Synchrony-Division Neural Multiplexing: An Encoding Model

Mohammad R. Rezaei[1,2,5], Milos R. Popovic[2,5], Steven A Prescott[2,3,4], Milad Lankarany[1,2,3,4,5]*

[1] Krembil Research Institute – University Health Network (UHN), Toronto, ON, Canada

[2] Institute of Biomedical Engineering, University of Toronto, Toronto, ON, Canada

[3] Department of Physiology, University of Toronto, Toronto, ON, Canada

[4] Neurosciences and Mental Health, The Hospital for Sick Children, Toronto, ON, Canada

[5] KITE Research Institute, Toronto Rehabilitation Institute - University Health Network (UHN), Toronto, ON, Canada

Keywords: Neural coding, information representation, multiplexed coding, synchronous and asynchronous spikes, general linear model

Abstract

Cortical neurons receive mixed information from collective spiking activities of primary sensory neurons in response to a sensory stimulus. A recent study demonstrated that the time underlying the onset-offset of a tactile stimulus and its varying intensity can be respectively represented by synchronous and asynchronous spikes of S1 neurons in rats. This evidence capitalized on the ability of an ensemble of homogeneous neurons to multiplex, a coding strategy that was referred to as synchrony division multiplexing (SDM). Although neural multiplexing can be conceived by distinct functions of individual neurons in a heterogeneous neural ensemble, the extent to which nearly identical neurons in a homogeneous neural ensemble encode multiple features of a mixed stimulus remains unknown. Here, we present a computational framework to provide a system level understanding on how an ensemble of homogeneous neurons enable SDM. First, we simulate SDM with an ensemble of homogeneous conductance-based model neurons receiving a mixed stimulus comprising slow and fast features. Using feature estimation techniques, we show that both features of the stimulus can be inferred from the generated spikes. Second, we utilize linear nonlinear (LNL) cascade models and calculate temporal filters and static nonlinearities of differentially synchronized spikes. We demonstrate that these filters and nonlinearities are distinct for synchronous and asynchronous spikes. Finally, we develop an augmented LNL cascade model as an encoding model for the SDM by combining individual LNLs calculated for each type of spike. The augmented LNL model reveals that a homogeneous neural ensemble can perform two different functions, namely, temporal- and rate- coding, simultaneously.

I. Introduction

Transmitting multiple signals over a single communication channel increases channel bandwidth and enhances coding efficiency [1, 2]. Similar to digital communication systems, the brain utilizes different forms of multiplexing – in different brain regions and in regard to different stimuli – to represent multiple features of a stimulus by a neural code [2]. For example, in the auditory sensory system, the frequency and intensity of a periodic stimulus are encoded by the phase-locked spikes and the probability of spiking per stimulus cycle, respectively[3]. Similarly, the frequency and intensity of vibrotactile stimuli are represented by the timing and rate of spikes in the somatosensory cortex [4]. Recently, differentially synchronized spiking neurons of the primary somatosensory cortex was shown to enable multiplexed coding of low- and high- contrast features of tactile stimuli [5]. Despite various forms of neural multiplexing, a thorough understanding of how the brain enables multiplexing remained undiscovered. Specifically, functional characteristics – in the sense of linear or nonlinear filtering properties – of a neural ensemble that multiplexes different features of a stimulus is yet to be uncovered.

Different features of stimuli, like the intensity, frequency, onset and offset, etc., dictate which multiplexing strategies are most appropriate [4]. In addition to the properties of stimuli, heterogeneity of neurons in a population code enables different neurons to encode different stimulus features. The functional properties of a heterogeneous neural ensemble, which includes neurons with different functions, e.g., integrators vs. coincidence detectors, might be fully described by the dynamics of individual neurons. For example, an ensemble of heterogeneous cochlear nuclei in the auditory cortex is composed of two anatomically distinct sub-nuclei, namely, the magnocellular- and the angular- nucleus, each of which selectively encodes a specific feature of the stimulus. The magnocellular nucleus selectively encodes stimulus frequency with a temporal code by implementing high-pass filter whereas angular nucleus selectively encodes stimulus intensity with a rate code by implementing a low-pass filter [6]. In contrast to heterogeneous neural ensemble, functional characteristics of an ensemble of homogenous neurons, which includes neurons with nearly identical functions, cannot be identified based on the properties of individual neurons solely [4, 5]. For example, in synchrony division multiplexing

(SDM) [5], information about slow and fast stimulus features were respectively represented by asynchronous and synchronous spikes of the same neurons. Thus, this form of multiplexing suggests that both slow and fast features of the stimulus can be encoded by homogeneous (identical) neurons that operate in a hybrid mode [5], i.e., neither low-pass nor high-pass filtering the stimulus [7, 8]. Thus, a challenging question is that whether multiplexing (like SDM) in a homogeneous neural ensemble reveals system-level functions beyond those performed by individual neurons [5].

In this paper, we utilize conductance-based and linear nonlinear (LNL) cascade models to establish a theoretical framework to address this question [9-12]. First, we use conductance-based models and construct a homogeneous neural ensemble that multiplexes slow and fast features of a common stimulus using asynchronous and synchronous spikes, respectively. Using the LNL cascade models, we explore whether different linear filters and static nonlinearities are associated with different types of spikes. We show that a low-pass filter followed by a nonlinearity with mild slope generates asynchronous spikes whereas a high-pass filter followed by a nonlinearity with steep slope detects fast features of the stimulus by generating synchronous spikes. Then, we develop an augmented LNL model for SDM by integrating LNL models underlying each type of spike.

II. Results

In the present paper, we developed an augmented LNL cascade model as an encoding model for the SDM [5]. Conductance-based neuron models were used to create an ensemble of homogeneous neurons whose input (mixed stimulus) – output (spikes) relationship was estimated by the augmented LNL model.

As shown in **Figure 1A**, we construct an ensemble of homogenous neurons with 30 Morris-Lecar (ML) neuron models (see Methods) all of which receive a common mixed signal comprising slow and fast features as well as independent physiologically-realistic conductance noise [5, 13]. The parameters of the ML model were selected in a way that all neurons operate in a hybrid mode [14]. Spikes generated by an ensemble of ML neurons were used to fit the augmented LNL model in which two separate LNL models were combined to represent rate and

temporal codes simultaneously. This study shows that an ensemble of homogeneous neurons utilizes different strategies to generate synchronous and asynchronous spikes which enable simultaneous coding of fast and slow features of a mixed stimulus, respectively. Although the biophysical mechanisms underlying implementation of SDM by an ensemble of homogeneous neurons is still unknown, the two-stream augmented LNL model provides a system-level understanding of the SDM function.

II. A. Different temporal filters map distinct features of a mixed stimulus

To explore how slow and fast features of the stimulus are encoded by spikes of an ensemble of neurons, we used well-known feature space estimators like the spike-triggered average (STA) [15, 16] or information-theoretic spike-triggered average and covariance (iSTAC) to reveal temporal characteristics of neurons in response to a stimulus[17].

