupted the phase of each modulation frequency at each power frequency but preserved the modulation strength in each participant's data – this is equal to disrupting temporal dynamics. We then averaged the surrogated data over all the participants in the spectral domain and derived a new group-level

MPS. We conducted this procedure 1000 times and constructed a null distribution of MPS at each power frequency and each modulation frequency, from which we derived a threshold of a one-sided alpha level of 0.05.

The reason that we chose the surrogate test in the spectral domain but not in the temporal domain is that we would like to specifically examine the temporal alignment of modulations on induced power across the participants while keeping the modulation strength unchanged. To conduct a surrogate test in the temporal domain, we would have to change the time range of individual participants' EEG data by shifting the time range of interest forward or backward in each surrogate. The shifted time range may include noise onset/offset responses or data from the fixation period, so both the modulation strength and phase of induced power in the shifted time range are changed and are different across participants.

We selected four significant clusters (> 2 spectral bins) that were robustly shown in each time of the surrogate test with high modulation power (Cluster 1, 2, 3, and 4 in Fig. 2B). Other significant modulation bins were also observed, but they either had low modulation strength in the MPS or fell below Sampling bound (likely caused by the ripple artefacts of the Fourier transform). The power frequency ranges/the modulation frequency ranges for the four clusters are as following: Cluster 1, 1-4 Hz/0-1.33 Hz; Cluster 2, 22-27 Hz/0-1.33 Hz; Cluster 3, 9-14 Hz/1.8-2.8 Hz; Cluster 4, 24-26 Hz/1.8-2.8 Hz.

The modulation frequency range for each cluster was not selected strictly according to the significant spectral bins in the MPS of Fig. 2B. The significant modulation frequencies across different power frequencies in one cluster were not the same, so it was implausible to filter the induced power strictly according to the significant frequency bins when the temporal domain of the induced power was investigated. Defining a 'square' region of interest in the MPS for each cluster facilitated filtering procedures in the following analyses of extracting temporal dynamics; simplistic filtering procedures avoided potential artefacts introduced by filtering each power frequency using a different modulation frequency range. Therefore, we chose the spectral bins for the clusters in the following manner: 1) the significant modulation frequencies of Cluster 1 and 2 were below Trend bound and potentially included trending responses that lasted longer than the EEG data period (1.5 s) included in the MPS analyses, so we used 0 Hz as the low cutoff modulation frequency and set the high cutoff frequency as 1.33 Hz; 2) Cluster 3 and 4 were centered around 2 Hz of the modulation frequency and we selected the same modulation frequency range for both the clusters, so that later we could compare temporal dynamics of induced power within the same modulation frequency range.

The MPS analyses were first conducted at Cz to investigate whether our hypothesized neural phenomena (Fig. 1D) could be observed. We later examined other EEG channels and created a topography for each cluster to reveal distributions of modulation strength across EEG channels. We filtered the induced power according to each defined cluster and conducted the surrogate test described above on each EEG channel. We derived thresholds at one-sided alpha level of 0.05 to determine whether the power modulation was significant for each cluster on each EEG channel, which was indicated by the green dots in Fig. 2C. When we plotted the topographies in Fig. 2C, we normalized the modulation strength by subtracting the mean of the null distributions out of the empirical modulation strength so that the distribution of the modulation strength across EEG channels was highlighted in the topographies.

Temporal dynamics of clusters and power coupling in the no-cue session

The above analyses in the spectral domain did not reveal temporal details of the induced power fluctuations, so we next investigated temporal dynamics of induced power within those four clusters in the no-cue session. We first averaged the induced power within the power frequency range of one ROI/cluster and then filtered the averaged power using a band-pass filter (for Cluster 3 and 4, 1.8 – 2.8 Hz) or a low-pass filter (for Cluster 1 and 2, < 1.33 Hz) (two-pass Butterworth filter of an order of 2 implemented in Fieldtrip).

We next checked whether each cluster was independent from or coupled with other clusters. By doing this, we can validate previous findings on the coupling of neural signals between the delta and beta bands in prospective timing paradigms and validate our data-driven procedures here. More importantly, considering that the Fourier transform arbitrarily projects signals onto an orthogonal space of sinusoid waves, a complex neural signal may be broken into several clusters. For example, Cluster 3 and 4 are probably in a harmonic relationship: Cluster 4 is likely the harmonic of Cluster 3 as the power frequency of Cluster 4 (24-26 Hz) is approximately twice of the power frequency of Cluster 3 (9 – 14 Hz). If this is the case, Cluster 3 and 4 should be coupled.

We calculated phase locking values (PLV) on induced power across participants to measure the coupling between clusters at each time point. We applied the Hilbert transformation on the induced power of each cluster and then derived the phase value at each time point by calculating angle of the Hilbert coefficient. The phase difference between each pair of clusters for each participant was then used to calculate PLV across all the participants using the following formula, in which $\boldsymbol{S}$ denotes the total number of participants, $\boldsymbol{s}$ the

participant number, ***c*** the cluster, ***a*** and ***b*** numbers of two different clusters, ***t*** the time point, and $\boldsymbol{\phi}$ the phase at each time point:

$$PLV_{c_{ab},t} = |\sum_{s=1}^{S} e^{(\phi_{s,t,c_a} - \phi_{s,t,c_b})*i}|/S$$

If two clusters are consistently coupled across participants, the phase differences across all the participants at one time point should concentrate around a specific phase value – a high PLV value (low circular variance across participants). Accordingly, we constructed a null hypothesis to determine the significant time points of PLV values through a surrogate test in the spectral domain. In each participant's data, we reset the phase of each cluster at one time point randomly by multiplying a complex number of a random phase and a norm of 1 with the Hilbert coefficient. By doing this, we disrupted the phase series of each cluster independently and hence the coupling strength between two clusters. We then calculated new PLV values of each pair of clusters across all the participants. We conducted this procedure 1000 times and constructed a null distribution of PLV at each time point, from which we derived a threshold of a one-sided alpha level of 0.05 for each cluster pair.

Merge Cluster 2, 3, and 4 into Cluster 234 and further analyses in the no-cue session

From Fig. 2D, we observed that Cluster 1 slowly changed along the entire trials and showed a global trajectory capturing the length of the noise piece, whereas the dynamics of Cluster 2, 3 and 4 seemingly did not correspond to the temporal structure in each trial. However, a peak after stimulus onset was observed in Cluster 2 around the first temporal location, Cluster 3 and 4 were coupled and both showed peaks around the second and the third temporal locations (Fig. 2D&S2C), and Cluster 2 and 4 are within the same EEG power frequency band (Fig. 2B). Overall, the above observation gave us a clue that Cluster 2, 3 and 4 may collectively capture the three temporal locations of the target tone. We tried averaging the EEG power over Cluster 2, 3, and 4 and merged them into Cluster 234. Surprisingly, the merged EEG power dynamics showed a salient neural signature marking the temporal locations, even if only the white noise was presented. By plotting Cluster 1 and Cluster 234 together, it can be clearly seen that Cluster 1 in the delta band reflected the global trial structure (the length of noise pieces) and that Cluster 234 captured the local temporal locations of the target tone. As we aimed to find neural signatures tracking the temporal locations and similar findings in the delta band, corresponding to Cluster 1, have been

reported before to encode event duration or fixed intervals (van Wassenhove and Lecoutre 2015; Breska and Deouell 2017) , we next focused on Cluster 234.

The dynamics of Cluster 234 contains a downward linear trend and a rhythmic component (Fig. 2). To single out the rhythmic component, we fitted a line to Cluster 234 from -1 s to 0 s and subtracted out the fitted line. The resulted neural power modulation, Corrected cluster 234, clearly showed peaks after each temporal location. We then individually fitted a Gaussian function to Corrected cluster 234 from each tone onset position to 0.5 s afterwards. From the fitted Gaussian curves, we derived the peak latency (the position of the mean in the fitted Gaussian curve) and the standard deviation (SD). We averaged each neural peak of Corrected cluster 234 within one SD centered around the peak latency (-0.5 SD + tone location to 0.5 SD + tone location) for each participant and conducted a one-way rmANOVA with Temporal location as the main factor.

To quantify and visualize the neural tracking of the temporal locations across all EEG channels, we calculated the modulation strength of Corrected cluster 234 from 0.4 s to 1.9 s at each channel (SD over this period of signal) and plotted the topography of the SD. The time range was chosen based on the first zero-crossing point after the noise onset and the time point after the third peak showing the lowest power magnitude. The first zero-crossing point indicated the rising of the EEG power above zero and the second time point the EEG power falling towards zero. The SD over this time range represented the degree of EEG power fluctuations, which can be seen as the amplitude summed over all the frequency components of Corrected cluster 234 within this time range.

Block number on power modulation of Corrected cluster234 in the no-cue session

We next examined how learning from the first to the third block affected the sequential temporal anticipation. The participants probably could not precisely anticipate the temporal locations of the target tone in the first block of the no-cue session, but gradually learned the exact temporal locations in the following blocks. We expected to find that the magnitude of Corrected cluster 234 increases from the first block to the third block.

We extracted the induced power from the spectral regions defined in Cluster 2, 3, and 4 in the data from each block and merged the three clusters into Cluster 234 in each block. As described in the above analyses, the MPS calculation was not affected by differences of baseline corrections, so the baseline differences between the blocks did not contribute to the following analyses. To correct Cluster234 in each block, we adopted the fitted baseline from the above analyses on the trials pooled across all three blocks, so that Corrected cluster 234 in

each block went through the same baseline correction procedure. Similarly, following the above procedures, we fitted Gaussian curves to peaks after each temporal location in each block (Fig. 3&S5) to extract peak magnitudes. We conducted a two-way rmANOVA on the peak magnitudes with Temporal location and Block number as the main factors to examine the learning effect (Fig. 3B).

Correlation between behavioral performance and Corrected cluster234 in the no-cue session

We correlated d-prime values with the peak magnitude of Corrected cluster 234 in each block to examine whether the identified neural signature coding temporal locations in the no-tone trials correlated with the tone detection performance in the tone trials (Fig. 3&S4). It is possible that, since the induced power of Corrected cluster 234 was derived in the no-tone trials, the power peaks may be caused by false alarms – the participants thought they heard a tone even though there was no tone, so we then correlated the power peaks with the false alarm rates.

Quantify peak latency change of Corrected cluster 234 across blocks and temporal locations

It can be seen that the peak latencies of Corrected cluster 234 decreased from the first temporal location to the third and were modulated by the block number (Fig. 2G&3A). We here verified this observation quantitively. One issue here was that the peak latencies determined by the Gaussian fits were derived from the group-averaged induced power; individual participant's data were noisy so that the Gaussian fits to each participant's neural power were not accurate. This prevented us from drawing conclusions by conducting parametric statistic tests over participants (e.g. rmANOVA). To circumvent this issue, we resorted to group-level permutation tests.

