frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 03 September 2020
doi: 10.3389/fncom.2020.00079

Excitation-Inhibition Balanced Neural Networks for Fast Signal Detection

Gengshuo Tian[1], *Shangyang Li*[1,2], *Tiejun Huang*[1] *and Si Wu*[1,2]*

[1] School of Electronics Engineering and Computer Science, Peking University, Beijing, China, [2] IDG/McGovern Institute for Brain Research, Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China

Excitation-inhibition (E-I) balanced neural networks are a classic model for modeling neural activities and functions in the cortex. The present study investigates the potential application of E-I balanced neural networks for fast signal detection in brain-inspired computation. We first theoretically analyze the response property of an E-I balanced network, and find that the asynchronous firing state of the network generates an optimal noise structure enabling the network to track input changes rapidly. We then extend the homogeneous connectivity of an E-I balanced neural network to include local neuronal connections, so that the network can still achieve fast response and meanwhile maintain spatial information in the face of spatially heterogeneous signal. Finally, we carry out simulations to demonstrate that our model works well.

Keywords: E-I balanced network, optimal noise structure, Fokker-Planck equation, fast tracking, asynchronous state

1. INTRODUCTION

To survive in natural environments, animals have developed, through millions of years evolution, the ability to process sensory inputs rapidly. For instance, studies have shown that human subjects can perform complex visual analyses within 150 ms (Thorpe et al., 1996), and the response latency of neurons in the visual cortex of monkeys is as short as tens of milliseconds (Raiguel et al., 1999; Sugase et al., 1999).

Meanwhile, many artificial engineering systems have high demands for real-time processing of rapidly varying signals. This is exemplified by the recently developed Spike Camera (Dong et al., 2017), which has a sampling rate of up to 40,000 frames per second (fps), far surpassing conventional cameras' 60 fps. This allows it to capture high-speed objects and their textual details, which can be used on real-time motion detection, tracking, and recognition if we have the appropriate algorithms and computing platforms. However, the processing speed of traditional algorithms often cannot meet such demands.

The balance of excitation and inhibition is a general property of neural systems. The excitation-inhibition (E-I) balanced neural network was first proposed to explain the irregular firing of cortical neurons widely observed in the cortex (Softky and Koch, 1993; Shadlen and Newsome, 1994), and was later confirmed by a large amount of experimental data (Haider et al., 2006; Okun and Lampl, 2008; Dorrn et al., 2010; Graupner and Reyes, 2013). Theoretical studies have found that the asynchronous irregular firing state spontaneously emerges in a network of excitatory and inhibitory neurons with random connections satisfying some very loose balancing conditions (van Vreeswijk and Sompolinsky, 1996; van Vreeswijk and Sompolinsky, 1998; Renart et al., 2010). The effects of this chaotic state on optimal coding (Denève and Machens, 2016), working memory (Lim and Goldman, 2014), and neuronal tuning (Hansel and van Vreeswijk, 2012), as well as its coexistence with attractor dynamics (Litwin-Kumar and Doiron, 2012) have been widely studied.

OPEN ACCESS

Edited by:
Pei-Ji Liang,
Shanghai Jiao Tong University, China

Reviewed by:
Robert Rosenbaum,
University of Notre Dame, United States
Lianchun Yu,
Lanzhou University, China

***Correspondence:**
Si Wu
siwu@pku.edu.cn

Received: *27 June 2020*
Accepted: *27 July 2020*
Published: *03 September 2020*

Citation:
Tian G, Li S, Huang T and Wu S (2020) Excitation-Inhibition Balanced Neural Networks for Fast Signal Detection. Front. Comput. Neurosci. 14:79. doi: 10.3389/fncom.2020.00079

In the present study, we focus on the fast tracking ability of E-I balanced networks, where the population firing rate of the network is proportional to the input amplitude and tracks input changes rapidly (van Vreeswijk and Sompolinsky, 1996; Renart et al., 2010), and investigate how E-I balanced neural networks can be used for fast signal detection in brain-inspired computation. Neuromorphic computing, which mimics the structures and computational principles of the neural system, is receiving increasing attention in artificial intelligence (AI), as it has the potential to overcome the von Neumann bottleneck in modern computers that limits their processing speed (Indiveri and Liu, 2015). The fast response property of the E-I balanced network makes it a naturally compatible candidate to be implemented in neuromorphic systems to achieve rapid information processing.

In the following sections, we show that the asynchronous firing state of the network generates an optimal noise structure which enables the network to track input changes rapidly. We then extend the homogeneous connectivity of the classical E-I balanced neural network to include local neuronal connections, so that the network can achieve fast response and meanwhile maintain the spatial information when presented with spatially heterogeneous signals. Finally, we carry out simulations to demonstrate the performance of our model.

2. FAST RESPONSE OF A HOMOGENEOUS E-I BALANCED NETWORK

To illustrate the mechanism of the fast response property, we first investigate a homogeneously connected E-I balanced network.

2.1. Intuition on the Mechanism of Fast Response

The fast response property of an E-I balanced network is at the population level. To understand this, let us consider a non-leaky linear integrate-and-fire neuron, whose dynamics is given by

$$\tau \frac{\mathrm{d}v}{\mathrm{d}t} = I, \tag{1}$$

where τ is the integration time constant of the neuron, v the membrane potential, and I the input current. When v reaches the threshold θ, the neuron generates an action potential, and v is reset to the reset potential v_0. Thus, for a constant input I_0, the time it takes for a neuron to generate a spike starting from v_0 is

$$T = \tau \frac{\theta - v_0}{I_0}.$$

It can be seen that the response time of a single neuron is limited by τ (**Figure 1A**).

However, when a neural population receives a signal, if the noise in the system keeps membrane potentials of different neurons at different levels, there will always be a few neurons whose potentials are near the threshold that can quickly respond to input changes. In such a case, the network as a whole can respond to input changes very fast, whose reaction time is only restricted by insurmountable factors such as axonal conduction delays, rather than the membrane time constant τ of individual neurons (**Figure 1B**). The key of this mechanism is to prevent synchronous firing of neurons and maintain a stable distribution of membrane potentials in the neural population, and asynchronous firing happens to be one of the hallmarks of an E-I balanced network (Renart et al., 2010), which we shall discuss in more detail below.

2.2. The Balancing Condition

We first present the conditions for maintaining an E-I balanced neural network and the stationary population firing rates under those conditions in the large N limit, where N is the number of neurons (van Vreeswijk and Sompolinsky, 1998; Rosenbaum et al., 2017). Consider a network of size N, with $N_E = q_E N$ being excitatory and $N_I = q_I N$ inhibitory, where $q_E + q_I = 1$. The input current received by neuron i in population a ($a = E$ being excitatory and $a = I$ being inhibitory) can be written as

$$I_i^a(t) = F_i^a(t) + R_i^a(t), \quad a = E, I, \tag{2}$$

where F_i^a is the feedforward (i.e., external) input, and R_i^a is the recurrent input from other neurons in the network with the form

$$R_i^a(t) = \sum_{b=E,I} \sum_j J_{ij}^{ab} \sum_k \frac{1}{\tau_{b,s}} e^{-(t-t_{j,k})/\tau_{b,s}}, \quad a = E, I, \tag{3}$$

where j indexes presynaptic neurons, $\tau_{b,s}$ is the synaptic time constant of the presynaptic population b, and $t_{j,k}$ is the spike time of the k'th spike of neuron j.