The STA filter is a precise and unbiased predictor for a neural population given a stationary and single dimension stimulus [16]. However, it fails to provide precise predictions when the dimensionality of the stimulus is larger than one. For example, in retinal ganglion cells the STA cannot predict the neural response of both ON and OFF cells given a mixed input comprising more than one feature. To explore other possible subspace features of the neural response, we used the iSTAC method and calculated the optimal subspace features. The iSTAC quantifies the significance of subspaces based on the mutual information between stimulus and neural response [17]. In this method, we choose eigenvectors of the spike-triggered stimulus ensemble matrix more precisely by minimizing the Kullback-Leibler (KL) divergence between the eigenvectors of ensemble matrix and the raw stimulus distributions (see the method section for more details). In fact, the iSTAC maximizes information based on the first two moments of the spike-triggered stimulus ensemble and provides a unifying information-theoretic framework that captures the ensemble neuron activity in different subspaces. This provides an implicit model of the contribution of the nonlinear function mapping the feature space to the neural response. As shown in **Figure 1B (left)**, the iSTAC matrix calculated for the mixed stimulus has two significant eigenvalues whose underlying

eigenvectors reveal two distinct temporal filters, namely, ν_1 and ν_2. The projection of the spike-triggered stimulus ensemble on ν_1 and ν_2, shown in **Figure 1B (right)**, reveals two distinct clusters.

In synchrony-division multiplexing, a mixed input signal containing slow and fast features drives an ensemble of neurons. The fast component of the stimulus whose neural representation is synchronous and sparse does not appear in the STA, as it averages out sparsely-occurring fast features of the stimulus [5]. However, unlike the fast signal, neural representation of the slow signal is asynchronous and dense , thus the STA filter mainly contains information of the slow components of the mixed signal [5]. Unlike the STA filter, the most informative subspaces selected by the iSTAC method behave as multi-space feature estimators and illustrate slow and fast features of the mixed stimulus.

Figure 1C shows that the STA filter calculated for the mixed stimulus mainly captures the slow feature of the signal, but cannot truly capture the dynamics of synchronous spikes. Unlike the STA filter, ν_1 and ν_2 of the iSTAC method illustrates slow and fast features of the mixed stimulus. As can be observed in **Figure 1C**, ν_1 is similar to the STA filter and represents the slow component of the stimulus, and ν_2 describes fast features of the stimulus (note that the STA filter was duplicated in **Figure 1C** (left and right) and compared with both ν_1 and ν_2) .

II. B. Low-dimension feature space of the neural response can be characterized by the STAs of synchronous and asynchronous spikes

Recently, it has been shown that synchronous and asynchronous spikes encode information of slow and fast features of a mixed stimulus (equivalent to that used in the present study), respectively [13, 18, 19]. Using an information theoretic approach, it was shown that synchronous and asynchronous spikes carry information in different time scales. By classifying spikes of a population of neurons into synchronous and asynchronous spikes, it was demonstrated that the STA filters underlying these spikes, namely, μ_{Async} and μ_{Sync}, reflect the fast and slow features of the stimulus, respectively. **Figure 2A** shows classified synchronous (red) and asynchronous (blue) spikes in the raster plot.

Here, we compared these filters with those obtained by the iSTAC method. First, we tested if the projection of the spike-triggered stimulus ensemble on μ_{Async} and μ_{Sync} creates two distinct clusters similar to that projected on v_1 and v_2. As shown in **Figure 2B (left)**, two distinct and separable clusters were generated by μ_{Async} and μ_{Sync}. More importantly, one can distinguish between these clusters by projecting synchronous- and asynchronous-spike-triggered stimulus ensemble on μ_{Async} and μ_{Sync}. **Figure 2B (right)** reveals that these stimulus ensembles are separable and mutually exclusive. **Figure 2C** shows temporal patterns of μ_{Async} and μ_{Sync} versus the STA filter. As expected, μ_{Async}resembles the STA filter, indicating slow features of the stimulus, and μ_{Sync}(similar to v_2) describes abrupt changes in the stimulus.

Furthermore, to investigate the functional roles of the above filters, we tested how they contribute in the signal reconstruction. The reconstructed signal was obtained by the convolution of spikes - either all spikes for STA (**Fig 3A**) or v_1 and v_2 (**Fig 3B**) or asynchronous and synchronous spikes for μ_{Async} and μ_{Sync} respectively (**Fig 3C**). **Figure 3** illustrates 10 sec sample of the reconstructed signal using these methods. As clear in **Figure 3.B**, the signal reconstructed by v_1 and v_2(iSTAC method) resemble that generated by μ_{Async} and μ_{Sync}, and both of these signals better capture fast features than that obtained by the STA filter, indicating the functional relevance between these filters.

II. C. Different nonlinear functions are associated with synchronous and asynchronous spikes

Given different temporal filters underlying synchronous and asynchronous spikes, we sought how these filters map fast and slow features of the mixed stimulus to the firing rate of an ensemble of conductance-based model neurons. Moreover, since the dynamics of a neural ensemble is not fully linear, these linear filters are not sufficient to project the stimulus to spikes. We utilized a well-known phenomenological model, namely, the LNL cascade model, which uses a linear stimulus filter followed by a static nonlinear transformation, to estimate the firing rate of an ensemble of neurons. **Figures 4** and **5** (panel A in both figures) show the

LNL diagram for asynchronous and synchronous spikes, respectively. We tested if a pair of linear filter and static nonlinearity is different for synchronous and asynchronous spikes given a common mixed signal. We obtained static nonlinearity functions for synchronous and asynchronous spikes by applying μ_{Sync} and μ_{Asycn} filters to the mixed stimulus (s) and mapping their outputs (through the nonlinearity) to the PSTHs of synchronous and asynchronous spikes, respectively:

$$PSTH_{Async} = f_{Async}\left(\mu_{Async} * s\right) \quad (1.a)$$

$$PSTH_{Sync} = f_{Sync}\left(\mu_{Sync} * s\right) \quad (1.b)$$

where $f_{Async}(x)$ and $f_{Sync}(x)$ are the nonlinearities associated with asynchronous and synchronous spikes, respectively.

Figures 4(B) and **5(B)** show respectively the raw nonlinearities for asynchronous and synchronous spikes that correspond to the mapping of every single point of the output of linear filters (x-axis) to the values of PSTHs (y-axis). For the nonlinearities underlying asynchronous and synchronous spikes, we fitted the ReLU nonlinearity and sigmoid functions, respectively [20]. The nonlinearly associated with asynchronous, $f_{Async}(x)$, has mild slope and broad dynamic range, enabling rate-modulated coding. In contrast, the nonlinearly underlying synchronous spikes, $f_{Sync}(x)$), has steep slope and narrow dynamic range, enabling event (abrupt changes) detection. Although more sophisticated nonlinear functions could provide better fits, we chose simple and well-established nonlinear functions to highlight the difference in shapes of nonlinearities underlying rate- versus temporal- codes in the context of SDM. The instantaneous firing rates of each type of spike can be constructed by passing the output of temporal filter through the fitted nonlinearities. These firing rates were estimated and drawn against the PSTHs of asynchronous and synchronous spikes in **Figures 4(C-D)** & **5(C-D)**, respectively. As shown in these figures, the nonlinear functions and estimated PSTHs underlying temporal filters obtained by the iSTAC (*V1* and *V2*) and classified spikes (μ_{Async} and μ_{Sync}) are nearly identical.