We first used a permutation test to examine one observation from Fig. 3A: the peak latencies were different between blocks. We constructed a null hypothesis that, at each temporal location, there is no peak latency difference between blocks. One reason for constructing this null hypothesis first was that the null distribution constructed at each temporal location could be viewed as a distribution that averages out the factor of the block number. Later, by comparing three null distributions of peak latency in the three temporal locations, we could determine whether the peak latencies of Corrected cluster 234 varied with the temporal locations.

To construct the null hypothesis of the peak latency at each temporal location across blocks, we permuted the block label in a way that each participant's data in each block were

randomly assigned with a new block number. We grouped the permuted data into three new blocks at each temporal location and derived Corrected cluster 234 in each new block. We fitted a Gaussian curve to each power peak to derive the peak latency by following the above procedure. We conducted this permutation procedure 1000 times and derived nine null distributions of peak latencies (3 temporal locations * 3 blocks). As the block labels were shuffled and were not meaningful any more in the null distributions, we pooled the null distributions over three blocks and had one null distribution of 3000 samples for each temporal location. We derived thresholds at a two-sided alpha level of 0.05 (the 76th latency and the 2925th latency) from each null distribution (Fig. 3D). The thresholds derived can be viewed as significant boundaries – if a peak latency from one block of the empirical data falls within the boundaries, the peak latency does not significantly differ from peak latencies of the other blocks; if a peak latency is outside of those significant boundaries, it is then significantly different from the other peak latencies. The boundaries, in a sense, represented an interval that the peak latencies of the three blocks are uniform, statistically speaking, so we called them as the 'uniform' bound and the permutation test as the 'uniform' test.

The null distributions of peak latency at each temporal location represented distributions of possible peak latencies without considering the factor of the block number, as each distribution was derived from random combinations of the participants' data drawn from the three blocks. Therefore, by randomly selecting a sample from each null distribution and examining how the peak latencies changed from the null distribution of the first temporal location to the third, we could further derive a distribution of the peak latency change that could be used to determine whether the peak latencies significantly varied with the temporal locations. One simple way to quantify the latency change was to fit a straight line to the peak latencies with the temporal location as the independent variable of values from 1 to 3 and the latency as the dependent variable. If the slope is significantly negative, it can be concluded that the peak latency decreases significantly with the temporal locations. Following this logic, we fitted a line to peak latencies randomly selected from the three null distributions and constructed a distribution of slope of 3000 samples. We derived a threshold at a two-sided alpha level of 0.05 from the slope distribution and compared the thresholds with zero to determine whether the peak latencies significantly shortened from the first temporal location to the third (Fig. 3F).

Modulation spectrum analysis on inter-trial phase coherence (ITPC) in the no-cue session

Besides the power measurement of neural signals, neural phases can also reflect temporal anticipation (reference) – the anticipation for a future event may cause neural phases to concentrate around a certain phase value, as suggested in previous literature (Samaha et al. 2015; Breska and Deouell 2017). If this is the case, we would expect that the neural phases in the no-tone trials were consistently locked to the tagged temporal locations and hence the variance of neural phases across trials decreased before or after the temporal locations. Accordingly, a 2-Hz component should be observed in the inter-trial phase coherence (ITPC) that measures consistency of neural phases across trials. Here, we conducted the modulation spectrum analysis on ITPC over the no-tone trials pooled from the three blocks in the no-cue session.

The Fourier coefficients were derived in *Time-frequency decomposition of EEG signals*. We derived the neural phase at each time-frequency bin by calculating angle of the Fourier coefficient. The formula of ITPC calculation at a time–frequency bin is shown below with ***N*** denoting the total number of the trials, ***n*** the trial number, ***f*** the frequency point, ***t*** the time point, and $\boldsymbol{\phi}$ the phase value at each time-frequency bin:

$$ITPC_{t,f} = |\sum_{n=1}^{N} e^{\phi_{n,t,f}*i}|/N$$

We used the valid trials across the three blocks in the no-cue session after *the EEG preprocessing* to calculate ITPC. We selected the time range between 0.25 s and 1.75 after the noise onset to avoid effects of onset and offset responses of the white noise. Then, following the procedures in *Modulation power spectrum (MPS) of induced power in the no-cue session*, we conducted MPS calculation as well as surrogate tests on ITPC values to determine significant modulation clusters. Only one salient cluster was shown below Trending bound and in the theta and delta bands along the axis of ITPC frequency (Fig. S2). Since no 2-Hz modulation component was found in the ITPC values, showing that the neural phases did not code the temporal locations, we did not conduct any further analyses on the neural phases.

MPS of EEG power in the cued session

The cued session served as a control condition or a comparison to the no-cue session, in a manner that we could test whether the 2-Hz neural signature observed in the no-cue session was related to the specific task demand and the corresponding cognitive processes. In the cued session, as the participants knew the temporal location of the forthcoming target tone at

the beginning of each trial, the participants did not need to anticipate the three temporal locations sequentially, but only needed to pay attention to the cued location and ignored the un-cued temporal locations. Hence, we should not observe in the cued session the 2-Hz neural signature revealed in the no-cue session. Alternatively, if the 2-Hz neural signature was still robustly observed in the cued session, this neural signature could be simply related to the trial structure but the anticipating process.

We followed the procedure of MPS analyses in the no-cue session and calculated MPS over all the trials of the three cued temporal locations when the cues were invalid and the participants were only listening to white noise. Although a few significant bins appeared around the 2 Hz of modulation frequency, they were not robustly shown in every surrogate test. The number of the significant spectral bins was larger than 2 in one or two repetitions of the surrogate test but was not in other repetitions. And, the modulation strength of the spectral bins around 2 Hz was much weaker than the corresponding spectral bins in the no-cue session. Therefore, we did not consider that spectral bins around 2 Hz of the modulation frequency to be robust.

We found three robust clusters in the MPS in the cued session below Trend bound (1.33 Hz) (Fig. S3). Cluster 1 found here overlapped with the Cluster 1 found in the no-cue session and similarly represented encoding of the noise length on a supra-second scale. Cluster 2 and 3 did not overlap with any clusters found in the no-cue session and their temporal dynamics did not straightforwardly correlate with the trial structure or the behavioral results. We then combined Cluster 2 and 3 into Cluster 23, but still did not observe any correspondence with the trial structure or the task. We did not further explore this set of results in the cued session in the current study, since the results here have demonstrated that the 2-Hz neural signature marking the temporal locations in the no-cue session indeed related to the sequential temporal anticipation for the temporal locations.

Train recurrent neural network models to accomplish the tone-detection task in the no-cue session

To gain insights on the functional aspects of our neural findings and uncover computational mechanisms underlying the sequential temporal anticipation, we turned to recurrent neural networks (RNNs) trained under the same behavioral paradigm in the current study. One advantage of employing neural network models is that RNNs are flexible to train and can be tested on variations of the current behavioral task to test different hypotheses (Yang and Wang 2021). Moreover, RNN structures can be manipulated and perturbed; RNN inner

dynamics is accessible to be examined. If the trained RNNs can indeed accomplish the behavioral task used in the current study and show a task performance comparable to the human participants' performance, we then study RNN hidden activities and structures to gain insights on what procedures are conducted to accomplish the tone detection task and anticipate the temporal locations (Yamins et al. 2014; Yang et al. 2019).

We chose to train continuous-time recurrent neural networks (CRNNs) with mini-batch gradient descent learned by backpropagation. The structure of CRNNs is defined as the following equation:

$$\tau \frac{dx_i}{dt} = -x_i + \sum_{j=1}^{N} W_{ij}^{rec} r_j + \sum_{k=1}^{N_{in}} W_{ik}^{in} u_k + \xi_i$$

$$r_i = [x_i]_+$$

$$z_l = \sum_{i=1}^{N} W_{li}^{out} r_{i,T}$$

The number of hidden units, $\boldsymbol{N}$, was 100. We tested CRNNs with fewer hidden units (64 and 32) and found out that those CRNNs could not be trained to achieve the human-level performance when the external noise level was high in our settings. The time step size, $\boldsymbol{\Delta t}$, was 25 ms (a sampling rate of 20 Hz), which was sufficient to capture the dynamics at 2 Hz shown in our neural findings and made the training efficiently. $\boldsymbol{\tau}$ is the time constant of the network units and was set to 100 ms. $\boldsymbol{u}$ is the input vector, so $\boldsymbol{k}$ can only be 1. $\boldsymbol{W^{in}}$ is a $1 \times N$ matrix of connection weights from the inputs to N network units. $\boldsymbol{W^{rec}}$ is an $N \times N$ matrix of recurrent connection weights between network units. $\boldsymbol{\xi}$ are N independent Gaussian white noise processes with zero mean and unit variance that represent noise intrinsic to the network. The rectified linear activation function (ReLU) was applied on hidden units. $\boldsymbol{W^{out}}$ is an $N \times 2$ matrix of connection weights from N network units to the outputs (two choices: yes or no). The outputs, $\boldsymbol{z}$, were generated after the input is over; the ending time point is indicated by $\boldsymbol{T}$. The CRNN, a continuous-time formulation of a standard RNN, was converted to a discrete-time version through the Euler method for implementation.

The connections between hidden units were constraint according to Dale's rule: 80% of units were fixed to be excitatory units and 20% of units were inhibitory units (Song, Yang, and Wang 2016). We initialized the weights by generating random weights using the normal distribution with the mean of zero and the standard deviation (SD) of square root of 5, taking absolute values of all weights, multiplying 20% of weights with -1 to set them as inhibitory

units, and equalizing the gain between excitatory units and inhibitory units. The diagonal of the weight matrix (self-connections of units) was fixed at 0. The internal noise level of ξ was set to 0.01, so that we could focus on studying the effect of noise level of inputs by varying the noise level of inputs systemically.

The inputs fed to CRNNs was constructed in NeuroGym (https://neurogym.github.io/) to mimic the trial structure in the no-cue session (Fig. 1A&4A). The magnitude of the fixation period in each trial was set as zero with additive gaussian noise of the mean of zero and SD of 0.05. The length of the fixation period was varied from 1000 ms to 1500 ms in different batches during training in a step size of 25 ms. For example, in one batch, the fixation period was 1000 ms; in another batch, the fixation period became 1175 ms. The stimulus period was 2 s and the noise was gaussian noise with the mean of 1 and various levels of SD according to the following training designs. The target magnitude was set as 1 and was added to the stimulus period at either 500 ms, 1000 ms, or 1500 ms. The CRNNs made decision by the end of the stimulus period to answer whether there was a target in a trial – a yes-no response. We did not simulate the reaction time but only aimed to simulate the procedure of human participants detecting a tone and indicating their decision by pressing a button as soon as the white noise was over. The loss was calculated only by the end of trials using cross entropy.