Since in both the cortex and industrial applications, the number of neurons in a network is large, we may examine the balanced network in the $N \to \infty$ limit. Expressing the relevant quantities in orders of N can help elucidate the mechanism. Neurons in the network are connected randomly, with the connection probability determined solely by the neuron types. The probability that neuron j in population b connects to neuron i in population a is p_{ab} for all i, j. Note that here p_{ab} is constant, and does not tend to 0 as $N \to \infty$. This regime is usually referred to as dense connectivity, in contrast to sparse connectivity where the number of presynaptic neurons for each postsynaptic neuron is kept constant as $N \to \infty$ (van Vreeswijk and Sompolinsky, 1996; Brunel, 2000). If a connection exists, its strength is set to be $J_{ij}^{ab} = j_{ab}/\sqrt{N}$; otherwise $J_{ij}^{ab} = 0$. Here $j_{ab} \sim \mathcal{O}(1)$. ($\mathcal{O}$ denotes scaling with respect to $N \to \infty$ throughout this paper.) This scaling is a hallmark of balanced networks. Note that in some earlier works, especially those that employ a sparse connectivity regime (van Vreeswijk and Sompolinsky, 1996), this scaling is often written as $J \sim \mathcal{O}(\sqrt{K_{ab}})$, where K_{ab} is the average number of presynaptic inputs from population b for a neuron in population a. Here, since we have $K_{ab} = p_{ab}N$ and $p_{ab} \sim \mathcal{O}(1)$, these two scalings are essentially the same.

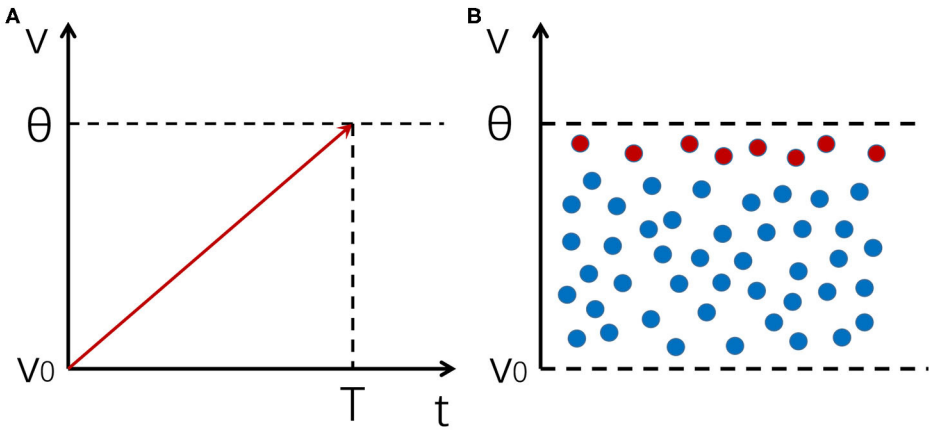

FIGURE 1 | An illustration of the mechanism of fast response for a neural population. **(A)** The integration and firing process of a neuron receiving a noiseless input. The integration time is constrained by the membrane time constant. **(B)** A distribution of membrane potentials across a neural population enables it to respond to input changes rapidly. Red dots represent neurons whose potentials are close to the firing threshold, which are the first ones to respond to input changes.

Using the mean-field approximation, the time- and population-averaged input current received by a neuron in population a can be written as,

$$\overline{I_a} = \overline{F_a} + \overline{R_a} = \sqrt{N}(f_a\mu_0 + w_{aE}r_E + w_{aI}r_I), \quad a = E, I, \quad (4)$$

where r_b is the mean firing rate of population b, $b = E, I$, and $w_{ab} = p_{ab}j_{ab}q_b \sim \mathcal{O}(1)$. Here, we have written $\overline{F_a}$ as $\overline{F_a} = \sqrt{N}f_a\mu_0$, where $f_a, \mu_0 \sim \mathcal{O}(1)$, because if we notice that long-distance projections are mainly excitatory, and assume that the feedforward inputs originated from another neural population of size $\mathcal{O}(N)$ and that the feedforward synaptic strength is also of order $\mathcal{O}(1/\sqrt{N})$, then $\overline{F_a} \sim \mathcal{O}(\sqrt{N})$ is a natural consequence. This is exactly the case in the Spike Camera data scenario that we shall examine later in section 3.

Therefore, to keep I (and thus r) bounded when $N \to \infty$, we must have

$$w_{aE}r_E + w_{aI}r_I + f_a\mu_0 \sim \mathcal{O}(\frac{1}{\sqrt{N}}), \quad a = E, I.$$

Letting $N \to \infty$, we get approximate firing rates in the large N limit

$$\begin{aligned} \lim_{N\to\infty} r_E &= \frac{f_E w_{II} - f_I w_{EI}}{w_{EI}w_{IE} - w_{EE}w_{II}}\mu_0, \\ \lim_{N\to\infty} r_I &= \frac{f_I w_{EE} - f_E w_{IE}}{w_{EI}w_{IE} - w_{EE}w_{II}}\mu_0. \end{aligned} \quad (5)$$

To keep the above limits positive and yield a stable solution, it is necessary and sufficient to let (van Vreeswijk and Sompolinsky, 1998)

$$\frac{f_E}{f_I} > \frac{w_{EI}}{w_{II}} > \frac{w_{EE}}{w_{IE}}.$$

This is the condition for the balanced firing state.

It is worth noting that whatever the neuronal transfer function is, the population firing rate in the large N limit is always linearly proportional to μ_0. That is, Equation (5) always holds. This is a direct result of Equation (4), where the total input current is the linear sum of the three $\mathcal{O}(\sqrt{N})$ order terms. The balanced firing state is a stable solution dynamically formed by the network (van Vreeswijk and Sompolinsky, 1998; Renart et al., 2010), and therefore requires no fine tuning of parameters such as j_{ab}, which is different from some other models that also try to recreate the asynchronous irregular firing state (e.g., Brunel, 2000).

It should be pointed out that Equation (5) only gives the $\mathcal{O}(1)$ order term of r_a. To satisfy the specific transfer function of neurons while maintaining the balance of the $\mathcal{O}(1)$ order term, the firing rates are adjusted by an $\mathcal{O}(1/\sqrt{N})$ term, which results in a $\mathcal{O}(1)$ order correction to I (Equation 4). We will come back to this in the specific case presented in the next section.

2.3. The Mechanism of Fast Response

As previously mentioned, the asynchronous firing of neurons is the key for fast response of the network. When the balancing conditions presented in the previous section are met, the network can achieve asynchronous irregular firing (Renart et al., 2010). We next use a network of non-leaky linear integrate and fire neurons to study the mechanism of fast response in more detail. Notably, this simple neuron model has already been implemented in a neuromorphic system (Fusi and Mattia, 1999). While not biologically realistic, this model captures the key characteristics of integrate-and-fire neurons crucial for neuromorphic computing.