II. D. An augmented LNL cascade model for synchrony-division multiplexing

The LNL cascade models were utilized to encode specific features of a mixed stimulus by synchronous or asynchronous spikes. As shown in the previous sections, temporal filters and nonlinear transformations of either types of spikes were distinct and estimated by separate LNL cascade models. Here, we sought if a combination of these cascade models, i.e., an augmented LNL model, could accurately encode different features of a mixed stimulus through different types of spikes. We developed a two-stream LNL cascade model that combines the PSTHs of synchronous and asynchronous spikes to reconstruct the mixed PSTH of both type of spikes, as:

$$PSTH_{total} = \sum_{i\in\{Sync,Async\}} \omega_i \times G_i * f_i(\mu_i * s) \quad (2)$$

where ω is the combination weight for each stream of reconstructed PSTHs. To reduce the model complexity and promote smoothness in the output, we applied parameterized Gaussian kernels, $G_{Async} = Gaussian\,(0, \sigma_{Async})$ and $G_{Sync} = Gaussian\,(0, \sigma_{Sync})$, to the reconstructed PSTHs in each stream [21, 22]. The augmented LNL model simultaneously encodes slow and fast features of the stimulus by asynchronous and synchronous spikes, respectively. **Figure 6(A)** shows the block diagram of the augmented LNL model. This model implies that the low-pass filter and shallow non-linearity underlying asynchronous spikes are required to produce the rate code. In contrast, the high-pass filter and sigmoid nonlinearity of synchronous spikes are necessary to preserve reliable spike times underlying fast features of the stimulus. Taken together, the augmented LNL model makes the coexistence of the rate- and temporal- codes happen to encode multiple features of a mixed stimulus. To capitalize on the significance of temporal filters and nonlinearity transformations of each type of spike in estimating the total firing rate of a neural ensemble, we compared the performance of augmented LNL model with that of a conventional one stream LNL. **Figure 6(B-D)** shows the estimated firing rate of three methods, namely, Poisson GLM and augmented LNL models (**Figure 6(C-D)**) (used Section II.C), against the PSTH of ensemble of neurons. As can be observed, the firing rate estimated by the augmented LNL models can better

capture both the rate of asynchronous spikes and timing of synchronous events compared to those estimated by one stream Poisson GLM.

III. Discussion

The ability of an ensemble of homogeneous cortical neurons to multiplex multiple features of a mixed stimulus was studied in [5]. However, specific encoding functions underlying these neurons were not determined. In this paper, we presented a computational framework to provide a system level understanding of the encoding mechanism underlying SDM. We used conductance-based neuron models to construct a homogenous neural ensemble that encodes slow and fast features of a mixed stimulus through asynchronous and synchronous spikes, respectively. To elucidate the contribution of slow and fast features of the mixed stimulus on the spikes generated by model neurons, we calculated most significant subspaces (eigenvectors) of spike-triggered stimulus matrix using the iSTAC method. We demonstrated that the calculated first and second eigenvectors resemble slow and fast features of the stimulus, respectively. Furthermore, the projection of spike-triggered stimulus matrix on these eigenvectors created two distinct clusters. We tested if these clusters can be characterized by synchronous and asynchronous spikes. By computing the spike-triggered average (STA) filters of synchronous and asynchronous spikes, and projecting this matrix on these filters, we clearly separated those clusters. Furthermore, we fitted a LNL model to each type of spikes. Similar to distinct temporal filters for synchronous and asynchronous spikes, their static nonlinearities have different properties. We found that the nonlinearity associated with asynchronous spikes is very shallow and can be approximated by a linear function. On the other hand, the nonlinearity associated with synchronous spikes has a very steep slope and can be approximated by highly nonlinear functions like sigmoid function. Finally, we developed an augmented LNL model to capture both dynamical characteristics of synchronous and asynchronous spikes and reconstruct the PSTH of all spikes.

Subspace Feature Estimators; iSTAC vs. STC

To explore more than one subspace features for stimulation-evoked neural responses, we compared the performance of the STC and iSTAC methods. One can

find most informative subspaces that maximize the mutual information between stimulus and response [23, 24]. Nevertheless, an accurate estimation of mutual information requires a large amount of data although no guarantee for optimal estimation can be necessarily expected[24]. A conventional way to find these subspaces is to calculate those related to the most significant eigenvectors of the spike-triggered covariance (STC) matrix [16]. The eigenvectors of the STC matrix provide analytic expressions for filter estimation using the moments of the stimulus and spike-triggered stimulus distribution [16, 17]. However, this method does not incorporate joint information between the mean and the variance, and also there is no specific measurement for selecting the most significant subspaces based on that information. We calculated the most important eigenvectors of the STC matrix underlying the mixed stimulus and neural response (see **II.B**, details are provided in **Methods**). The first eigenvector of the STC matrix was similar to the STA of all spikes (see **Figure 7**) including both slow and fast features of the stimulus. Unlike the first eigenvector of the STC matrix, the second eigenvector comprised fast features of the stimulus. As shown in Figure **7**, the 2-D projection of the spike-triggered stimulus matrix on eigenvectors of the STC matrix cannot be clearly separated in two distinct clusters.

To avoid this problem, we used the iSTAC method that allows to choose eigenvectors of the spike-triggered stimulus ensemble matrix more precisely by minimizing the Kullback-Leibler (KL) divergence between the eigenvectors of this matrix and that obtained by raw stimulus distributions [17]. It is to be noted that the whitening transformation is usually used before finding subspaces of the spike response. One can use whitening transformation to calculate uncorrelated and normalized subspaces (for both STC and iSTAC methods). However, due to the type of the mixed stimulus (i.e., structured and not a random process), we found that eliminating this transformation results in more representative subspace features as shown in **Figure 1**. We compared the 2-D projection of the spike-triggered stimulus matrix on eigenvectors of the iSTAC method with and without whitening. It can be clearly observed that the iSTAC without whitening can better separate the 2-D space.

Choice of Static Nonlinearity in the LNL Model

The static nonlinearities obtained in the augmented LNL model can describe why synchronous and asynchronous spikes are associated with different functions. For example, the smoothness and linear behavior of $f_{Async}(x)$, for $x>0$, generates a smooth PSTH for asynchronous spikes which linearly encodes to the intensity of the stimulus. In contrast, the sigmoid-like nonlinearity of synchronous spikes, $f_{Sync}(x)$, maintains the PSTH of synchronous spikes very sparse which nonlinearly detects abrupt changes of stimulus. It is worth mentioning that more flexible nonlinear functions could provide better fits for representing the PSTH of synchronous and asynchronous spikes. Of note, one can use deep neural networks (DNNs) to give more flexibility to the models. A DNN is simply a high-dimensional non-linear function estimator which gives a multilayer nonlinear function in the form of a neural network [25, 26]. To use that in our augmented LNL model or any LNL model, one can easily replace the nonlinearity estimator with a DNN.

Generalized Linear Model (GLM) for Augmented LNL

The proposed augmented LNL can also be interpreted in the GLM framework. From this point of view, synchronous and asynchronous PSTHs are modeled by two separated GLMs with different random processes, which eventually combine their PSTHs linearly. The first GLM, filters the mixed stimulus with the first eigenvector of iSTAC and then by passing it through a nonlinearity and then a Gaussian random process (with a linear nonlinearity as its link function), it models the PSTH related to asynchronous spikes. Likewise, the second GLM, filters the mixed stimulus with the second eigenvector of iSTAC and then by passing it through a nonlinearity and then a Bernoulli random process (with a sigmoid nonlinearity as its link function), it models the PSTH related to synchronous spikes (See the method section for more details about GLMs). Or, we even simply can interpret the augmented LNL as a single Poisson GLM, with two input filters (the first two eigenvectors of iSTAC) and a Poisson random process at the end (see the appendix A for more details). To reach the optimal parameters set for the model and avoid computational complexity, we use parametric models for the static nonlinearities [9]. We also can use more flexible parametric function (with parameter set θ) like ex-quadratic function $f_\theta(x)$ as the static nonlinearities. By using ex-quadratic function as nonlinearities we

eventually need to optimize a convex cost function, Which gives the optimum parameters set θ for the nonlinearity and can be optimally optimized by a maximum-likelihood (ML) algorithm (details in the Appendix B) [9, 27, 28].