During training CRNNs, each mini-batch included 16 trials with each trial equally likely coming from each of four conditions (three tone trial conditions and one no-tone trial condition) (Fig. 4A). We set the maximum training sessions to 40000 and terminated the training if the mean accuracy over last 100 mini-batches was above 99 percentage correct. We used ADAM (Kingma and Ba 2014) during training and scheduled the learning rate to drop in three steps: 0.001 for 10000 sessions, then 0.0001 for 20000 following sessions, and 0.00001 for the last 10000 sessions. All the training procedures were implemented in PyTorch 1.7.0 (Paszke et al. 2019).

The main parameter we varied during training and testing CRNNs was the external noise level of the stimulus period. The external noise was added to the stimuli and was shared across all the hidden units of CRNNs. We trained five CRNNs at five external noise levels, separately: 0.01, 0.05, 0.1, 0.15, and 0.2 SD, and tested each trained CRNN at 7 noise levels: 0.01, 0.05, 0.1, 0.15, 0.2, 0.25, and 0.3 SD.

CRNN behavioral performance, hidden activities, and modulation spectra

We trained five CRNNs at different noise levels, respectively, and tested them across all the 7 testing noise levels. We generated 27 sets of behavioral data at each testing noise level to match the number of recruited human participants. Following the same procedure in *Behavioral data analysis*, we calculated d-prime values from CRNN outputs. Then, we selected one testing noise level for each of the five CRNNs according to how close the group-level d-prime value derived at this testing noise level matched the group-level mean of the human behavioral data averaged over the three temporal locations in Fig. S1C.

We fed each of the five trained CRNNs with 40 no-tone trials at the chosen testing noise level and extracted activities of their hidden units. As neural signals recorded by EEG reflect summarized neural activities over both excitatory and inhibitory synapses, similarly, we averaged the CRNN hidden activities over both excitatory and inhibitory units. We analyzed the averaged hidden activities following the procedure used in calculating the modulation spectrum of EEG signals. We selected the averaged hidden activity of each trial from 0.25 s to 1.75 s after noise onset and applied the Fourier transform, and then extracted the amplitude of the Fourier coefficients for each trial and averaged the modulation spectra over all the 40 no-tone trials to generate an averaged modulation spectrum of the hidden activities. We found that the CRNN trained at the noise level of 0.15 SD and tested at the noise level of 0.2 SD showed the largest 2 Hz modulation strength (Fig. 4C), so we selected this CRNN for the subsequent investigations as it both matched the human behavioral performance and the neural findings (Fig. 2). The hidden activities of CRNNs shown in Fig. 4D and Fig. 4E were derived from averaging hidden activities over 40 no-tone trials.

Train and test CRNNs with a uniform probabilistic distribution of tone position from 0.25 s to 1.75 s

We trained a CRNN with a new trial structure in which the target tone can appear any time from 0.25 s to 1.75 s with an equal probability (Fig. 4F), so that the newly trained CRNN can be compared with the CRNNs trained with a target tone appearing at one of the three temporal locations. The motivation here is twofold: 1) If the 2 Hz modulation component was driven by the specific temporal locations, the CRNN trained with the new trial structure should not contain the 2 Hz modulation component; 2) We aimed to test tone detection performance of the CRNN trained with the three temporal locations at other time points.

In the original trial structure for training CRNNs (Fig. 4A), 75 percent of trials had the target tone imbedded and other 25 percent of trials were no-tone trials. Accordingly, here, we fixed 25 percent of trials to be no-tone trials and, in the other 75 percent of trials, a target

randomly appeared any time from 0.25 s to 1.75 s. In each training batch, the trial type was randomly sampled with a probability of 0.75 to be tone trials and 0.25 to be no-tone trials. We chose the training noise level to be 0.15 SD. The other parameters and procedures for training CRNNs were the same as in the previous CRNN training procedures. After 40000 batches, the training accuracy reached 0.97.

We tested both the newly trained CRNN and the CRNN trained at 0.15 SD with three temporal locations at a testing noise level of 0.2 SD. We chose a different way to analyze the tone detection performance, in contrast with the above behavioral analyses (Fig. 4B). We independently generated 40 no-tone trials and 40 tone trials with the target at one time point, and asked CRNNs to generate outputs. For each time point, we measured a hit rate from the tone trials and a false alarm rate from the no-tone trials and calculated a d-prime value. This d-prime calculation was conducted from 0.25 s to 1.75 s with a step of 0.025 s (Fig. 4G).

Perturb E/I ratio of the CRNN trained at the noise level of 0.15 SD

To account for the neural power precession observed in Fig. 3D, we perturbed excitation-inhibition balance of trained CRNNs. The perturbed CRNNs may generate new hidden dynamics that potentially capture the phenomenon of the shortened peak latencies of Corrected cluster 234.

We conducted the perturbation on the CRNN trained at the noise level of 0.15 SD with three temporal locations and derived the hidden activities at the testing noise level of 0.2 SD, as the above analyses showed that this CRNN captured the neural dynamics successfully (Fig. 4C-E). To perturb the excitation-inhibition balance, we changed the weights of excitatory and inhibitory units. For example, we increased the weights of excitatory units by 10 percent so that the excitatory units sent stronger projections to all units. However, only increasing excitatory unit weights would disturb the CRNN and potentially drive it to a chaotic state or to a silent state. Hence, we changed the weights of excitatory and inhibitory units by the same amount but in opposite directions. For example, when the perturbation was 10 percent, we increased the excitatory weights by 10 percent and, correspondingly, decreased the inhibitory weights by 10 percent.

We perturbed the excitation-inhibition balance from -40 percent (negative gain for excitatory weights and positive gain for inhibitory weights) to 40 percent with a step of 5 percent. After we perturbed CRNNs, we asked the CRNNs to generate behavioral outputs and calculated d-prime values. Meaningful positive d-prime values were generated when the perturbations were -10, -5, 5, 10, and 15 percent, so we only analyzed the CRNNs under

those five perturbations. We derived the averaged hidden activities of the perturbed CRNNs from 40 no-tone trials at a noise level of 0.2 SD (Fig. 4H) and calculated d-prime values at three temporal locations from 27 simulated datasets following the procedure used in analyzing human behavioral data (Fig. 4I).

References

Arnal, Luc H., Keith B. Doelling, and David Poeppel. 2014. "Delta–Beta Coupled Oscillations Underlie Temporal Prediction Accuracy." *Cerebral Cortex* 25 (9): 3077–85. https://doi.org/10.1093/cercor/bhu103.

Arnal, Luc H., and Anne-Lise Giraud. 2012. "Cortical Oscillations and Sensory Predictions." *Trends in Cognitive Sciences* 16 (7): 390–98. https://doi.org/10.1016/j.tics.2012.05.003.

Assaneo, M. Florencia, Pablo Ripollés, Joan Orpella, Wy Ming Lin, Ruth de Diego-Balaguer, and David Poeppel. 2019. "Spontaneous Synchronization to Speech Reveals Neural Mechanisms Facilitating Language Learning." *Nature Neuroscience* 22 (4): 627–32. https://doi.org/10.1038/s41593-019-0353-z.

Auksztulewicz, Ryszard, Nicholas E. Myers, Jan W. Schnupp, and Anna C. Nobre. 2019. "Rhythmic Temporal Expectation Boosts Neural Activity by Increasing Neural Gain." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 39 (49): 9806–17. https://doi.org/10.1523/JNEUROSCI.0925-19.2019.

Bauer, Markus, Christian Kluge, Dominik Bach, David Bradbury, Hans Jochen Heinze, Raymond J. Dolan, and Jon Driver. 2012. "Cholinergic Enhancement of Visual Attention and Neural Oscillations in the Human Brain." *Current Biology: CB* 22 (5): 397–402. https://doi.org/10.1016/j.cub.2012.01.022.

Bauer, Markus, Max-Philipp Stenner, Karl J. Friston, and Raymond J. Dolan. 2014. "Attentional Modulation of Alpha/Beta and Gamma Oscillations Reflect Functionally Distinct Processes." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 34 (48): 16117–25. https://doi.org/10.1523/JNEUROSCI.3474-13.2014.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society* 57 (1): 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Bolger, Deirdre, Jennifer T. Coull, and Daniele Schön. 2014. "Metrical Rhythm Implicitly Orients Attention in Time as Indexed by Improved Target Detection and Left Inferior Parietal Activation." *Journal of Cognitive Neuroscience* 26 (3): 593–605. https://doi.org/10.1162/jocn_a_00511.

Bree, Sander van, Ediz Sohoglu, Matthew H. Davis, and Benedikt Zoefel. 2021. "Sustained Neural Rhythms Reveal Endogenous Oscillations Supporting Speech Perception." *PLoS Biology* 19 (2): e3001142. https://doi.org/10.1371/journal.pbio.3001142.

Breska, Assaf, and Leon Y. Deouell. 2017. "Neural Mechanisms of Rhythm-Based Temporal Prediction: Delta Phase-Locking Reflects Temporal Predictability but Not Rhythmic Entrainment." *PLoS Biology* 15 (2): e2001665. https://doi.org/10.1371/journal.pbio.2001665.

Breska, Assaf, and Richard B. Ivry. 2020. "Context-Specific Control over the Neural Dynamics of Temporal Attention by the Human Cerebellum." *Science Advances* 6 (49). https://doi.org/10.1126/sciadv.abb1141.

Brunel, Nicolas, and Xiao-Jing Wang. 2003. "What Determines the Frequency of Fast Network Oscillations with Irregular Neural Discharges? I. Synaptic Dynamics and Excitation-Inhibition Balance." *Journal of Neurophysiology* 90 (1): 415–30. https://doi.org/10.1152/jn.01095.2002.

Buzsáki, György, and Rodolfo Llinás. 2017. "Space and Time in the Brain." *Science*, October. https://science.sciencemag.org/content/358/6362/482.abstract.

Chen, L. 1982. “Topological Structure in Visual Perception.” *Science* 218 (4573): 699–700. https://doi.org/10.1126/science.7134969.

Correa, Angel, and Anna C. Nobre. 2008. “Neural Modulation by Regularity and Passage of Time.” *Journal of Neurophysiology* 100 (3): 1649–55. https://doi.org/10.1152/jn.90656.2008.

Coull, J. T., and A. C. Nobre. 1998. “Where and When to Pay Attention: The Neural Systems for Directing Attention to Spatial Locations and to Time Intervals as Revealed by Both PET and FMRI.” *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 18 (18): 7426–35. https://www.ncbi.nlm.nih.gov/pubmed/9736662.

Coull, Jennifer T., Julien Cotti, and Franck Vidal. 2016. “Differential Roles for Parietal and Frontal Cortices in Fixed versus Evolving Temporal Expectations: Dissociating Prior from Posterior Temporal Probabilities with FMRI.” *NeuroImage* 141 (November): 40–51. https://doi.org/10.1016/j.neuroimage.2016.07.036.