The neuronal dynamics is given by Equation (1). For simplicity, let $v_0 = 0$. It can be easily seen that the transfer function of this neuron is threshold-linear, i.e.,

$$r = \begin{cases} \dfrac{\bar{I}}{\theta\tau}, & \bar{I} \geqslant 0, \\ 0, & \bar{I} < 0. \end{cases} \quad (6)$$

Substituting this into Equation (4) yields the population firing rates of excitatory and inhibitory neurons

$$r_E = \frac{(f_E w_{II} - f_I w_{EI}) - \frac{1}{\sqrt{N}} f_E \theta \tau_I}{(w_{EI} w_{IE} - w_{EE} w_{II}) + \frac{1}{\sqrt{N}} \theta (w_{EE}\tau_I + w_{II}\tau_E) - \frac{1}{N}\theta^2 \tau_I \tau_E} \mu_0,$$
$$r_I = \frac{(f_I w_{EE} - f_E w_{IE}) - \frac{1}{\sqrt{N}} f_I \theta \tau_E}{(w_{EI} w_{IE} - w_{EE} w_{II}) + \frac{1}{\sqrt{N}} \theta (w_{EE}\tau_I + w_{II}\tau_E) - \frac{1}{N}\theta^2 \tau_I \tau_E} \mu_0. \tag{7}$$

Comparing the above result with Equation (5), we can see that they are indeed $\mathcal{O}(1/\sqrt{N})$ order corrections to the $N \to \infty$ limit, as stated at the end of the last section. Note that the firing rates still linearly encode the external input, which is a result of the threshold-linear transfer function. We also check that even when the external input is small or the number of neurons is not large, the linear encoding property still holds, which expands the dynamic range of the network. However, for other non-linear neuron models, this linear encoding property may not hold.

Equation (7) is derived from the mean-field approximation, that is, it is the result of averaging over time and neurons when the system reaches a stable state. To study how the instantaneous firing rate of the population changes with time when external input changes, we need more detailed analysis. We shall use the Fokker-Planck equation (Risken, 1996) to study the membrane potential distribution $p_a(v, t)$ (Brunel and Hakim, 1999; Fusi and Mattia, 1999; Brunel, 2000; Huang et al., 2011).

First, we examine the input received by a single neuron as described in Equation (2). We consider an external input signal with additive white Gaussian noise

$$F_i^a(t) = \sqrt{N} f_a \mu_F(t) + \sigma_{aF}(t)\xi_i^{aF}(t), \quad a = E, I, \quad i = 1, \cdots, N_a, \tag{8}$$

where ξ_i^{aF} is a Gaussian white noise of magnitude 1 that is independent across neurons. Note that the signal mean is of order $\mathcal{O}(\sqrt{N})$, while the variance is of order $\mathcal{O}(1)$. This is because if we continue to use the settings considered before, and view the feedforward input as coming from Poisson spike trains generated by $\mathcal{O}(N)$ neurons firing at rates of order $\mathcal{O}(1)$, and transmitted through synapses with the strength of order $\mathcal{O}(1/\sqrt{N})$, then the resulting input's variance is the sum of $\mathcal{O}(N)$ number of terms with the same order as the square of synaptic strengths ($\mathcal{O}(1/N)$), and is therefore of order $\mathcal{O}(1)$. This characteristic is also present in the later analysis of recurrent inputs.

Next, we examine the recurrent inputs. When the network enters the balanced state, since the neurons fire asynchronously (Renart et al., 2010), and the effect of each spike is small, we could use Gaussian white noise to approximate the variations of recurrent inputs, and rewrite the second term in Equation (2) as (Brunel, 2000)

$$R_i^a(t) = \sqrt{N}\mu_{aR}(t) + \sigma_{aR}(t)\xi_i^{aR}(t), \quad a = E, I, \tag{9}$$

where

$$\mu_{aR} = w_{aE} r_E + w_{aI} r_I, \quad \sigma_{aR}^2 = j_{aE} w_{aE} r_E + j_{aI} w_{aI} r_I, \tag{10}$$

and ξ_i^{aR} is Gaussian noise of magnitude 1. The terms ξ_i^{aR} and ξ_i^{aF} are independent due to the asynchronous firing state, and can therefore be merged into one noise source. Thus, we transform Equation (2) into

$$I_i^a(t) = \mu_a(t) + \sigma_a(t)\xi_i^a(t), \quad a = E, I, \tag{11}$$

where

$$\mu_a = \sqrt{N}(w_{aE} r_E + w_{aI} r_I + f_a \mu_F)$$
$$\sigma_a^2 = j_{aE} w_{aE} r_E + j_{aI} w_{aI} r_I + \sigma_{aF}^2, \tag{12}$$

and ξ_i^a is Gaussian white noise of magnitude 1. Also note that the mean of the signal is consistent with Equation (4), and the mean and variance are both of order $\mathcal{O}(1)$.

Since the balanced state implies asynchronous firing (Renart et al., 2010), the noise ξ_i^a of different neurons can be seen as independent. Then, the excitatory (inhibitory) population can be viewed as i.i.d. samples of the same random process. The membrane potential distribution of population a, $p_a(v, t)$, can thus be derived from Equation (11). We obtain the Fokker-Planck equation (Brunel, 2000; Huang et al., 2011)

$$\tau_a \frac{\partial p_a(v,t)}{\partial t} = -\mu_a \frac{\partial p_a(v,t)}{\partial v} + \frac{\sigma_a^2}{2\tau_a} \frac{\partial^2 p_a(v,t)}{\partial v^2}, \quad a = E, I. \tag{13}$$

A few boundary conditions can be naturally imposed (Brunel and Hakim, 1999; Brunel, 2000):

$$p_a(v, t) = 0, \quad \forall v \geqslant \theta. \tag{14}$$
$$p_a(0^-, t) = p_a(0^+, t), \tag{15}$$
$$\frac{\partial p_a(0^+,t)}{\partial v} - \frac{\partial p_a(0^-,t)}{\partial v} = \frac{\partial p_a(\theta,t)}{\partial v}. \tag{16}$$
$$\int_{-\infty}^{\theta} p_a(v, t)\mathrm{d}v = 1. \tag{17}$$

In Equation (13), letting $\partial p_a/\partial t = 0$, and using the above boundary conditions, we get the stationary solution

$$p_{a0}(v) = \begin{cases} \frac{1}{\theta}[1 - \exp(-2\tau_a\beta_a)] \exp\left(\frac{2\tau_a v}{\beta_a}\right), & v < 0 \\ \frac{1}{\theta}\left[1 - \exp\left(\frac{-2\tau_a(\theta - v)}{\beta_a}\right)\right], & 0 \leqslant v \leqslant \theta \\ 0, & v > \theta \end{cases} \tag{18}$$

where $\beta_a := \sigma_a^2/\mu_a$ is the variance-to-mean ratio (VMR). This result is confirmed by simulations (**Figure 2A**).

The population firing rate, i.e., the flux at θ, is

$$r_a = -\frac{\sigma_a^2}{2\tau_a^2} \frac{\partial p_{a0}(v)}{\partial v}\bigg|_\theta = \frac{\mu_a}{\theta\tau_a}. \tag{19}$$

which is consistent with Equation (6).