Materials and Methods

Simulated mixed input

According to the feasibility of neural systems to multiplexed coding, we simulated the activity of a homogeneous neural ensemble in response to a mixed-stimulus to explore how much information can be encoded by different patterns of spikes. Each neuron received a mixed signal (I_{mixed}) which consists of a fast signal (I_{fast}) and a slow signal (I_{slow}). I_{fast} Stands for the timing of fast events or abrupt changes in the stimulus and was generated by convolving a randomly (Poisson) distributed Dirac-delta function with a synaptic waveform (normalized to the peak amplitude), $\tau_{rise} = 0.5\ ms$, and $\tau_{fall} = 3\ ms$. Fast events occurred at a rate of $\sim 1\ Hz$ and were scaled by $a_{fast} = 85\ pA$.

I_{slow} was generated by an OU process as follows.

$$\frac{dI_{slow}}{dt} = -\frac{I_{slow}(t) - \mu}{\tau} + \sigma\frac{\sqrt{2}}{\tau}\xi(t), \qquad \xi \sim (0,1) \tag{3}$$

where ξ is a random number drawn from a Gaussian distribution, $\tau = 100\ ms$ is the time constant of the slow signal that produces a slow-varying random walk with an average of $\mu = 15\ pA$ and a standard deviation of $\sigma = 60\ pA$. The mixed signal (I_{mixed}) was obtained by adding I_{fast} and I_{slow}, were generated independently.

An independent noise (equivalent to the background synaptic activity) was added to each neuron, thus each neuron receives a mixed signal plus noise. Similar to, the noise (I_{noise}) was generated by an OU process of $\tau = 5\ ms, \mu = 0\ pA$, and $\sigma = 10\ pA$.

Simulated neural ensemble and its response to mixed input

The neural ensemble consists of 30 neurons, each of them was modeled by Morris-Lecar equations [13, 29]. The equations of a single model neuron receiving a mixed-signal plus noise can be written as follows.

$$C\frac{dV}{dt} = I_{mixed}(t) + I_{noise}(t) - \bar{g}_{Na}m_{\infty}(V)(V - E_{Na}) - \bar{g}_{K}w(V - E_{K}) - g_{L}(V - E_{L}) - \bar{g}_{AHP}z(V - E_{K}) - g_{exc}(V - E_{exc}) - g_{inh}(V - E_{inh}) \quad (4)$$

where,

$$\frac{dw}{dt} = \phi\frac{w(V) - w}{\tau_{W(V)}} \quad (5)$$

$$\frac{dz}{dt} = \frac{\frac{1}{1 + e^{(\beta_z - V)/\gamma}} - z}{\tau_z} \quad (6)$$

$$m_{\infty}(V) = 0.5\left[1 + tanh\left(\frac{V - \beta_m}{\gamma_m}\right)\right] \quad (7)$$

$$w_{\infty}(V) = 0.5\left[1 + tanh\left(\frac{V - \beta_w}{\gamma_w}\right)\right] \quad (8)$$

$$\tau_w(V) = \frac{1}{cosh\left(\frac{V - \beta_w}{2\beta_w}\right)} \quad (9)$$

where $\{\underline{g}_{Na} = 20, \underline{g}_{k} = 20, \underline{g}_{L} = 20,\ \underline{g}_{AHP} = 25, g_{exc} = 1.2, g_{inh} = 1.9\}\frac{mS}{cm^2}$, $\beta_m = -1.2,\ \gamma_m = 18, \beta_w = -19, \gamma_w = 10, \beta_z = 0, \gamma_z = 2, \tau_a = 20\ ms, \phi = 0.15, and\ C = 2\frac{\mu F}{cm^2}$. These parameters were set to ensure a neuron operates in a hybrid mode [30], i.e., an operating mode between integration and coincidence detection [5, 31]. The inclusion of background excitatory and inhibitory synaptic conductance in (2) reproduced a "balanced" high conductance state. The surface area of the neuron was set to 200 μm^2 so that I_{mixed} is reported in pA, rather than as a density[32, 33]. **Figure 1.A** Shows the mixed stimulus and the spiking activates of the ensemble of neurons in response to this stimulus.

Generalized Linear Model (GLM) details

GLM model is a generalization of traditional linear models, which gives the neural encoding models more flexibility to capture nonlinear dynamics of neural activity. GLM contains three stages. The first stage is a linear mapping which consists of a set of d linear-filters, let's assume $\boldsymbol{K} = [k_1, \dots, k_D]$, that maps high dimensional sensory stimulus $s \in R^M$ into a low dimensional stimulus feature map $x \in R^D$:

$$\boldsymbol{x} = \boldsymbol{K^T s} \tag{10}$$

The second stage is a pointwise nonlinearity, $f: R^D \rightarrow R$, which maps the linear features of d dimensions into a nonnegative spike rate:

$$\boldsymbol{\lambda} = f(\boldsymbol{x}) \tag{11}$$

In the final stage, the number of spikes generated by a random process:

$$P_\theta(\boldsymbol{Y} = \boldsymbol{r}|\ \boldsymbol{s}) \tag{12}$$

Where Y is random variable related to spikes occurrence, $\boldsymbol{r}$ is instantaneous firing rate, and the $\theta$is parameter set of the random process

In simple words, by using GLM we approximate the instantaneous firing rate by considering feature from D dimensions instead of M dimensions:

$$P(\boldsymbol{Y}|\boldsymbol{s}) \sim P(\boldsymbol{Y}|\boldsymbol{K^T s}) \tag{13}$$

So, there are two set of parameters, the estimators ($\boldsymbol{K}$) and the pointwise nonlinearity (f), which can be optimized to reach the desired model.

STA and STC estimator

If we assume that $p(s)$ is has zero mean, then the STA can be defined as the average of the stimulus given the instantaneous firing rate:

$$\mu = \frac{1}{n_{sp}} \sum_{\{s_i|spike\}} s_i\ ,\ n_{sp} = \sum_{t=1}^{N} r_t \tag{14}$$

Where N is the total number of time points. The STA is an unbiased, consistent estimation which gives the direction in the stimulus space along which the means of $P(\boldsymbol{s}|spike)$ and $P(\boldsymbol{s})$ differ most. The problem is The STA gives a single direction in stimulus space and leads to a single estimator which is not efficient to capture all information in the stimulus space (we previously discussed we have a mixed stimulus in this research). To involve other possible directions with maximally differences in variances between $P(\boldsymbol{s}|spike)$ and $P(\boldsymbol{s})$ we can use eigenvectors of the STC matrix, defined as:

$$\Lambda = \frac{1}{n_{sp}} \sum_{\{s_i|spike\}} (s_i - \mu)(s_i - \mu)^T. \tag{15}$$

The STA and eigenvectors of the STC matrix can provide a basis, K, for a reduced dimensional model of the neural response.