Cravo, André M., Gustavo Rohenkohl, Karin Moreira Santos, and Anna C. Nobre. 2017. “Temporal Anticipation Based on Memory.” *Journal of Cognitive Neuroscience* 29 (12): 2081–89. https://doi.org/10.1162/jocn_a_01172.

Dale, Henry. 1935. “Pharmacology and Nerve-Endings.” *Proceedings of the Royal Society of Medicine* 28 (3): 319–32. https://doi.org/10.1177/003591573502800330.

Ding, Nai, Lucia Melloni, Hang Zhang, Xing Tian, and David Poeppel. 2016. “Cortical Tracking of Hierarchical Linguistic Structures in Connected Speech.” *Nature Neuroscience* 19 (1): 158–64. https://doi.org/10.1038/nn.4186.

Doelling, Keith B., M. Florencia Assaneo, Dana Bevilacqua, Bijan Pesaran, and David Poeppel. 2019. “An Oscillator Model Better Predicts Cortical Entrainment to Music.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (20): 10113–21. https://doi.org/10.1073/pnas.1816414116.

Doelling, Keith B., and David Poeppel. 2015. “Cortical Entrainment to Music and Its Modulation by Expertise.” *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1508431112.

Donhauser, Peter W., and Sylvain Baillet. 2020. “Two Distinct Neural Timescales for Predictive Speech Processing.” *Neuron* 105 (2): 385-393.e9. https://doi.org/10.1016/j.neuron.2019.10.019.

Faisal, A. Aldo, Luc P. J. Selen, and Daniel M. Wolpert. 2008. “Noise in the Nervous System.” *Nature Reviews. Neuroscience* 9 (4): 292–303. https://doi.org/10.1038/nrn2258.

Friston, Karl. 2005. “A Theory of Cortical Responses.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 360 (1456): 815–36. https://doi.org/10.1098/rstb.2005.1622.

Ghitza, Oded. 2012. “On the Role of Theta-Driven Syllabic Parsing in Decoding Speech: Intelligibility of Speech with a Manipulated Modulation Spectrum.” *Frontiers in Psychology* 3 (July): 238. https://doi.org/10.3389/fpsyg.2012.00238.

Grabenhorst, Matthias, Laurence T. Maloney, David Poeppel, and Georgios Michalareas. 2021. “Two Sources of Uncertainty Independently Modulate Temporal Expectancy.” *Proceedings of the National Academy of Sciences of the United States of America* 118 (16). https://doi.org/10.1073/pnas.2019342118.

Grabenhorst, Matthias, Georgios Michalareas, Laurence T. Maloney, and David Poeppel. 2019. "The Anticipation of Events in Time." *Nature Communications* 10 (1): 1–13. https://www.nature.com/articles/s41467-019-13849-0.

Haegens, Saskia, Lisa Luther, and Ole Jensen. 2012. "Somatosensory Anticipatory Alpha Activity Increases to Suppress Distracting Input." *Journal of Cognitive Neuroscience* 24 (3): 677–85. https://doi.org/10.1162/jocn_a_00164.

Haegens, Saskia, and Elana Zion Golumbic. 2018. "Rhythmic Facilitation of Sensory Processing: A Critical Review." *Neuroscience and Biobehavioral Reviews* 86 (March): 150–65. https://doi.org/10.1016/j.neubiorev.2017.12.002.

Harald Baayen, R., and Petar Milin. 2010. "Analyzing reaction times." *International journal of psychological research* 3 (2): 12–28. https://doi.org/10.21500/20112084.807.

Heilbron, Micha, and Maria Chait. 2018. "Great Expectations: Is There Evidence for Predictive Coding in Auditory Cortex?" *Neuroscience* 389 (October): 54–73. https://doi.org/10.1016/j.neuroscience.2017.07.061.

Henry, Molly J., Björn Herrmann, and Jonas Obleser. 2014. "Entrained Neural Oscillations in Multiple Frequency Bands Comodulate Behavior." *Proceedings of the National Academy of Sciences of the United States of America* 111 (41): 14935–40. https://doi.org/10.1073/pnas.1408741111.

Henry, Molly J., and Jonas Obleser. 2012. "Frequency Modulation Entrains Slow Neural Oscillations and Optimizes Human Listening Behavior." *Proceedings of the National Academy of Sciences of the United States of America* 109 (49): 20095–100. https://doi.org/10.1073/pnas.1213390109.

Herbst, Sophie K., Lorenz Fiedler, and Jonas Obleser. 2018. "Tracking Temporal Hazard in the Human Electroencephalogram Using a Forward Encoding Model." *ENeuro* 5 (2). https://doi.org/10.1523/ENEURO.0017-18.2018.

Herrmann, Björn, Molly J. Henry, Saskia Haegens, and Jonas Obleser. 2016. "Temporal Expectations and Neural Amplitude Fluctuations in Auditory Cortex Interactively Influence Perception." *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2015.09.019.

Hillyard, S. A., E. K. Vogel, and S. J. Luck. 1998. "Sensory Gain Control (Amplification) as a Mechanism of Selective Attention: Electrophysiological and Neuroimaging Evidence." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 353 (1373): 1257–70. https://doi.org/10.1098/rstb.1998.0281.

Huron, David. 2008. *Sweet Anticipation: Music and the Psychology of Expectation*. MIT Press. https://play.google.com/store/books/details?id=sgr-DwAAQBAJ.

Issa, John B., Gilad Tocker, Michael E. Hasselmo, James G. Heys, and Daniel A. Dombeck. 2020. "Navigating Through Time: A Spatial Navigation Perspective on How the Brain May Encode Time." *Annual Review of Neuroscience* 43 (July): 73–93. https://doi.org/10.1146/annurev-neuro-101419-011117.

Janssen, Peter, and Michael N. Shadlen. 2005. "A Representation of the Hazard Rate of Elapsed Time in Macaque Area LIP." *Nature Neuroscience* 8 (2): 234–41. https://doi.org/10.1038/nn1386.

Jensen, Ole, and Ali Mazaheri. 2010. "Shaping Functional Architecture by Oscillatory Alpha Activity: Gating by Inhibition." *Frontiers in Human Neuroscience* 4 (November): 186. https://doi.org/10.3389/fnhum.2010.00186.

Jones, M. R., G. Kidd, and R. Wetzel. 1981. “Evidence for Rhythmic Attention.” *Journal of Experimental Psychology. Human Perception and Performance* 7 (5): 1059–73. https://doi.org/10.1037//0096-1523.7.5.1059.

Kingdom, Frederick A. A., and Nicolaas Prins. 2016. *Psychophysics: A Practical Introduction*. Academic Press. https://play.google.com/store/books/details?id=3sHQBAAAQBAJ.

Kingma, Diederik P., and Jimmy Ba. 2014. “Adam: A Method for Stochastic Optimization.” *ArXiv [Cs.LG]*. arXiv. http://arxiv.org/abs/1412.6980.

Klimesch, Wolfgang. 2012. “Alpha-Band Oscillations, Attention, and Controlled Access to Stored Information.” *Trends in Cognitive Sciences* 16 (12): 606–17. https://doi.org/10.1016/j.tics.2012.10.007.

Klimesch, Wolfgang, Paul Sauseng, and Simon Hanslmayr. 2007. “EEG Alpha Oscillations: The Inhibition–Timing Hypothesis.” *Brain Research Reviews* 53 (1): 63–88. https://doi.org/10.1016/j.brainresrev.2006.06.003.

Kontsevich, L. L., and C. W. Tyler. 1999. “Bayesian Adaptive Estimation of Psychometric Slope and Threshold.” *Vision Research* 39 (16): 2729–37. https://doi.org/10.1016/s0042-6989(98)00285-5.

Kotz, Sonja A., and Maren Schmidt-Kassow. 2015. “Basal Ganglia Contribution to Rule Expectancy and Temporal Predictability in Speech.” *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior* 68 (July): 48–60. https://doi.org/10.1016/j.cortex.2015.02.021.

Lakatos, Peter, Annamaria Barczak, Samuel A. Neymotin, Tammy McGinnis, Deborah Ross, Daniel C. Javitt, and Monica Noelle O’Connell. 2016. “Global Dynamics of Selective Attention and Its Lapses in Primary Auditory Cortex.” *Nature Neuroscience* 19 (12): 1707–17. https://doi.org/10.1038/nn.4386.

Larrouy-Maestri, Pauline, and Peter Q. Pfordresher. 2018. “Pitch Perception in Music: Do Scoops Matter?” *Journal of Experimental Psychology. Human Perception and Performance* 44 (10): 1523–41. https://doi.org/10.1037/xhp0000550.

Lashley, Karl Spencer. 1951. *The Problem of Serial Order in Behavior*. Bobbs-Merrill. https://play.google.com/store/books/details?id=fPh4nQEACAAJ.

Lee, Caroline S., Mariam Aly, and Christopher Baldassano. 2021. “Anticipation of Temporally Structured Events in the Brain.” *ELife* 10 (April). https://doi.org/10.7554/eLife.64972.

Lerdahl, Fred, and Ray Jackendoff. 1983. “An Overview of Hierarchical Structure in Music.” *Music Perception: An Interdisciplinary Journal* 1 (2): 229–52. https://doi.org/10.2307/40285257.

Macmillan, Neil A., and C. Douglas Creelman. 2004. *Detection Theory: A User’s Guide*. Psychology Press. https://play.google.com/store/books/details?id=P094AgAAQBAJ.

Maris, Eric, and Robert Oostenveld. 2007. “Nonparametric Statistical Testing of EEG- and MEG-Data.” *Journal of Neuroscience Methods*. https://doi.org/10.1016/j.jneumeth.2007.03.024.

Martens, Sander, and Addie Johnson. 2005. “Timing Attention: Cuing Target Onset Interval Attenuates the Attentional Blink.” *Memory & Cognition* 33 (2): 234–40. https://doi.org/10.3758/bf03195312.

Mattiesing, Rozemarijn M., Wouter Kruijne, Martijn Meeter, and Sander A. Los. 2017. “Timing a Week Later: The Role of Long-Term Memory in Temporal Preparation.” *Psychonomic Bulletin & Review* 24 (6): 1900–1905. https://doi.org/10.3758/s13423-017-1270-3.

McClure, Samuel M., Gregory S. Berns, and P. Read Montague. 2003. "Temporal Prediction Errors in a Passive Learning Task Activate Human Striatum." *Neuron* 38 (2): 339–46. https://doi.org/10.1016/s0896-6273(03)00154-5.

Mello, Gustavo B. M., Sofia Soares, and Joseph J. Paton. 2015. "A Scalable Population Code for Time in the Striatum." *Current Biology: CB* 25 (9): 1113–22. https://doi.org/10.1016/j.cub.2015.02.036.