It can be seen from Equation (18) that the membrane potential distribution is determined by the VMR β_a. The ideal noise structure is thus obtained when VMR stays constant

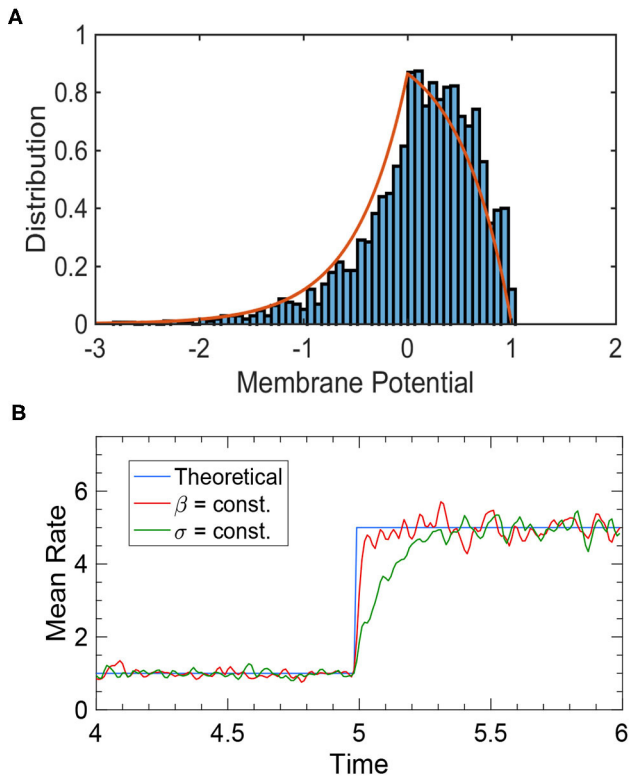

FIGURE 2 | Simulation results of an uncoupled neural population. **(A)** The membrane potential distribution of a neural population receiving independent white noise-corrupted signals with a constant VMR of 1. The red curve is the theoretical prediction given by Equation (18), and the blue histogram is the actual simulation result. **(B)** The tracking performance of a neural population depends on the input noise structure. The blue curve is the theoretical prediction of steady-state firing rate given by Equation (19). The red curve is the network performance when the VMR is constant ($\beta = 1$), which tracks the input change almost instantaneously. The green curve is the network performance when the noise variance, rather than the VMR, is constant ($\sigma \equiv 1$), where a significant delay is present. Other parameters are: $N = 2,500$, $\tau = 1$, $\theta = 1$, and μ changing from 1 to 5 at time $t = 5$.

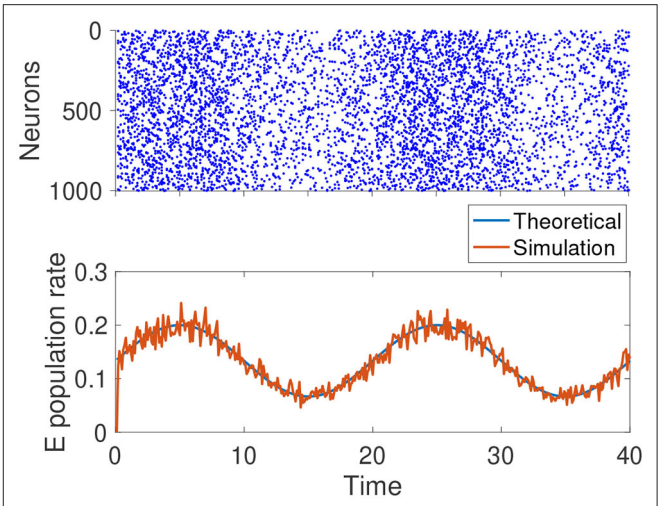

FIGURE 3 | Simulation results of a homogeneous E-I balanced network tracking a time-varying input. The network receives a sinusoidal input centered at $\mu_F = 0.1$ with an amplitude of 0.05. $\sigma_{aF}^2/\mu_F = 0.1$ remains constant. The blue curve is the theoretic prediction given by Equation (7). The red curve is the instantaneous average firing rate of excitatory neurons. The parameters are: $N = 1 \times 10^4, q_I = 0.2, p_{ab} = 0.25, \theta = 15, \tau_E = 15, \tau_I = 10, \tau_{E,s} = 6, \tau_{I,s} = 5, f_E = 3, f_I = 2, j_{EE} = 0.25, j_{EI} = -1, j_{IE} = 0.4, j_{II} = -1$.

(Huang et al., 2011), because it ensures that when the external input μ_F changes, the system remains in a stationary state where Equation (19), and thus Equation (7), always holds. In this way, the population rate can track input changes instantaneously and linearly encode μ_F at all times. **Figure 2B** illustrates how the response time of the population rate is determined by input noise structure.

From Equations (12) and (19), we know that when the network is at the stationary state,

$$\beta_a = \frac{\sigma_a^2}{\mu_a} = \frac{j_{aE} w_{aE} r_E + j_{aI} w_{aI} r_I + \sigma_{aF}^2}{\theta \tau_a r_a}, \quad a = E, I.$$

From Equation (7), we know $r_E, r_I \propto \mu_F$. For the σ_{aF}^2 term, if we continue to assume that the external input comes from the Poisson spike trains of another population of neurons, and the changes in μ_F are due to the firing rate of that population, then we have $\sigma_{aF}^2 \propto \mu_F$. Thus, when μ_{aF} changes, β_a remains constant. This is the ideal noise structure, and the population rate of the network can track the external input instantaneously. In reality, the ideal noise structure can only be approximately satisfied, but the tracking speed of the network is still reasonably fast, as confirmed by **Figure 3**.

It should be pointed out that the neuron model we used in this section does not have a lower bound to its membrane potential. In real applications, a reflecting barrier can be imposed at the reset potential v_0 (Fusi and Mattia, 1999). We verify that this does not affect our main results. The neuron model used in the following sections has a reflecting barrier.

3. PROCESSING SPATIALLY HETEROGENEOUS INPUT WITH LOCAL CONNECTIVITY

In the above, we have studied an E-I balanced neural network with homogeneous connectivity, which is able to track input changes rapidly. However, when the external input is spatially heterogeneous, that is, when different neurons receive inputs of different magnitudes, this homogeneous connectivity generates statistically equivalent recurrent inputs for each neuron that cannot balance the external inputs. The same R_a^i's cannot balance different F_a^i's, causing neurons to receive inputs of order $\mathcal{O}(\sqrt{N})$ and fire pathologically. In addition, the random long-range connections between neurons spread out local activities to the entire network, which blurs the spatial location of inputs. In applications, however, we often need to know not only when the signal occurs but also where it occurs. To solve this problem, we need to introduce local connectivity in the network. Previous

studies have shown that if appropriate local connectivity is included, the network can maintain the balanced firing state as well as retain the spatial information of the input (Rosenbaum and Doiron, 2014; Rosenbaum et al., 2017), which enables the network to achieve both fast tracking and spatial location encoding. Below, we briefly introduce the balancing conditions and the response property of an E-I balanced neural network with local connectivity.

Here, each neuron is assigned a location (x, y) on the plane, and local connectivity is achieved by a connection probability that decays with the spatial distance between pairs of neurons instead of being homogeneous as in the previous sections, so that neurons closer to each other have higher probabilities to connect with each other. Specifically, the probability of a connection between neurons i and j follows

$$\mathbb{P}(j \text{ connects to } i) \propto G_b\left(d_{ij}\right), \tag{20}$$

where G_b is a 2-dimensional Gaussian shaped function whose spatial spread is determined by the presynaptic population b, and d_{ij} is the distance between the neurons.