iSTAC estimator

There are two major problems with STA/STC, when we consider more than one direction for stimulus space. The first one is that STA is not orthogonal to STC eigenvectors and this increases the risk of losing information, and the second one is that the measure we use to select eigenvectors of STC is based on eigenvalues, which does not truly represent most informative directions. As we mentioned before the objective in iSTAC is to reduce KL divergence between Gaussian approximations to the spike-triggered and raw stimulus distributions. Therefore, we define Q as a Gaussian approximation of $P(\boldsymbol{s}|spike)$ based on the information contained only in the mean and covariance of the P as:

$$Q(\boldsymbol{s}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Lambda|^{\frac{1}{2}}}\, e^{-\frac{1}{2}(\boldsymbol{s}-\mu)^T\Lambda^{-1}(\boldsymbol{s}-\mu)} \tag{16}$$

Where n is dimensionality of stimulus space. Now we drive KL divergence between Q and P as:

$$D(Q,P) = \int Q(s)\; log\; log\; \left(\frac{Q(s)}{P(s)}\right)\; ds \tag{17}$$

By considering that mean of P and P. Q is zero and have identity covariance (if not, we can use "whitening" technique) we can rewrite D in a simpler form as:

$$D(Q,P) = \frac{1}{2}(Tr(\Lambda) - log\; log\; |\Lambda| + \mu^T\mu - n) \tag{18}$$

Where $Tr(.)$ And $|.|$ indicate trace and determinant, respectively.

Based on the fact that we are interested in d subspaces we can approximate the D with:

$$D_{[K]}(P,Q) = \frac{1}{2}(Tr[K^T(\Lambda + \mu^T\mu)K - log\; log\; |K^T\Lambda\, K| - d\,] \tag{19}$$

where d is the dimension of the interested subspaces.

So, in terms of finding the d most informative subspaces decomposed by STA and eigenvectors of STC we need to find $D_{[K]}(P,Q)$ for all subspaces and select the first d ones.

An important advantage of the iSTAC approach over traditional STA/STC analysis is that it makes statistically efficient use of changes in both mean and covariance of the response spaces.

Figures

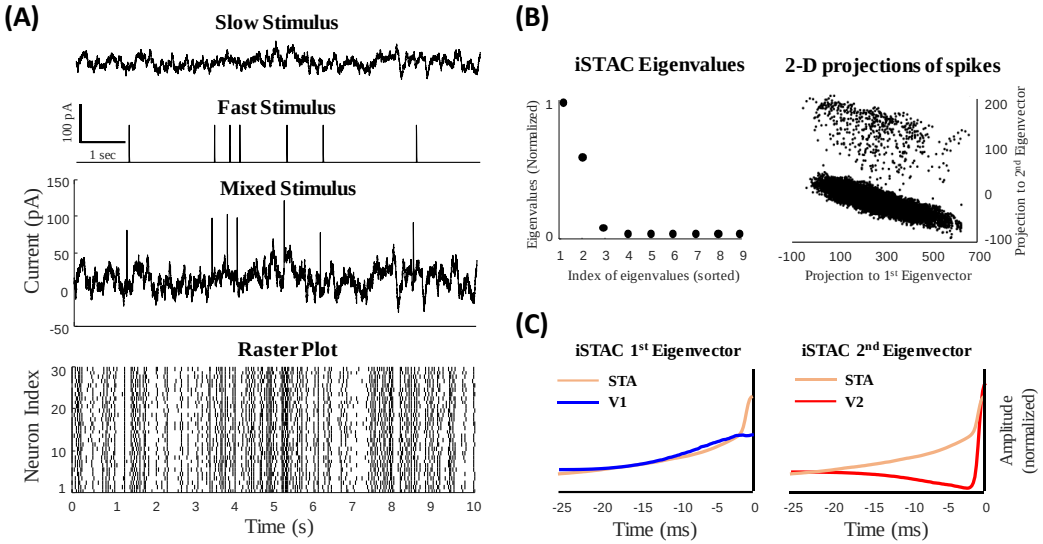

Figure 1. Slow and fast features of a mixed signal can be inferred from responses of a homogeneous ensemble of neurons using the iSTAC method. (**A**) Slow and fast signals constructing a mixed signal. (A, Bottom) Sample raster plot of 30 model neurons receiving the common mixed signal (and independent noise). Spikes evoked by the fast and slow signals cannot be distinguished visually. (**B**) The iSTAC method was applied to spike-triggered mixed signal and eigenvalues and eigenvectors were obtained (see Methods). (B, Left) The eigenvalues of the iSTAC matrix reveals two significant components of the population code. (B, Right) The projection of spike-triggered mixed signal onto the main eigenvectors of the iSTAC matrix. Two clusters can be visually distinguished. (**C**) The 1st and 2nd eigenvectors of the iSTAC matrix, *V1* and *V2*, respectively, are shown against the spike-triggered average (STA). *V1* resembles the STA filter reflecting slowly-varying changes in the signal. Unlike *V1*, *V2* resembles a high-pass filter (differentiator) which reflects fast features of the mixed signal.

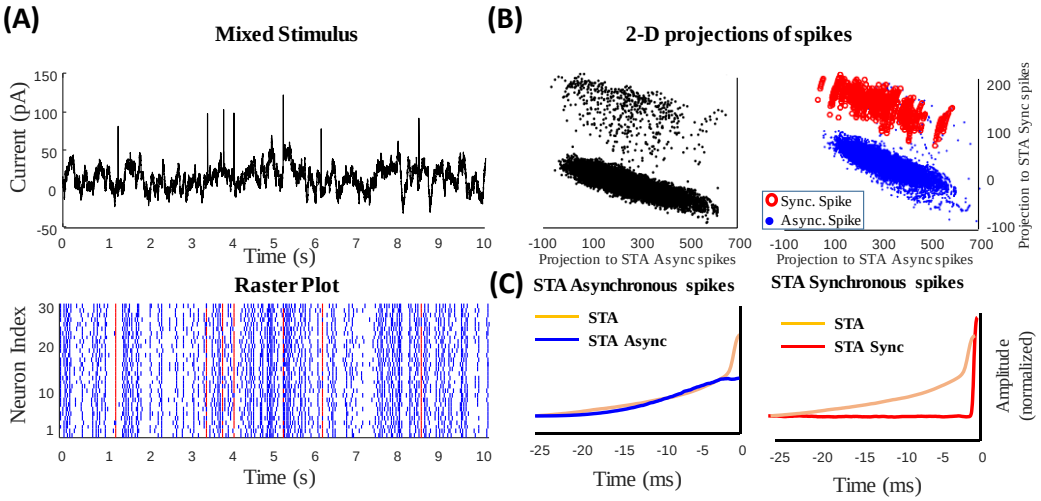

Figure 2. Synchronous and asynchronous spikes represent information of slow and fast features of the mixed signal, respectively. (**A**) Synchronous (red) and asynchronous (blue) spikes are distinguished by setting a threshold on the instantaneous firing rate calculated by a narrow kernel (see Methods). Synchronous spikes evoked by the fast signals can be distinguished visually. (**B**) The projection of spike-triggered mixed signal onto the STA_{Sync} and STA_{Async}. Two (visually) distinguishable clusters belong to asynchronous spikes representing the slow feature of the signal (blue dots) and synchronous spikes representing the fast features (red circles). (**C**) The spike-triggered average of synchronous (red) and asynchronous (blue) spikes, namely, STA_{Sync} and STA_{Async}, respectively, was shown against the STA of all spikes (similar to Figure 1. C).