Merchant, Hugo, and Kielan Yarrow. 2016. "How the Motor System Both Encodes and Influences Our Sense of Time." *Current Opinion in Behavioral Sciences* 8 (April): 22–27. https://doi.org/10.1016/j.cobeha.2016.01.006.

Morillon, Benjamin, and Sylvain Baillet. 2017. "Motor Origin of Temporal Predictions in Auditory Attention." *Proceedings of the National Academy of Sciences of the United States of America* 114 (42): E8913–21. https://doi.org/10.1073/pnas.1705373114.

Morillon, Benjamin, Charles E. Schroeder, and Valentin Wyart. 2014. "Motor Contributions to the Temporal Precision of Auditory Attention." *Nature Communications*. https://doi.org/10.1038/ncomms6255.

Morillon, Benjamin, Charles E. Schroeder, Valentin Wyart, and Luc H. Arnal. 2016. "Temporal Prediction in Lieu of Periodic Stimulation." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 36 (8): 2342–47. https://doi.org/10.1523/JNEUROSCI.0836-15.2016.

Motanis, Helen, and Dean V. Buonomano. 2015. "Neural Coding: Time Contraction and Dilation in the Striatum." *Current Biology: CB*. https://doi.org/10.1016/j.cub.2015.02.057.

Näätänen, R., P. Paavilainen, T. Rinne, and K. Alho. 2007. "The Mismatch Negativity (MMN) in Basic Research of Central Auditory Processing: A Review." *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology* 118 (12): 2544–90. https://doi.org/10.1016/j.clinph.2007.04.026.

Navon, David. 1977. "Forest before Trees: The Precedence of Global Features in Visual Perception." *Cognitive Psychology* 9 (3): 353–83. https://doi.org/10.1016/0010-0285(77)90012-3.

Ng, Benedict Shien Wei, Tim Schroeder, and Christoph Kayser. 2012. "A Precluding but Not Ensuring Role of Entrained Low-Frequency Oscillations for Auditory Perception." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 32 (35): 12268–76. https://doi.org/10.1523/JNEUROSCI.1877-12.2012.

Niemi, Pekka, and Risto Näätänen. 1981. "Foreperiod and Simple Reaction Time." *Psychological Bulletin* 89 (1): 133–62. https://doi.org/10.1037/0033-2909.89.1.133.

Nobre, A. C. 2001. "Orienting Attention to Instants in Time." *Neuropsychologia* 39 (12): 1317–28. https://doi.org/10.1016/s0028-3932(01)00120-8.

Nobre, Ac, A. Correa, and Jt Coull. 2007. "The Hazards of Time." *Current Opinion in Neurobiology* 17 (4): 465–70. https://doi.org/10.1016/j.conb.2007.07.006.

Nobre, Anna C., and Freek van Ede. 2018. "Anticipated Moments: Temporal Structure in Attention." *Nature Reviews. Neuroscience* 19 (1): 34–48. https://doi.org/10.1038/nrn.2017.141.

Nobre, Anna C., and Mark G. Stokes. 2019. "Premembering Experience: A Hierarchy of Time-Scales for Proactive Attention." *Neuron* 104 (1): 132–46. https://doi.org/10.1016/j.neuron.2019.08.030.

Nozaradan, Sylvie, Michael Schwartze, Christian Obermeier, and Sonja A. Kotz. 2017. "Specific Contributions of Basal Ganglia and Cerebellum to the Neural Tracking of Rhythm." *Cortex; a Journal Devoted to the*

Study of the Nervous System and Behavior 95 (October): 156–68. https://doi.org/10.1016/j.cortex.2017.08.015.

Oostenveld, Robert, Pascal Fries, Eric Maris, and Jan-Mathijs Schoffelen. 2011. "FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data." *Computational Intelligence and Neuroscience*. https://doi.org/10.1155/2011/156869.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

Paton, Joseph J., and Dean V. Buonomano. 2018. "The Neural Basis of Timing: Distributed Mechanisms for Diverse Functions." *Neuron* 98 (4): 687–705. https://doi.org/10.1016/j.neuron.2018.03.045.

Poeppel, David, and M. Florencia Assaneo. 2020. "Speech Rhythms and Their Neural Foundations." *Nature Reviews. Neuroscience* 21 (6): 322–34. https://doi.org/10.1038/s41583-020-0304-4.

Prins, N. 2013. "The Psi-Marginal Adaptive Method (or: How to Give Nuisance Parameters the Attention They Deserve, No More and No Less)." *Journal of Vision*. https://doi.org/10.1167/13.9.277.

Reynolds, John H., and Leonardo Chelazzi. 2004. "Attentional Modulation of Visual Processing." *Annual Review of Neuroscience* 27: 611–47. https://doi.org/10.1146/annurev.neuro.26.041002.131039.

Rimmele, Johanna, Hajnal Jolsvai, and Elyse Sussman. 2011. "Auditory Target Detection Is Affected by Implicit Temporal and Spatial Expectations." *Journal of Cognitive Neuroscience* 23 (5): 1136–47. https://doi.org/10.1162/jocn.2010.21437.

Rimmele, Johanna M., Benjamin Morillon, David Poeppel, and Luc H. Arnal. 2018. "Proactive Sensing of Periodic and Aperiodic Auditory Patterns." *Trends in Cognitive Sciences* 22 (10): 870–82. https://doi.org/10.1016/j.tics.2018.08.003.

Salinas, E., and T. J. Sejnowski. 2001. "Correlated Neuronal Activity and the Flow of Neural Information." *Nature Reviews. Neuroscience* 2 (8): 539–50. https://doi.org/10.1038/35086012.

Samaha, Jason, Phoebe Bauer, Sawyer Cimaroli, and Bradley R. Postle. 2015. "Top-down Control of the Phase of Alpha-Band Oscillations as a Mechanism for Temporal Prediction." *Proceedings of the National Academy of Sciences of the United States of America* 112 (27): 8439–44. https://doi.org/10.1073/pnas.1503686112.

Schubotz, Ricarda I. 2007. "Prediction of External Events with Our Motor System: Towards a New Framework." *Trends in Cognitive Sciences* 11 (5): 211–18. https://doi.org/10.1016/j.tics.2007.02.006.

Sherman, Maxwell A., Shane Lee, Robert Law, Saskia Haegens, Catherine A. Thorn, Matti S. Hämäläinen, Christopher I. Moore, and Stephanie R. Jones. 2016. "Neural Mechanisms of Transient Neocortical Beta Rhythms: Converging Evidence from Humans, Computational Modeling, Monkeys, and Mice." *Proceedings of the National Academy of Sciences of the United States of America* 113 (33): E4885-94. https://doi.org/10.1073/pnas.1604135113.

Song, H. Francis, Guangyu R. Yang, and Xiao-Jing Wang. 2016. "Training Excitatory-Inhibitory Recurrent Neural Networks for Cognitive Tasks: A Simple and Flexible Framework." *PLoS Computational Biology* 12 (2): e1004792. https://doi.org/10.1371/journal.pcbi.1004792.

Spaak, Eelke, Floris P. de Lange, and Ole Jensen. 2014. "Local Entrainment of Alpha Oscillations by Visual Stimuli Causes Cyclic Modulation of Perception." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 34 (10): 3536–44. https://doi.org/10.1523/JNEUROSCI.4385-13.2014.

Spitzer, Bernhard, and Saskia Haegens. 2017. "Beyond the Status Quo: A Role for Beta Oscillations in Endogenous Content (Re)Activation." *Eneuro*. https://doi.org/10.1523/eneuro.0170-17.2017.

Strauß, Antje, Malte Wöstmann, and Jonas Obleser. 2014. "Cortical Alpha Oscillations as a Tool for Auditory Selective Inhibition." *Frontiers in Human Neuroscience*. https://doi.org/10.3389/fnhum.2014.00350.

Tal, Idan, Edward W. Large, Eshed Rabinovitch, Yi Wei, Charles E. Schroeder, David Poeppel, and Elana Zion Golumbic. 2017. "Neural Entrainment to the Beat: The 'Missing-Pulse' Phenomenon." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 37 (26): 6331–41. https://doi.org/10.1523/JNEUROSCI.2500-16.2017.

Tavano, Alessandro, Erich Schröger, and Sonja A. Kotz. 2019. "Beta Power Encodes Contextual Estimates of Temporal Event Probability in the Human Brain." *PloS One* 14 (9): e0222420. https://doi.org/10.1371/journal.pone.0222420.

Ten Oever, Sanne, Charles E. Schroeder, David Poeppel, Nienke van Atteveldt, Ashesh D. Mehta, Pierre Mégevand, David M. Groppe, and Elana Zion-Golumbic. 2017. "Low-Frequency Cortical Oscillations Entrain to Subthreshold Rhythmic Auditory Stimuli." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 37 (19): 4903–12. https://doi.org/10.1523/JNEUROSCI.3658-16.2017.

Teng, X., P. Larrouy-Maestri, and D. Poeppel. 2021. "Segmenting and Predicting Musical Phrase Structure Exploits Neural Gain Modulation and Phase Precession." *BioRxiv*. https://www.biorxiv.org/content/10.1101/2021.07.15.452556.abstract.

Teng, Xiangbin, Min Ma, Jinbiao Yang, Stefan Blohm, Qing Cai, and Xing Tian. 2020. "Constrained Structure of Ancient Chinese Poetry Facilitates Speech Content Grouping." *Current Biology: CB* 30 (7): 1299-1305.e7. https://doi.org/10.1016/j.cub.2020.01.059.

Teng, Xiangbin, Xing Tian, and David Poeppel. 2016. "Testing Multi-Scale Processing in the Auditory System." *Scientific Reports* 6 (October): 34390. https://doi.org/10.1038/srep34390.

Teng, Xiangbin, Xing Tian, Jess Rowland, and David Poeppel. 2017. "Concurrent Temporal Channels for Auditory Processing: Oscillatory Neural Entrainment Reveals Segregation of Function at Different Scales." *PLoS Biology* 15 (11): e2000812. https://doi.org/10.1371/journal.pbio.2000812.

Van Ede, Freek, Sammi R. Chekroud, Mark G. Stokes, and Anna C. Nobre. 2018. "Decoding the Influence of Anticipatory States on Visual Perception in the Presence of Temporal Distractors." *Nature Communications* 9 (1): 1–12. https://www.nature.com/articles/s41467-018-03960-z.

Wassenhove, Virginie van, and Lucille Lecoutre. 2015. "Duration Estimation Entails Predicting When." *NeuroImage* 106 (February): 272–83. https://doi.org/10.1016/j.neuroimage.2014.11.005.

Wilsch, Anna, Molly J. Henry, Björn Herrmann, Burkhard Maess, and Jonas Obleser. 2015. "Slow-Delta Phase Concentration Marks Improved Temporal Expectations Based on the Passage of Time." *Psychophysiology* 52 (7): 910–18. https://doi.org/10.1111/psyp.12413.