Similar to Equation (4), we again utilize the mean-field approximation. Only this time, we do not average over the entire population, but rather approximate the neural activity of population a near location $\mathbf{x}$ with the neural field

$$\overline{I_a}(\mathbf{x}) = \overline{F_a}(\mathbf{x}) + \overline{R_a}(\mathbf{x}) = \sqrt{N}[f_a(\mathbf{x}) + w_{aE} * r_E(\mathbf{x}) - w_{aI} * r_I(\mathbf{x})], \quad a = E, I, \tag{21}$$

where the feedforward input $\overline{F}_a(\mathbf{x}) = \sqrt{N} f_a(\mathbf{x})$, $w_{ab}(\mathbf{x}) = q_b j_{ab} p_{ab} G_b(\mathbf{x})$ is the mean connectivity a neuron in population a receives from neurons in population b at location $\mathbf{x}$, and $r_a(\mathbf{x})$ is the firing rate. The symbol $*$ denotes the spatial convolution against $\mathbf{x}$.

Similar to section 2.2, we have

$$w_{aE} * r_E(\mathbf{x}) - w_{aI} * r_I(\mathbf{x}) + f_a(\mathbf{x}) \sim \mathcal{O}(1/\sqrt{N}), \quad a = E, I. \tag{22}$$

Let $N \to \infty$ and perform 2-dimensional Fourier transform against $\mathbf{x}$, and we get

$$\tilde{w}_{aE}\tilde{r}_E - \tilde{w}_{aI}\tilde{r}_I + \tilde{f}_a = 0, \quad a = E, I,$$

where the symbol $\tilde{}$ denotes the spatial Fourier transform. This gives

$$\tilde{r}_E = \frac{\tilde{f}_E\tilde{w}_{II} - \tilde{f}_I\tilde{w}_{EI}}{\tilde{w}_{EI}\tilde{w}_{IE} - \tilde{w}_{EE}\tilde{w}_{II}}, \quad \tilde{r}_I = \frac{\tilde{f}_E\tilde{w}_{IE} - \tilde{f}_I\tilde{w}_{EE}}{\tilde{w}_{EI}\tilde{w}_{IE} - \tilde{w}_{EE}\tilde{w}_{II}}. \tag{23}$$

To ensure that the above Fourier transform exists, it is necessary that $\tilde{r}_a$ tends to 0 as the frequency tends to infinity. This requires that the external input f be "wider" than recurrent input w. This can be understood intuitively from Equation (22), where we see convolution makes $w_{ab} * r_b(\mathbf{x})$ wider than $w_{ab}(\mathbf{x})$, so for the terms to balance each other, f has to be "wider" than w. Also, to get a positive stable solution, the following condition has to be met:

$$\frac{\overline{f}_E}{\overline{f}_I} > \frac{\overline{w}_{EI}}{\overline{w}_{II}} > \frac{\overline{w}_{EE}}{\overline{w}_{IE}}, \tag{24}$$

where the bar represents spatial average. Also, to make the solution stable, w_{aE} has to be "wider" than w_{aI}. For a more detailed account of these conditions, see Rosenbaum and Doiron, 2014; Pyle and Rosenbaum, 2017.

Rosenbaum et al. (2017) proved the asynchronous firing state of the network with local connections under the above conditions. Thus, with the premise of asynchronous firing satisfied, our results regarding the optimum noise structure in section 2.3 still holds. Let the total input variance of the neuron in population a at location $\mathbf{x}$ be $\sigma_a^2(\mathbf{x})$, and the VMR be $\beta_a(\mathbf{x})$, and we have

$$\sigma_a^2(\mathbf{x}) = j_{aE} w_{aE} * r_E(\mathbf{x}) + j_{aI} w_{aI} * r_I(\mathbf{x}).$$

The threshold-linear transfer function gives us $\overline{I_a}(\mathbf{x}) = \theta\tau_a r_a(\mathbf{x})$, so we have

$$\beta_a(\mathbf{x}) = \frac{j_{aE} w_{aE} * r_E(\mathbf{x}) + j_{aI} w_{aI} * r_I(\mathbf{x})}{\theta\tau_a r_a(\mathbf{x})},$$

Here the division is point-wise at each $\mathbf{x}$. If $\beta_a(\mathbf{x})$ is constant at each $\mathbf{x}$ for arbitrary external input $f_a(\mathbf{x})$, it must be spatially invariant, that is, $\beta_a(\mathbf{x}) \equiv \beta_a$. We can thus move the denominator on the r.h.s. to the left, and perform Fourier transform to get

$$\beta_a\theta\tau_a\tilde{r}_a = j_{aE}\tilde{w}_{aE} * \tilde{r}_E(\mathbf{x}) + j_{aI}\tilde{w}_{aI} * \tilde{w}_I(\mathbf{x}), \quad a = E, I.$$

Substituting it in Equation (23), we get

$$-\frac{j_{EE}\tilde{w}_{EE}}{j_{EI}\tilde{w}_{EI}} = \frac{\tilde{f}_E\tilde{w}_{IE} - \tilde{f}_I\tilde{w}_{EE}}{\tilde{f}_E\tilde{w}_{II} - \tilde{f}_I\tilde{w}_{EI}},$$
$$-\frac{j_{II}\tilde{w}_{II}}{j_{IE}\tilde{w}_{IE}} = \frac{\tilde{f}_E\tilde{w}_{II} - \tilde{f}_I\tilde{w}_{EI}}{\tilde{f}_E\tilde{w}_{IE} - \tilde{f}_I\tilde{w}_{EE}}.$$

The above equations cannot be satisfied for all f_a, so this network structure cannot maintain an optimum noise structure and track any input instantly. However, for input changes that only concerns magnitude and not the spatial shape, $\beta_a(\mathbf{x})$ can remain constant and allow instant tracking. For other kinds of input changes, although instantaneous tracking is not possible, the response speed of the network is still significantly smaller than what the neuronal time constant allows, as we shall explore in the next section.