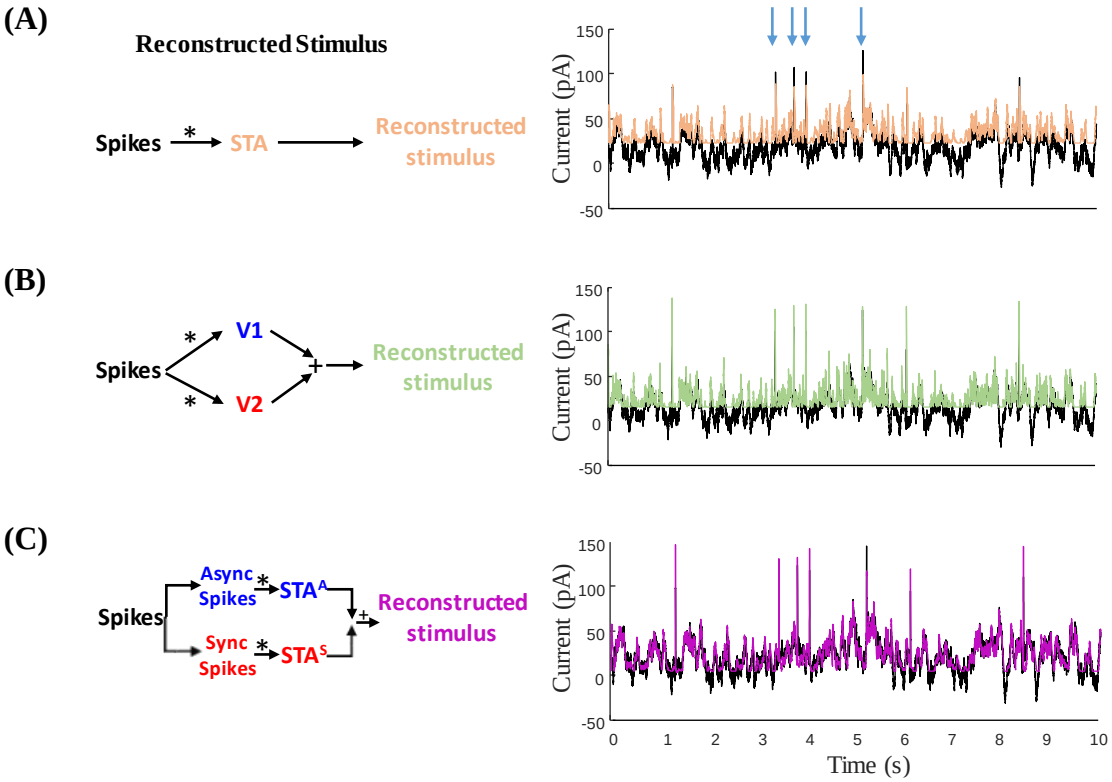

Figure 3. Block diagram of decoding the mixed signal from spikes by (**A**) the STA filter (light brown), (**B**) a weighted sum of the 1st and 2nd eigenvectors of the iSTAC method (light purple), and (**C**) a weighted sum of filtered asynchronous spikes (by STA $_{Async}$) and filtered synchronous spikes (by STA_{Sync}) (purple). Original mixed signal overlaid with black color in the plots. As can be seen in these figures, the reconstructed signal based on STA_{Sync} and STA_{Async} – similar to that obtained by eigenvectors of iSTAC method– can capture both slow and fast components of the signal accurately.

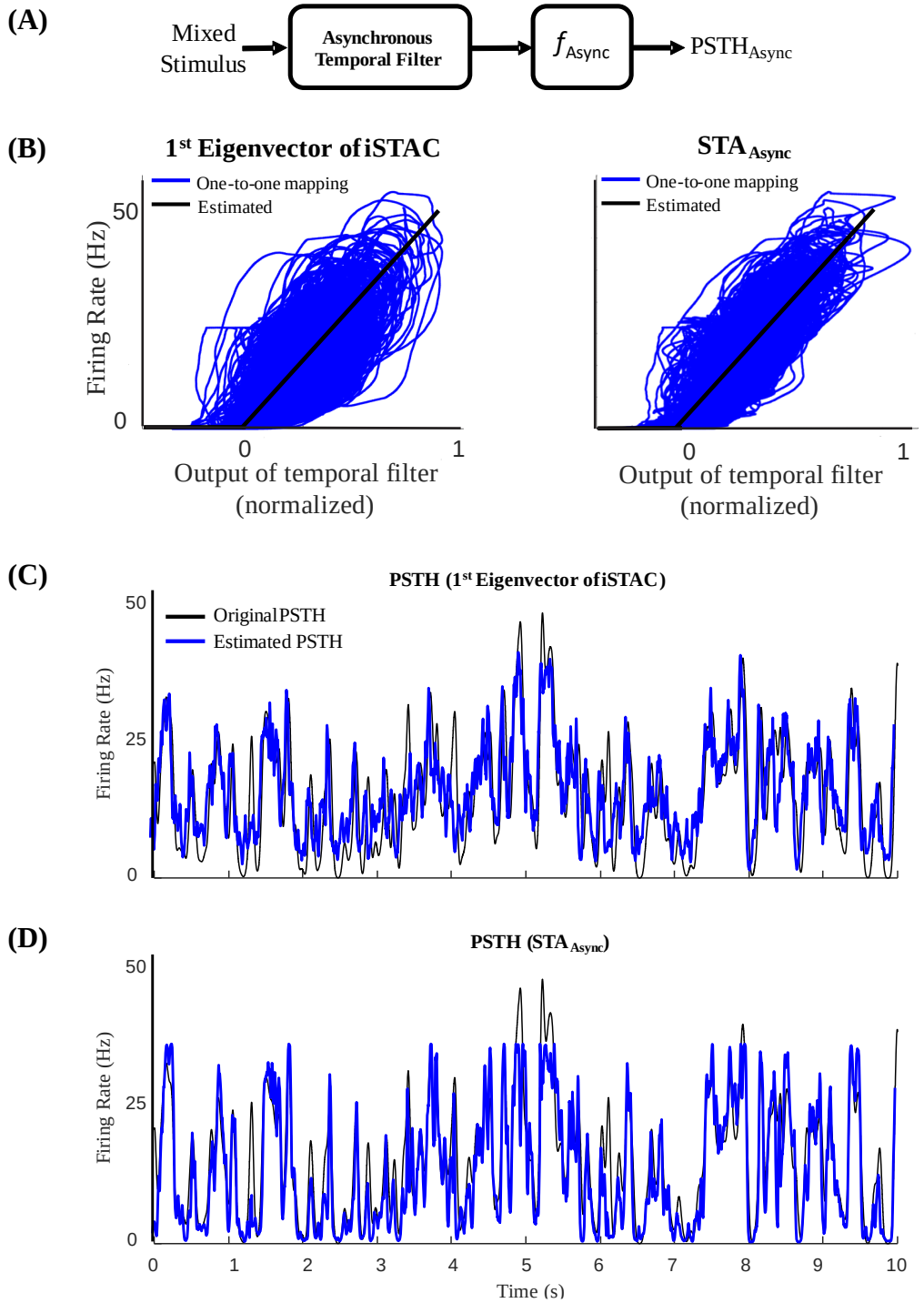

Figure 4. Static nonlinearities underlying asynchronous spikes. (**A**) Blog diagram of LNL model for asynchronous spikes. (**B**) Static nonlinearity calculated for the asynchronous spikes is obtained by mapping the output of filtered stimulus to the instantaneous firing rate of asynchronous spikes (calculated by a wide kernel, σ = 25 msec). Static nonlinearity calculated based on 1st eigenvectors of the iSTAC method, *V1*, (Left) and STA_{Async} (Right). The solid black shows fitted rectifiers. (**C**) The PSTHs constructed by the fitted nonlinearities based on *V1* were drawn against the PSTH of asynchronous spikes. **(D)** The PSTHs constructed by the fitted nonlinearities based on STA_{Async} were drawn against the PSTH of asynchronous spikes.