Wittmann, Marc. 2013. "The Inner Sense of Time: How the Brain Creates a Representation of Duration." *Nature Reviews. Neuroscience* 14 (3): 217–23. https://doi.org/10.1038/nrn3452.

Worden, Michael S., John J. Foxe, Norman Wang, and Gregory V. Simpson. 2000. "Anticipatory Biasing of Visuospatial Attention Indexed by Retinotopically Specific α-Bank Electroencephalography Increases over Occipital Cortex." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 20 (6): RC63–RC63. https://doi.org/10.1523/JNEUROSCI.20-06-j0002.2000.

Wöstmann, Malte, Mohsen Alavash, and Jonas Obleser. 2019. "Alpha Oscillations in the Human Brain Implement Distractor Suppression Independent of Target Selection." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 39 (49): 9797–9805. https://doi.org/10.1523/JNEUROSCI.1954-19.2019.

Wöstmann, Malte, Sung-Joo Lim, and Jonas Obleser. 2017. "The Human Neural Alpha Response to Speech Is a Proxy of Attentional Control." *Cerebral Cortex* 27 (6): 3307–17. https://doi.org/10.1093/cercor/bhx074.

Yamins, Daniel L. K., and James J. DiCarlo. 2016. "Using Goal-Driven Deep Learning Models to Understand Sensory Cortex." *Nature Neuroscience*. https://doi.org/10.1038/nn.4244.

Yamins, Daniel L. K., Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. 2014. "Performance-Optimized Hierarchical Models Predict Neural Responses in Higher Visual Cortex." *Proceedings of the National Academy of Sciences of the United States of America* 111 (23): 8619–24. https://doi.org/10.1073/pnas.1403112111.

Yang, Guangyu Robert, Madhura R. Joglekar, H. Francis Song, William T. Newsome, and Xiao-Jing Wang. 2019. "Task Representations in Neural Networks Trained to Perform Many Cognitive Tasks." *Nature Neuroscience* 22 (2): 297–306. https://doi.org/10.1038/s41593-018-0310-2.

Yang, Guangyu Robert, and Xiao-Jing Wang. 2021. "Artificial Neural Networks for Neuroscientists: A Primer." *Neuron* 109 (4): 739. https://doi.org/10.1016/j.neuron.2021.01.022.

Yekutieli, Daniel, and Yoav Benjamini. 1999. "Resampling-Based False Discovery Rate Controlling Multiple Test Procedures for Correlated Test Statistics." *Journal of Statistical Planning and Inference* 82 (1): 171–96. https://doi.org/10.1016/S0378-3758(99)00041-5.

Zanto, Theodore P., Joel S. Snyder, and Edward W. Large. 2006. "Neural Correlates of Rhythmic Expectancy." *Advances in Cognitive Psychology / University of Finance and Management in Warsaw* 2 (2): 221–31. https://doi.org/10.2478/v10053-008-0057-5.

Zumer, Johanna M., René Scheeringa, Jan-Mathijs Schoffelen, David G. Norris, and Ole Jensen. 2014. "Occipital Alpha Activity during Stimulus Processing Gates the Information Flow to Object-Selective Cortex." *PLoS Biology* 12 (10): e1001965. https://doi.org/10.1371/journal.pbio.1001965.

Supplementary Information

Xiangbin Teng and Ru-Yuan Zhang

Corresponding Author: Xiangbin Teng, xiangbin.teng@gmail.com

This file includes:

Figure S1 – S5

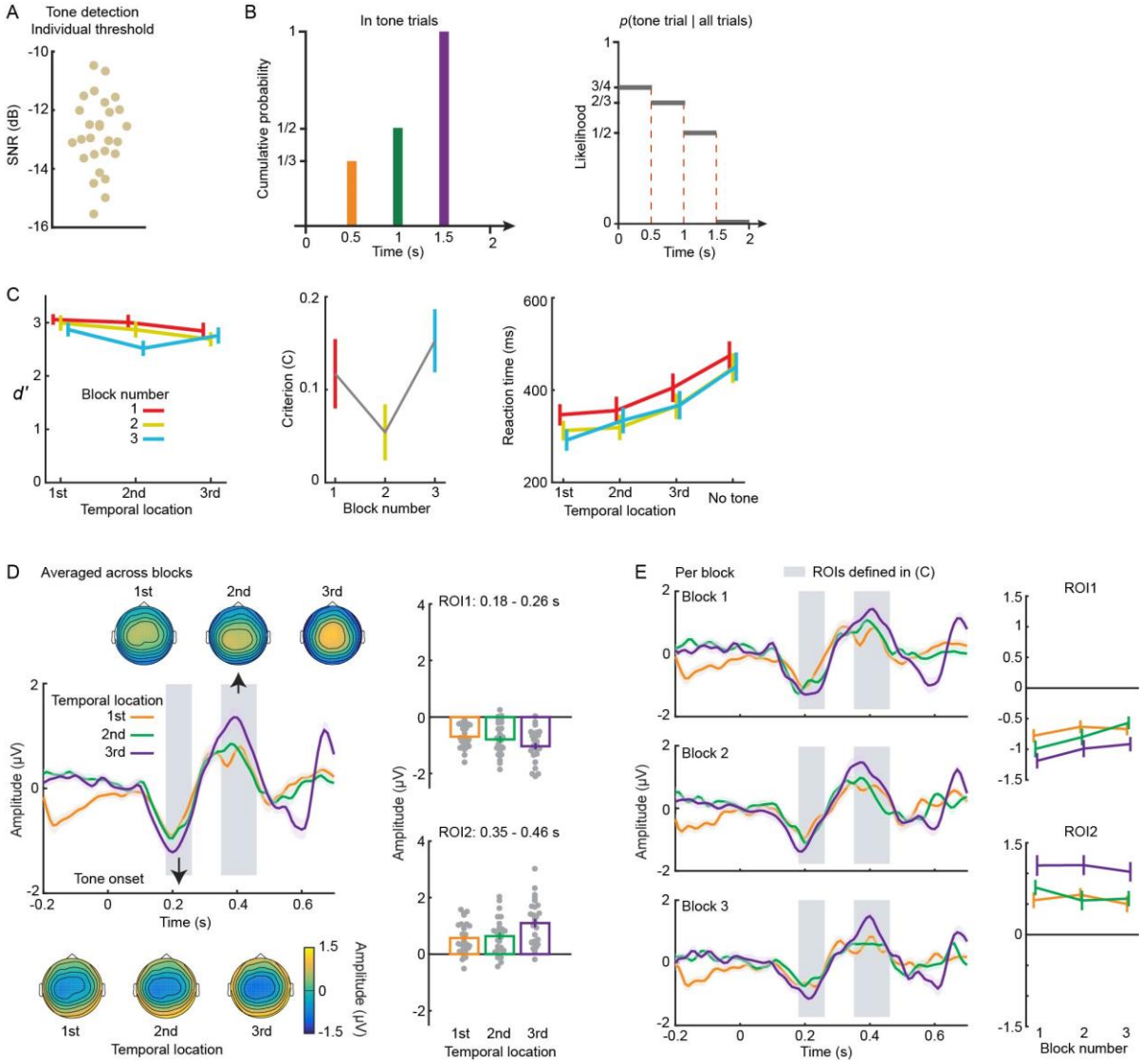

Fig. S1. Tone detection threshold, probabilities of conditions, and ERP results in the no-cue session. Related to Fig. 1&2.

(*A*) Individual tone detection thresholds measured before the no-cue session. Each dot represents each participant's threshold.

(*B*) Probabilities of conditions. The left plot shows the cumulative probability for the target tone to appear in the tone trials as the time is passing; the right plot shows that the probability of a trial is a tone trial, considering that there is a condition that there is no tone presented.

(*C*) Behavioral results. The plots, from left to right, show results of d-prime values, criteria, and reaction time, respectively. The line color denotes the block number. Error bars represent ±1 standard error of the mean over participants. We analyzed the tone detection performance under the signal detection framework and conducted a two-way repeated-measure ANOVA (rmANOVA) on the d-prime values with Temporal location and Block number as the main factors. We found a significant main effect for Block number ($F(2,34) = 5.74$, $p = .007$, $\eta_p^2 = 0.252$) with a significant linear trend showing that the participants' tone detection performance decreased with the block number ($F(1,17) = 9.94$, $p = .006$, $\eta_p^2 = 0.369$). Since we maintained the tone signal to noise ratio (SNR) across blocks, the block number effect must be caused by participants' decreased effort or fatigue due to the repetitive trial structure in this tone detection task. The main effect of Temporal location is not significant

($F(2,34) = 3.26$, $p = .051$, $\eta_p^2 = 0.166$), but the interaction between Temporal location and Block number was significant ($F(4,68) = 3.80$, $p = .008$, $\eta_p^2 = 0.183$). We then tested the main effect of Temporal location in each block (one-way rmANOVA) and found a significant main effect for Temporal location in the second block ($F(2,34) = 4.97$, $p = .020$, $\eta_p^2 = 0.226$) and in the third block ($F(2,34) = 7.01$, $p = .009$, $\eta_p^2 = 0.292$), but not in the first block ($F(2,34) = 2.94$, $p = .066$, $\eta_p^2 = 0.148$). To control false positive rate, adjusted False Discovery Rate (FDR) was used.

The criteria calculated for the three temporal locations shared the same no-tone trials and the differences between them was dominated by the hit rate in the tone trials. Instead of calculating one criterion for each temporal location, we averaged the hit rates over the three temporal locations in the tone trials and calculated a criterion for each block. We conducted a one-way rmANOVA on criteria with Block number as the main factor and did not find the main effect to be significant ($F(2,34) = 3.06$, $p = .056$, $\eta_p^2 = 0.105$).

We conducted a two-way rmANOVA on the raw reaction time with Trial condition (three different temporal locations in the tone trials and one no-tone condition) and Block number as the main factors. The main effect of Trial condition is significant ($F(3,51) = 28.68$, $p < .001$, $\eta_p^2 = 0.636$); no significance was found for the main effect of Block number ($F(2,34) = 1.65$, $p = .207$, $\eta_p^2 = 0.089$) nor for the interaction ($F(6,102) = .92$, $p = .485$, $\eta_p^2 = 0.051$). The tone detected in the early locations might give the participants more time to prepare for button responses, so the significant main effect of Trial condition was observed.

(*D*) ERPs to the target tone averaged across three blocks. We selected the channel of Cz for the statistical analyses. Through a permutation test, we identified two significant temporal regions showing a main effect of Temporal location ($p < 0.05$): region of interest 1 (ROI 1) from 0.18 s to 0.26 s and ROI 2 from 0.35 s to 0.46 s (left plot). We then averaged the ERP amplitudes within each temporal region to compare ERP magnitude between temporal locations. In both ROIs, we found that the absolute magnitude of ERPs to the target tone significantly increase with the temporal location with the ERP in the third tone location larger than the first and the second tone locations and the ERP in the second tone location larger than the first tone location ($p < 0.05$; adjusted FDR correction was applied).