4. SIMULATION RESULTS

One of the potential applications of the balanced network's fast response property is to process Spike Camera data in real time. Spike Camera is a newly developed neuromorphic hardware that encodes visual signals with spikes (Dong et al., 2017). It consists of artificial ganglion cells, each corresponding to a pixel, that linearly integrate the luminance intensity and fire a spike upon reaching the threshold, converting continuous visual information to discrete spikes. This event-based data transmission method significantly reduces the data volume and allows for a sampling rate of as high as 40,000 fps. Compared to another extremely

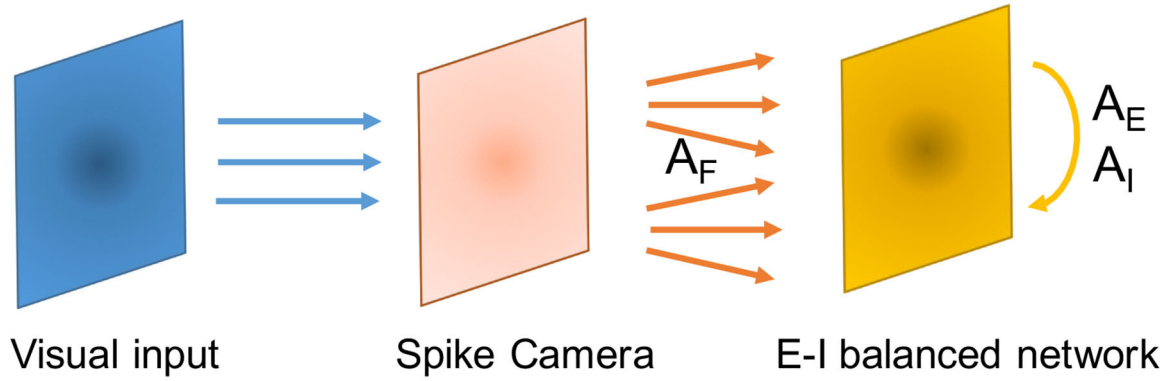

FIGURE 4 | Schematic of the network structure in the context of processing Spike Camera data.

high-speed camera, the Dynamic Vision Sensor (DVS) (Serrano-Gotarredona and Linares-Barranco, 2013), which only transmits changes in light intensity, Spike Camera can directly encode the absolute value of the luminance signal with its spiking rate while having an even higher sampling rate. In this section, we explore the tracking performance of our network under the setting of processing Spike Camera-like data.

4.1. Network Structure

We use a feedforward layer consisting of 50×50 non-leaky linear integrate-and-fire neurons to mimic the Spike Camera. Each neuron in this layer receives visual signal from its corresponding pixel location, and connects to the balanced network layer through feedforward connections $J_{ij}^{aF}, a = E, I$. The balanced network layer consists of 80×80 excitatory neurons and 40×40 inhibitory neurons. The neurons of each population is placed uniformly on a square area with a side length of 1. The neurons in the feedforward layer obeys Equation (1), and have a neuronal time constant of τ_F. To reflect the high sampling rate of Spike Camera, τ_F is set to be very small. The connection probability of the network obeys

$$\mathbb{P}(J_{ij}^{ab} = j_{ab}/\sqrt{N}) = p_{ab} G_b(d_{ij}^{ab}), \quad b = F, E, I, \quad a = E, I,$$

where F stands for the feedforward layer, G_b is a 2-dimensional Gaussian distribution centered at 0 with scale parameter A_b. To satisfy the balancing conditions, we let $A_F > A_E \geqslant A_I$ and make sure that Equation (24) holds. Since spatial location is discretized in the network, to keep the total connection probability from population b to population a at p_{ab}, we normalize G_b by letting $\sum_i G_b(d_{ij}^{ab}) = 1, \forall j$. **Figure 4** demonstrates this structure.

4.2. Tracking Time-Varying Stimuli

We test the tracking performance of our network with four example input stimuli. The first stimulus is the sudden appearance of an object, modeled as an abrupt change in input magnitude at the object's location. **Figure 5A** shows the network's response to this change summarized by the population rate of the excitatory neurons corresponding to the location of interest. We see that in this case, the network's activity tracks the stimulus change very quickly.

The second stimulus is similar to the previous one, except that the input magnitude continuously changes in a sinusoidal manner. **Figure 5B** shows the tracking performance of the network. It can be seen that the network can track the stimulus almost instantaneously, which is expected since β_E is constant here.

The third stimulus is an object moving quickly from left to right in the field of vision, which can be seen as a model of a typical motion tracking task. We use the coordinates of the center of the circular object to represent the location of the stimulus. The coordinates calculated from the Spike Camera data and the balanced network activity are then compared in **Figures 5C,D**. The network activity closely tracks the input, and the spatial information is preserved.

The last stimulus is similar to the previous one, except that the motion is circular instead of linear, which implies a constantly changing velocity. The same method is used to locate the stimulus, and the results are shown in **Figures 5E,F**. The performance is again very good.

4.3. Trackable Speeds

To explore the extent of the network's tracking ability, we next evaluate the temporal and spatial lags of the response. We first change the frequency of the sinusoidal signal in the second task in the previous section (**Figure 5B**) and calculate the phase lag of the balanced network's response. As can be seen in **Figure 6A**, while the phase lag $|\phi|$ increases when the signal frequency $1/T$ is higher, the delay is still very small overall.

Next, we vary the speed of the object's circular motion in the fourth task in the previous section (**Figures 5E,F**) and evaluate the spatial phase lag of the object location decoded from the balanced network activity compared to that of the Spike Camera layer. As shown in **Figure 6B**, the tracking error is small even when the object is moving very quickly.

Since the encoding happens at the population level, input changes have to be propagated through the population to be successfully tracked, and this process is mediated by synaptic