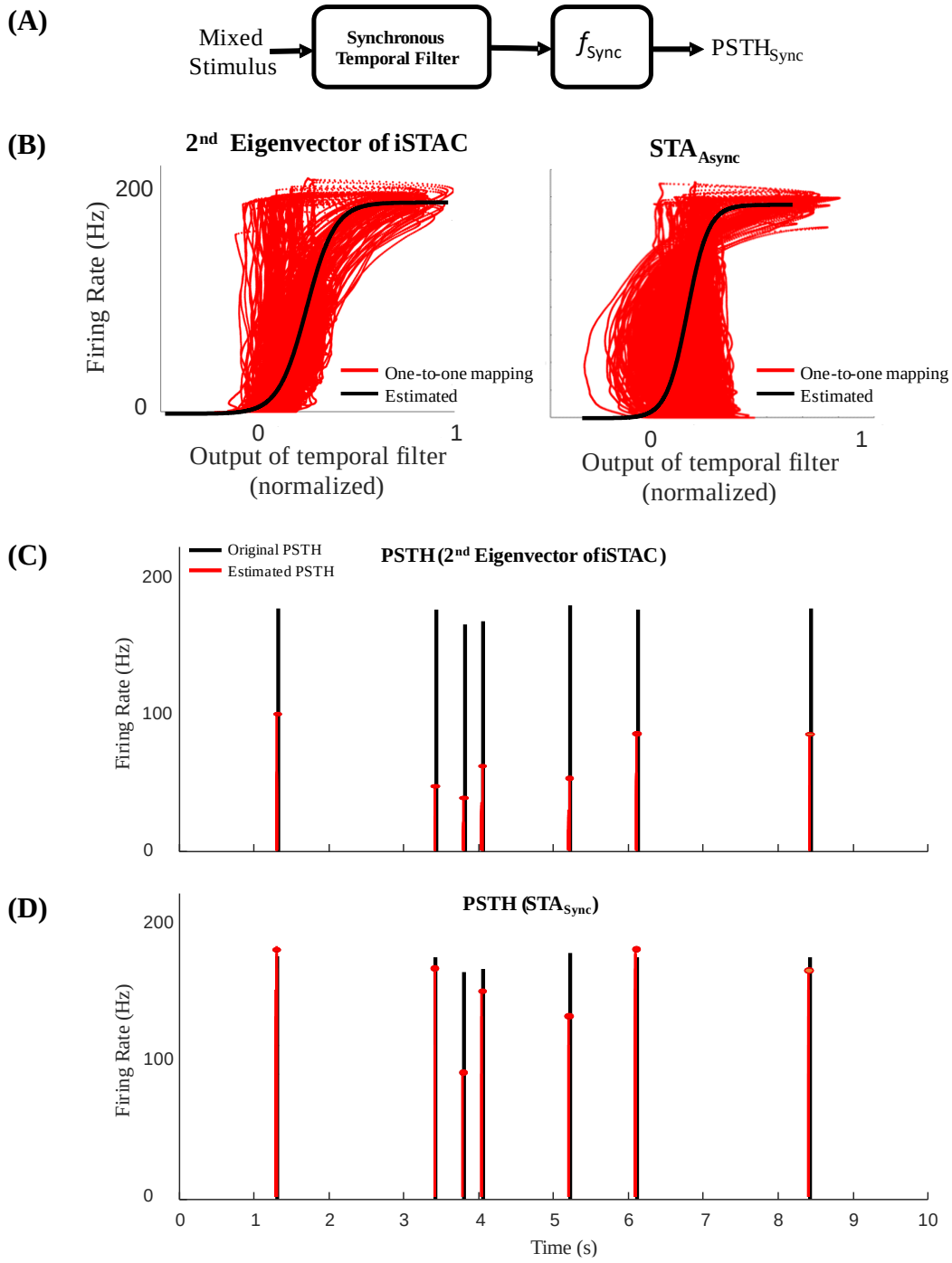

Figure 5. Static nonlinearities underlying synchronous spikes. (**A**) Blog diagram of LNL model for synchronous spikes. (**B**) Static nonlinearity calculated for the synchronous spikes is obtained by mapping the output of filtered stimulus to the instantaneous synchronous events (calculated by a narrow kernel, σ = 1 msec). Static nonlinearity calculated based on 2nd eigenvectors of the iSTAC method, *V2*, (Left) and STA_{Sync} (Right). The solid black shows fitted sigmoid functions. (**C**) The PSTHs constructed by the fitted nonlinearities based on *V2* were drawn against the PSTH of synchronous spikes. **(D)** The PSTHs constructed by the fitted nonlinearities based on *V2* were drawn against the PSTH of synchronous spikes.