(*E*) ERPs to the target tone in each block. We applied the ROIs 1 and 2 derived from (*D*) and examined the effect of the block number. We conducted a two-way rmANOVA with Temporal location and Block number as the main factors in each ROI. In ROI 1, the main effect of Block number is significant ($F(2,52) = 7.92$, $p = .001$, $\eta_p^2 = 0.233$) with the absolute ERP magnitude of ROI 1 decreasing with the block number - the ERP in the third block is smaller than the first and the second blocks and the ERP in the second block is smaller than the first block ($p < 0.05$; adjusted FDR correction was applied). We did not find a significant main effect of Block number in ROI 2 ($F(2,52) = 2.54$, $p = .089$, $\eta_p^2 = 0.089$) nor significant interactions in both ROIs ($p > 0.05$).

The shaded areas of color and the error bars represent ±1 standard error of the mean over participants.

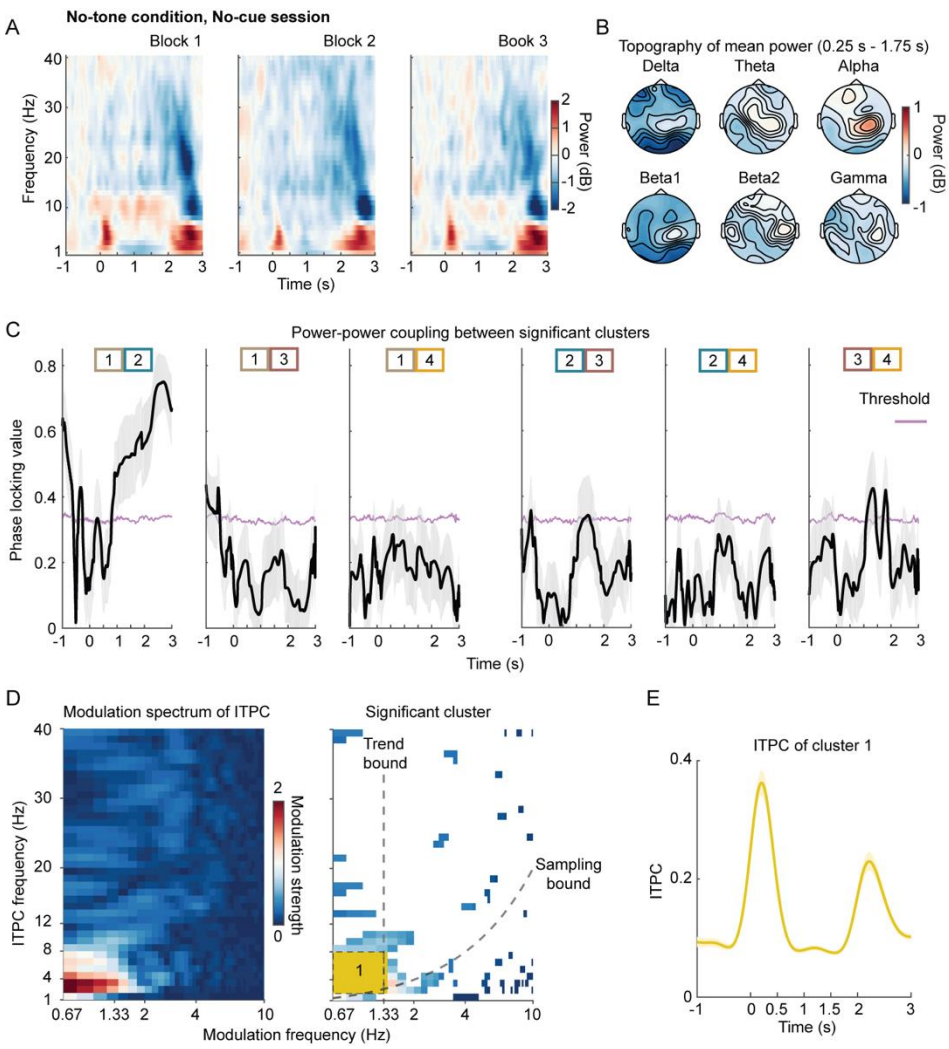

Fig. S2. Traditional power analyses in the no-cue session, power-power coupling, and modulation spectrum analyses on ITPC. Related to Fig. 2. (*A*) Induced power of no-tone trials in the no-cue session. From left to right, the plots represent spectrograms of induced power in the block 1, 2, and 3, respectively. (*B*) Topographies of routinely defined power frequency bands: delta (1-3 Hz), theta (4-7 Hz), alpha (8-12 Hz), beta1 (13-20 Hz), beta2 (21-30 Hz), and gamma (31-40 Hz). The induced power in each frequency band was averaged over three blocks from 0.25 s to 1.75 s after noise onset. (*C*) Power-power coupling between significant clusters shown in Fig. 2B&D. We calculated phase-locking values (solid black line in each block) and derived corresponding thresholds of a one-sided alpha level of 0.01 using surrogate tests (thin pink lines). The power dynamics of Cluster 1 and 2 is significantly coupled after noise onset. We also found significant couplings after noise onset between Cluster 2 and 3 and between Cluster 3 and 4, which suggests that Cluster 2, 3, and 4 together contribute a common neural process. (*D*) Modulation spectrum of ITPC. We conducted the modulation spectrum analyses on the neural phase measurement – inter-trial phase coherence (ITPC), but only found one cluster showing robust modulation strength below Trend bound. (*E*) Dynamics of ITPC Cluster 1. We filtered the ITPC according to the cluster defined in (*D*) and show the dynamics of ITPC. It can be seen that the ITPC cluster reflects increased phase coherences to the noise onset and offset, without showing signatures of coding local temporal locations in the white noise or of monitoring the elapsed time of the noise. The shaded areas of color represent ±1 standard error of the mean over participants.

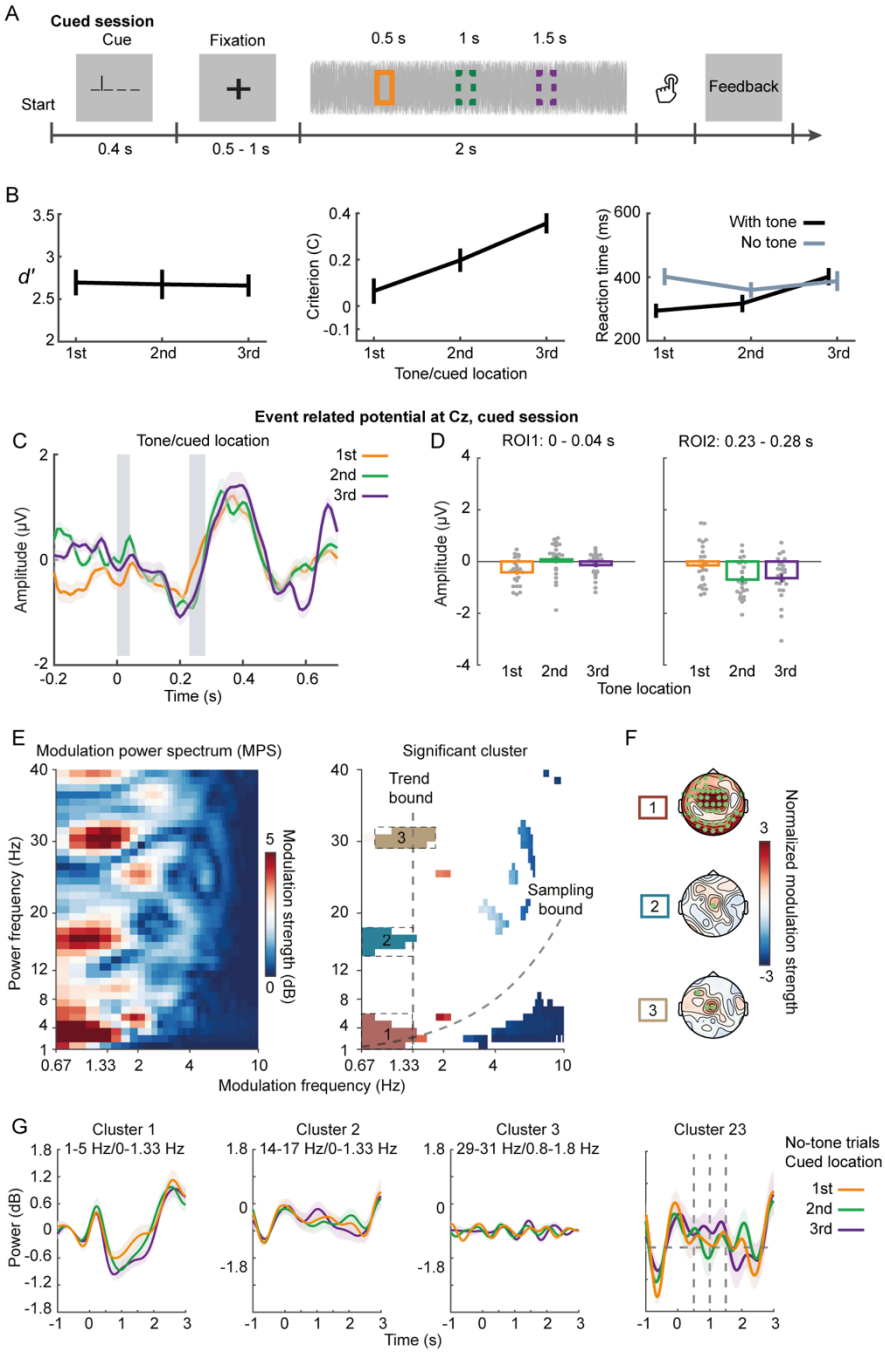

Fig. S3. Experimental paradigm and results in the cued session. Related to Fig. 1&2. (*A*) Experimental paradigm in the cued session. The trial structure here is the same as in the no-cue session, but the visual feedback revealing where the target tone is in the no-cue session was moved to the beginning of each trial and served as a visual cue to direct the participants' attention to a specific temporal location. The visual cue was valid in half of the trials (a tone appeared in the cued location) and invalid in the other half (no tone appeared). (*B*) Behavioral results. From left to right, the plots show d-prime values, criteria, and reaction time, respectively. After conducting a one-way rmANOVA, we found that the main effect of Temporal location on d-prime values was not significant ($F(2,32) = 0.38$, $p = .687$, $\eta_p^2 = 0.023$), which differs from the finding in the no-cue session. On criteria, the main effect of Temporal location is significant ($F(2,32) = 16.56$, $p < .001$, $\eta_p^2 = 0.399$) with an increasing linear trend ($F(1,25) = 25.67$, $p < .001$, $\eta_p^2 = 0.507$), suggesting that the participants became