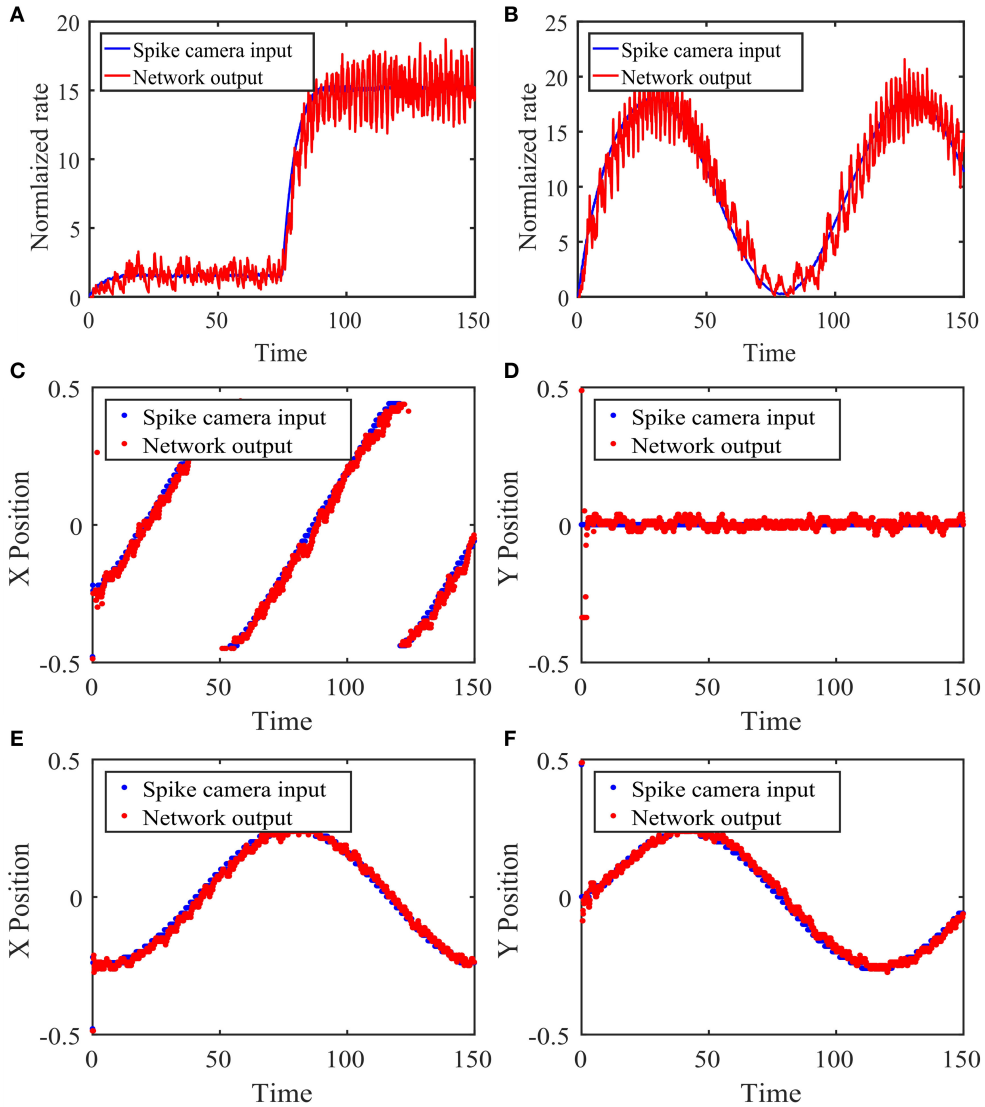

FIGURE 5 | Performance of the network with local connections in response to time-varying stimuli. **(A)** Network response to the sudden appearance of an object. The Spike Camera layer receives a disc-shaped visual input centered at (0.25, 0.5) with a radius of 0.05, whose magnitude changes abruptly from 1.5 to 15 at $t = 75$. A background noise is added. The blue curve is the firing rate of the area corresponding to the visual input in the Spike Camera layer. The red curve is the rate of the excitatory neurons at the same area in the balanced network layer, which is normalized for better comparison with the blue curve. **(B)** Same as panel **(A)**, except that the input amplitude follows the sinusoidal function $\mu(t) = A(\sin(B * 2\pi t/T)) + C, A = 30, B = 3/2, C = 30$. **(C,D)** The stimulus is an object moving across the visual field in constant velocity. The object has the same shape as panels **(A,B)**, with a magnitude of 10. Panels **(C,D)** show the tracking of the x and y coordinates, respectively. The blue curve is the object location decoded from the activity of the Spike Camera layer, and the red curve is that of the balanced network layer. **(E,F)** Same as panels **(C,D)**, except that the stimulus moves counterclockwise on a circle in constant speed. The network parameters are $\theta = 15, \tau_F = 1, \tau_E = 15, \tau_I = 10, \tau_{F,s} = \tau_{E,s} = 5, \tau_{I,s} = 2.5, p_{EF} = 0.05, p_{IF} = 0.025, p_{EE} = 0.02, p_{EI} = 0.08, p_{IE} = 0.06, p_{II} = 0.08, A_F = 0.05, A_E = 0.02, A_I = 0.02, j_{EF} = 140, j_{IF} = 93.3, j_{EE} = 80, j_{EI} = -320, j_{IE} = 40, j_{II} = -320$.

interactions. This lead us to suspect that the synaptic time constants $\tau_{b,s}$ could be a limiting factor for tracking performance. To study this, we varied $\tau_{b,s}$ in both the temporal and spatial tracking tasks. Indeed, as can be seen in **Figure 6**, a shorter synaptic time constant leads to better performance. In practice, the shape of the synaptic current can be designed to have a $\tau_{b,s}$ as small as possible. The real constricting factor is the synaptic transmission delay, which corresponds to the communication speed of the hardware, but this is expected to be insignificant given the highly compact nature of neuromorphic chips.

5. DISCUSSION AND CONCLUSION

This paper proposed an algorithm for fast response in neuromorphic systems based on E-I balanced networks,

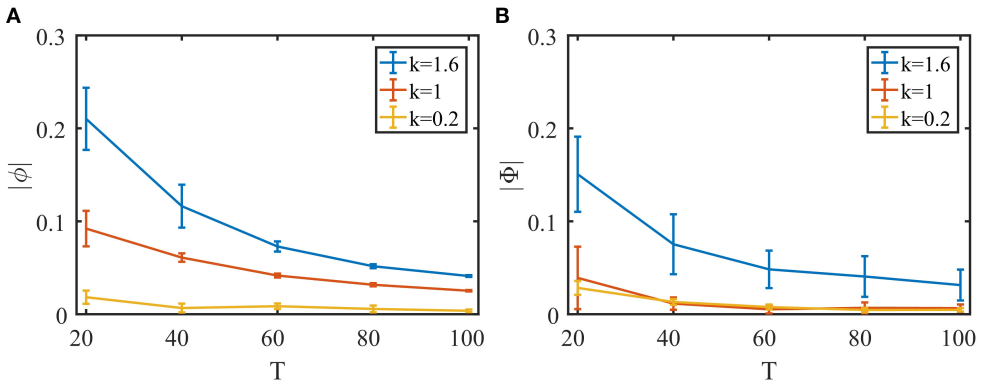

FIGURE 6 | Quantifying the network's performance with temporal and spatial phase lag. To examine the effect of the synaptic time constant on tracking performance, we define $\tau_{b,s} = k\tau_{b,s0}, b = F, E, I$, where $\tau_{b,s0}$ is the set of parameters used in **Figure 5**. **(A)** Temporal phase lag in the second task (**Figure 5B**) with different signal periods. **(B)** Spatial phase lag in the fourth task (**Figures 5E,F**) with different circular motion periods. Ten trials for each data point. Error bars show standard deviations.

systematically analyzed its fast response mechanism, and introduced local connections to maintain balance and retain spatial information in the face of spatially heterogeneous inputs. Simulations verified that the network indeed performs well with rapidly changing input stimuli.

There are still some questions left to explore. For instance, we have mentioned that the network cannot keep an optimal noise structure at all times, and thus the membrane potential distribution will change with the input. A study of the transient dynamics during such changes could help us further improve the network performance. As another example, notice that most of the theoretical analyses in the paper were conducted in the limit of $N \to \infty$. In real-world applications, we often have to track small objects, during which the number of neurons encoding it usually does not exceed a few hundred. Studying the finite-size effect could help us better understand the network dynamics.

Although we mainly discussed the case where the input comes from Spike Camera, the network structure we proposed is not limited to processing visual signal. The "location" of neurons can also correspond to tuning to different variables or representation of abstract features. To achieve real-time processing of high-frequency data, the fast response property is required for each computational process. There has been a lot of research discussing how to implement various computations on top of a balanced network (Barrett, 2012; Hansel and van Vreeswijk, 2012; Litwin-Kumar and Doiron, 2012; Lim and Goldman, 2014; Denève and Machens, 2016; Pyle and Rosenbaum, 2017). The asynchronous irregular state can be taken as a model of the spontaneous state in the cortex. With the spontaneous state as a global attractor, and the specific computations and memories as input-sensitive local attractors (Amit and Brunel, 1997; Litwin-Kumar and Doiron, 2012), the chaos in the network's balanced firing state can allow it to respond to specific inputs very rapidly and initiate the required computation. Besides the fast response property, the balanced state also has other computational advantages such as stochastic resonance (Barrett, 2012).

Neuromorphic computing systems colocalize computation and memory by mimicking neural structures like neurons and synapses. This allows it to circumvent the von Neumann bottleneck, granting it enormous potentials in processing speed (Indiveri and Liu, 2015). There has been a lot of work investigating possible mechanisms for fast neural response (e.g., Bharioke and Chklovskii, 2015; Yu et al., 2015) which could potentially complement the processing speed of neuromorphic systems, and the balance of excitation and inhibition we explored here is one of them. The model we proposed here, with its simple neuron model and connectivity structure, can be readily implemented in hardware and serve as a fast-responding module integrated in a general neuromorphic system for rapid information processing. This paper thus lays the groundwork for realizing various kinds of fast computation using balanced networks, especially in neuromorphic systems.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

SW designed the project. GT, SL, and SW wrote the paper. GT did the theoretical analyses. GT, SL, and SW carried out simulations and data analysis. TH contributed important ideas.

FUNDING

This work was supported by the National Natural Science Foundation of China (No: 61425025, TH) and Guangdong Province with Grant (No. 2018B030338001, SW). This work also received support from Huawei Technology Co., Ltd (YBN2019105137). The authors declare that this study received funding from Huawei Technology Co., Ltd. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

REFERENCES

Amit, D. J., and Brunel, N. (1997). Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cereb. Cortex* 7, 237–252. doi: 10.1093/cercor/7.3.237

Barrett, D. G. T. (2012). *Computation in balanced networks* (Ph.D. thesis). University College London, London, United Kingdom.

Bharioke, A., and Chklovskii, D. B. (2015). Automatic adaptation to fast input changes in a time-invariant neural circuit. *PLoS Comput. Biol.* 11:e1004315. doi: 10.1371/journal.pcbi.1004315

Brunel, N. (2000). Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. *J. Comput. Neurosci.* 8, 183–208. doi: 10.1023/A:1008925309027

Brunel, N., and Hakim, V. (1999). Fast global oscillations in networks of integrate-and-fire neurons with low firing rates. *Neural Comput.* 11, 1621–1671. doi: 10.1162/089976699300016179

Denéve, S., and Machens, C. K. (2016). Efficient codes and balanced networks. *Nat. Neurosci.* 19, 375–382. doi: 10.1038/nn.4243

Dong, S., Huang, T., and Tian, Y. (2017). "Spike camera and its coding methods," in *2017 Data Compression Conference (DCC)* (Snowbird, UT), 437. doi: 10.1109/DCC.2017.69

Dorrn, A. L., Yuan, K., Barker, A. J., Schreiner, C. E., and Froemke, R. C. (2010). Developmental sensory experience balances cortical excitation and inhibition. *Nature* 465, 932–936. doi: 10.1038/nature09119

Fusi, S., and Mattia, M. (1999). Collective behavior of networks with linear (VLSI) integrate-and-fire neurons. *Neural Comput.* 11, 633–652. doi: 10.1162/089976699300016601

Graupner, M., and Reyes, A. D. (2013). Synaptic input correlations leading to membrane potential decorrelation of spontaneous activity in cortex. *J. Neurosci.* 33, 15075–15085. doi: 10.1523/JNEUROSCI.0347-13.2013

Haider, B., Duque, A., Hasenstaub, A. R., and McCormick, D. A. (2006). Neocortical network activity *in vivo* is generated through a dynamic balance of excitation and inhibition. *J. Neurosci.* 26, 4535–4545. doi: 10.1523/JNEUROSCI.5297-05.2006

Hansel, D., and van Vreeswijk, C. (2012). The mechanism of orientation selectivity in primary visual cortex without a functional map. *J. Neurosci.* 32, 4049–4064. doi: 10.1523/JNEUROSCI.6284-11.2012

Huang, L., Cui, Y., Zhang, D., and Wu, S. (2011). Impact of noise structure and network topology on tracking speed of neural networks. *Neural Netw.* 24, 1110–1119. doi: 10.1016/j.neunet.2011.05.018

Indiveri, G., and Liu, S. (2015). Memory and information processing in neuromorphic systems. *Proc. IEEE* 103, 1379–1397. doi: 10.1109/JPROC.2015.2444094

Lim, S., and Goldman, M. S. (2014). Balanced cortical microcircuitry for spatial working memory based on corrective feedback control. *J. Neurosci.* 34, 6790–6806. doi: 10.1523/JNEUROSCI.4602-13.2014

Litwin-Kumar, A., and Doiron, B. (2012). Slow dynamics and high variability in balanced cortical networks with clustered connections. *Nat. Neurosci.* 15, 1498–1505. doi: 10.1038/nn.3220

Okun, M., and Lampl, I. (2008). Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. *Nat. Neurosci.* 11, 535–537. doi: 10.1038/nn.2105

Pyle, R., and Rosenbaum, R. (2017). Spatiotemporal dynamics and reliable computations in recurrent spiking neural networks. *Phys. Rev. Lett.* 118:018103. doi: 10.1103/PhysRevLett.118.018103

Raiguel, S. E., Xiao, D.-K., Marcar, V. L., and Orban, G. A. (1999). Response latency of macaque area MT/V5 neurons and its relationship to stimulus parameters. *J. Neurophysiol.* 82, 1944–1956. doi: 10.1152/jn.1999.82.4.1944

Renart, A., de la Rocha, J., Bartho, P., Hollender, L., Parga, N., Reyes, A., et al. (2010). The asynchronous state in cortical circuits. *Science* 327, 587–590. doi: 10.1126/science.1179850

Risken, H. (1996). *Fokker-Planck Equation*. Berlin; Heidelberg: Springer. doi: 10.1007/978-3-642-61544-3_4

Rosenbaum, R., and Doiron, B. (2014). Balanced networks of spiking neurons with spatially dependent recurrent connections. *Phys. Rev. X* 4:021039. doi: 10.1103/PhysRevX.4.021039

Rosenbaum, R., Smith, M. A., Kohn, A., Rubin, J. E., and Doiron, B. (2017). The spatial structure of correlated neuronal variability. *Nat. Neurosci.* 20, 107–114. doi: 10.1038/nn.4433

Serrano-Gotarredona, T., and Linares-Barranco, B. (2013). A 128 ×128 1.5 mw asynchronous frame-free dynamic vision sensor using transimpedance preamplifiers. *IEEE J. Solid State Circ.* 48, 827–838. doi: 10.1109/JSSC.2012.2230553

Shadlen, M. N., and Newsome, W. T. (1994). Noise, neural codes and cortical organization. *Curr. Opin. Neurobiol.* 4, 569–579. doi: 10.1016/0959-4388(94)90059-0

Softky, W. R., and Koch, C. (1993). The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *J. Neurosci.* 13, 334–350. doi: 10.1523/JNEUROSCI.13-01-00334.1993

Sugase, Y., Yamane, S., Ueno, S., and Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature* 400, 869–873. doi: 10.1038/23703

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature* 381, 520–522. doi: 10.1038/381520a0

van Vreeswijk, C., and Sompolinsky, H. (1996). Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science* 274, 1724–1726. doi: 10.1126/science.274.5293.1724

van Vreeswijk, C., and Sompolinsky, H. (1998). Chaotic balanced state in a model of cortical circuits. *Neural Comput.* 10, 1321–1371. doi: 10.1162/089976698300017214

Yu, L., Wang, L., Jia, F., and Jia, D. (2015). Stimulus-dependent frequency modulation of information transmission in neural systems. *arXiv [preprint]. arXiv:1507.08269*.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.