conservative and miss rates increased with the temporal location. Since the likelihood of a tone to appear increases with the elapsed time, participants should have become more aggressive and the criteria should have decreased, but we observed the opposite. This suggests that the participants may estimate the likelihood that a trial is a tone trial (Fig. S1B) and become more conservative, as the likelihood of tone trial indeed decreases as the time is passing. (*C*) ERPs to the target tone. Following the same procedure in Fig. S1C, we identified two ROIs showing a significant main effect of Temporal location in the tone trials; we did not find any ROIs in the no-tone trials. The post-hoc tests in both ROIs showed a significant difference between ERPs in the first temporal location and the other two locations ($p < 0.05$; adjusted FDR was applied), which was probably caused by the baseline differences – the first temporal location was affected by the ERP to noise onset whereas the other two locations were not. Note that the main effect of Temporal location shown in Fig. S1C was not observed here, suggesting that the participants might employ different cognitive processes to detect a tone in the cued session. (*E*) Raw MPS (left plot) and significant clusters (right plot) on the induced power of no-tone trials. We followed the same procedure shown in Fig. 2B and derived MPS on the no-tone trials in the cued session. In the right plot, three clusters were selected after the surrogate test. Cluster 1 here aligns with Cluster 1 in Fig. 2B, but Cluster 2 and 3 here are newly identified and different from the clusters in Fig. 2B, which suggests that the participants employed a different cognitive strategy to detect the target tone in the cued session and that Cluster 2, 3, and 4 identified in the no-cue session were not simply a result of the trial structure. Although around 2 Hz modulation frequency we observed significant modulation bins, they were not robustly shown in every surrogate test (see *Methods*) and had low modulation strength. Consequently, we did not identify them as significant clusters for further analyses. (*F*) Topographies of significant clusters. The green dots denote significant EEG channels. (*G*) Power dynamics of significant clusters. Cluster 2 and 3 do not clearly correspond to the temporal locations; after we combined Cluster 2 and 3 together, we still did not observe a neural signature coding the temporal locations. We show this set of results as a comparison with the no-cue session to demonstrate that the clusters observed in the no-cue session were caused by participants actively anticipating the temporal locations but were not simply caused by the trial structure. Hence, we did not further explore those findings. The shaded areas of color and the error bars represent ±1 standard error of the mean over participants.

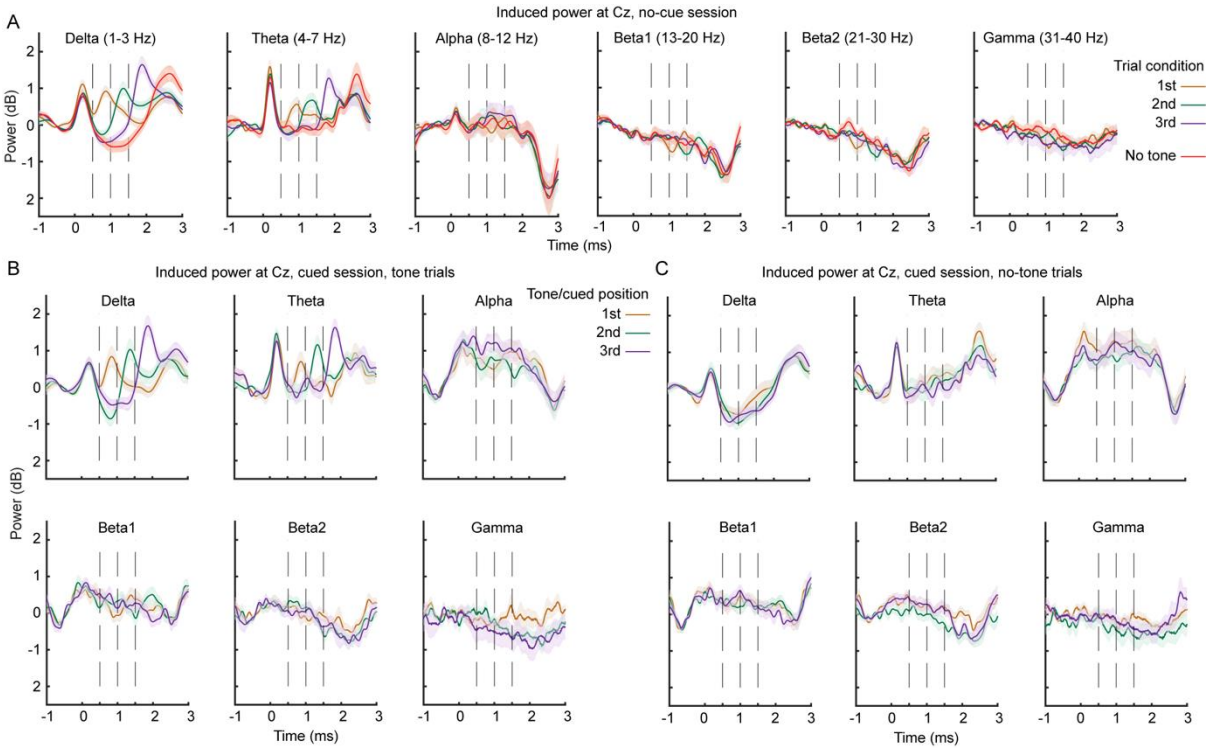

Fig. S4. Dynamics of induced power in the routinely defined frequency band. Related to Fig. 2&3. (*A*) Induced power in the no-cue session. Although the beta band was commonly defined as from 13 Hz to 30 Hz, we deliberately divided it into two frequency bands – beta1 and beta2 bands, after we derived the clusters shown in Fig. 2B. In the plots, the neural signature coding the local temporal locations is not seen since frequency bands were defined a priori without knowing which power frequencies were related to the experimental effects. This demonstrates the advantage of our data-driven method shown in Fig. 2B. (*B*) Induced power on the tone trials in the cued session. (*C*) Induced power on the no-tone trials in the cued session. Similar to (*A*), no clear patterns were shown to relate to the local temporal locations.

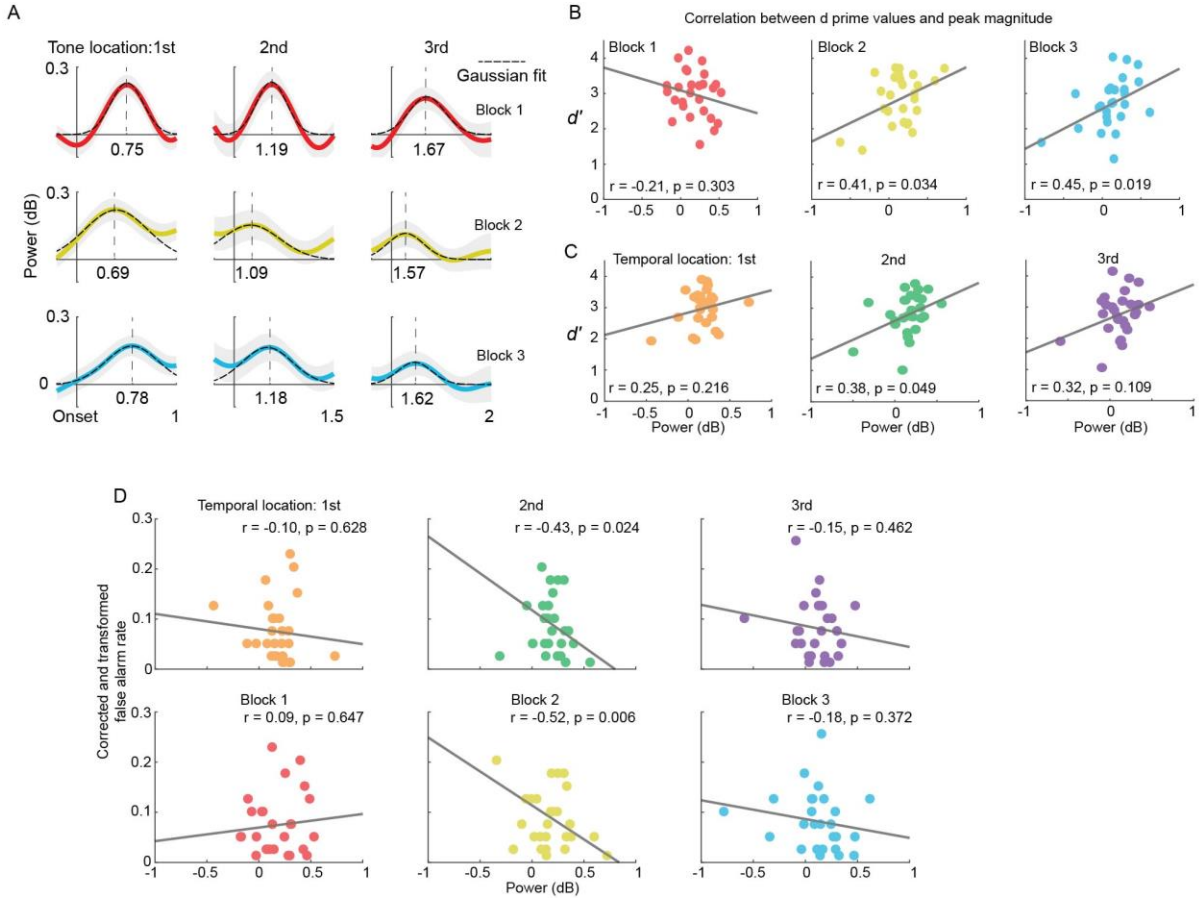

Fig. S5. Gaussian fits to neural peaks of Corrected cluster 234 in each block and correlations between behavioral measurements and neural data. Related to Fig. 3. (*A*) Gaussian fits to each neural peak of Corrected cluster 234 in each block. From the Gaussian fits, we derived peak latencies and peak widths (SD of the Gaussian fits). (*B*) Correlations between d-prime values and magnitude of neural peaks per block. The correlation in the first block was negative (not significant) whereas the correlations were positive and significant in the other two blocks. It is possible that, in the first block, the participants actively anticipated each temporal location as taught by the visual feedbacks but did not exploit the anticipated temporal locations for tone detection. By the second and the third blocks, the participants learned to exploit the timing information of temporal locations for tone detection. It suggests that the participants first learned the temporal locations of the target tone and then used such temporal information for a specific task. (*C*) Correlations between d-prime values and magnitude of neural peaks per temporal location. All the correlations are positive, which suggests that the neural coding of local temporal locations observed in the no-tone trials is indeed relevant to tone detection. (*D*) Correlations between false alarm rate and magnitude of neural peaks. The false alarm rates were transformed using the Fisher-Z transformation before the correlations were calculated. One potential explanation for the emergence of Corrected cluster 234 is that participants may wrongly consider a tone to be appeared in the no-tone trials and such 'hallucination' induced the neural peaks after each temporal location. However, the correlations here are mostly negative or close to zero, suggesting that this explanation based on the false alarm rate did not account for the emergence of Corrected cluster 234.