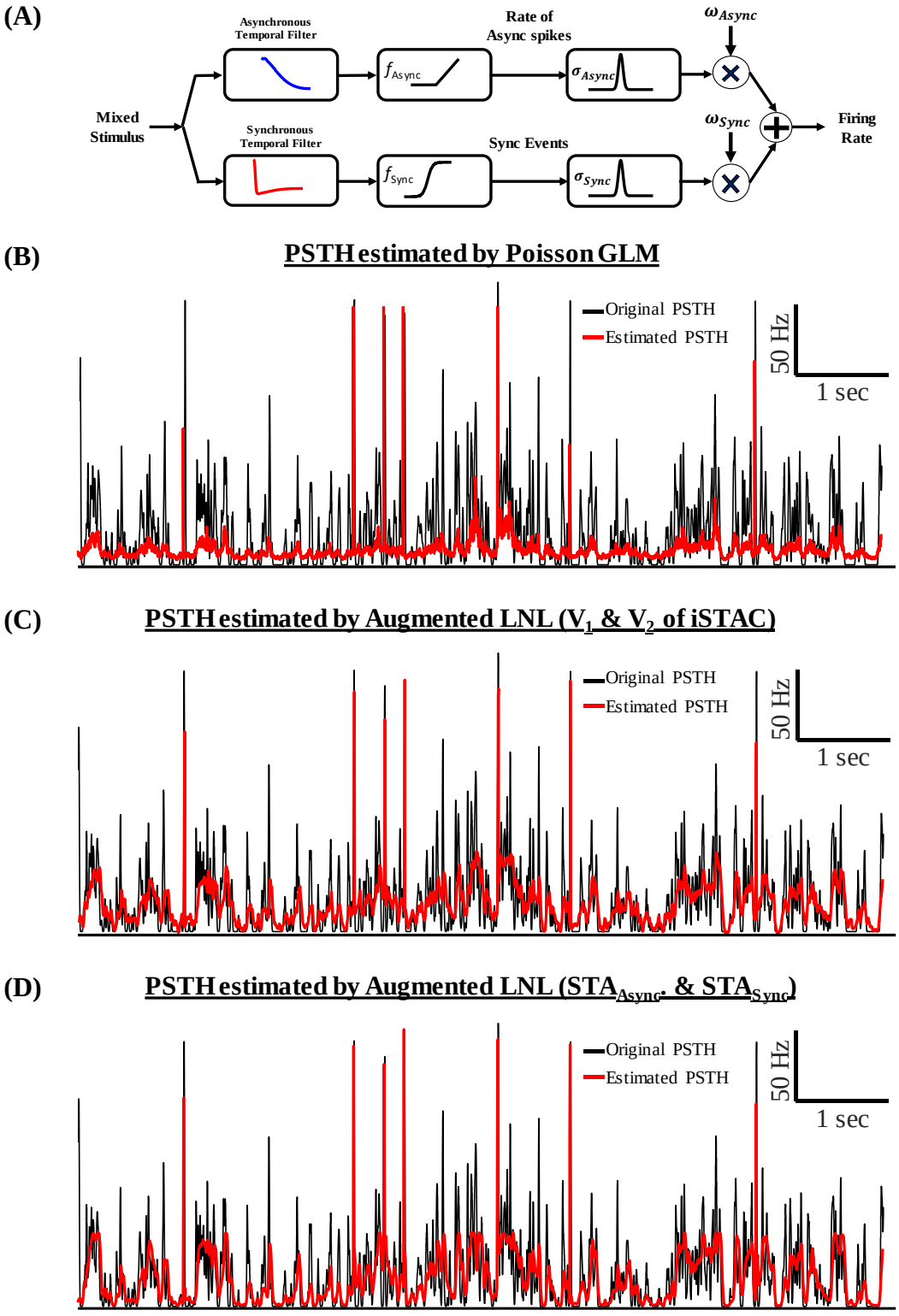

Figure 6. Two stream LNL model, referred to as augmented LNL model, enables co-existence of temporal- and rate-codes. (**A**) Blog diagram of the augmented LNL model for combining rate of asynchronous spikes and events of synchronous spikes. (**B**) The PSTHs estimated by a conventional Poisson GLM (red) were shown against the original PSTH (calculated by 1 msec Gaussian kernel). **(C)** The PSTHs estimated by the segmented LNL using temporal filters of iSTAC method. **(D)** The PSTHs estimated by LNL using the segmented LNL using STA_{Async} and STA_{Sync}.

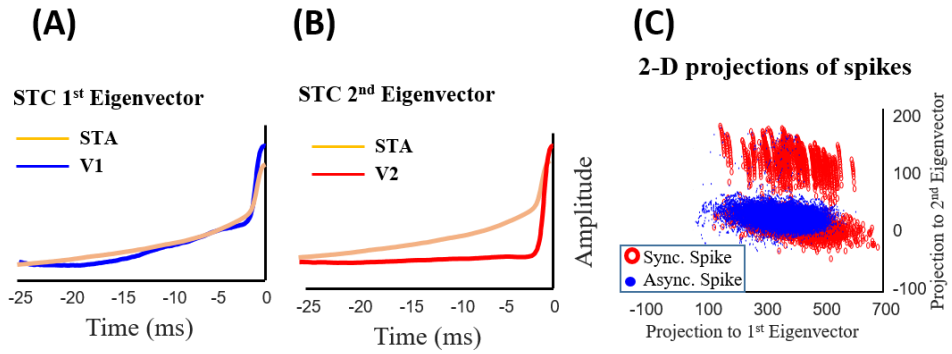

Figure 7. Slow and fast features decomposition of a mixed signal of a homogeneous ensemble of neurons using the STC method. (**A**) The projection of spike-triggered mixed signal onto the 1st eigenvector and (**B**) 2nd eigenvector of the STC matrix. (**C**) The 1st and 2nd eigenvectors of the STC matrix, *V1* and *V2*, respectively, are shown against the spike-triggered average (STA) calculated using all spikes.

References

1. Li, Y.G., J.H. Winters, and N.R. Sollenberger, *MIMO-OFDM for wireless communications: signal detection with enhanced channel estimation.* IEEE Transactions on communications, 2002. **50**(9): p. 1471-1477.
2. Laughlin, S.B. and T.J.J.S. Sejnowski, *Communication in neuronal networks.* 2003. **301**(5641): p. 1870-1874.
3. Johnson, D.H., *The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones.* The Journal of the Acoustical Society of America, 1980. **68**(4): p. 1115-1122.
4. Harvey, M.A., et al., *Multiplexing stimulus information through rate and temporal codes in primate somatosensory cortex.* PLoS Biol, 2013. **11**(5): p. e1001558.
5. Lankarany, M., et al., *Differentially synchronized spiking enables multiplexed neural coding.* 2019. **116**(20): p. 10097-10102.
6. Sullivan, W. and M. Konishi, *Segregation of stimulus phase and intensity coding in the cochlear nucleus of the barn owl.* Journal of Neuroscience, 1984. **4**(7): p. 1787-1799.
7. Saal, H.P., M.A. Harvey, and S.J. Bensmaia, *Rate and timing of cortical responses driven by separate sensory channels.* Elife, 2015. **4**: p. e10450.
8. Saal, H.P. and S.J. Bensmaia, *Touch is a team effort: interplay of submodalities in cutaneous sensibility.* Trends in neurosciences, 2014. **37**(12): p. 689-697.
9. Paninski, L., *Maximum likelihood estimation of cascade point-process neural encoding models.* Network: Computation in Neural Systems, 2004. **15**(4): p. 243-262.
10. Latimer, K.W., et al., *Multiple timescales account for adaptive responses across sensory cortices.* Journal of Neuroscience, 2019. **39**(50): p. 10019-10033.
11. Churchland, M.M., et al., *Techniques for extracting single-trial activity patterns from large-scale neural recordings.* Current opinion in neurobiology, 2007. **17**(5): p. 609-618.
12. Paninski, L., J. Pillow, and J.J.P.i.b.r. Lewi, *Statistical models for neural encoding, decoding, and optimal stimulus design.* 2007. **165**: p. 493-507.
13. Rezaei, M.R., M.R. Popovic, and M. Lankarany, *A Time-Varying Information Measure for Tracking Dynamics of Neural Codes in a Neural Ensemble.* Entropy, 2020. **22**(8): p. 880.
14. Prescott, S.A., et al., *Pyramidal neurons switch from integrators in vitro to resonators under in vivo-like conditions.* Journal of neurophysiology, 2008. **100**(6): p. 3030-3042.
15. Paninski, L., *Convergence properties of three spike-triggered analysis techniques.* Network: Computation in Neural Systems, 2003. **14**(3): p. 437-464.
16. Schwartz, O., et al., *Spike-triggered neural characterization.* Journal of vision, 2006. **6**(4): p. 13-13.
17. Pillow, J.W. and E.P. Simoncelli, *Dimensionality reduction in neural models: an information-theoretic generalization of spike-triggered average and covariance analysis.* Journal of vision, 2006. **6**(4): p. 9-9.
18. Lankarany, M. and S.A. Prescott, *Multiplexed coding through synchronous and asynchronous spiking.* BMC Neuroscience, 2015. **16**(1): p. 1-2.
19. Bojak, I. and T. Nowotny, *27th Annual Computational Neuroscience Meeting (CNS* 2018): Part Two*. 2018, BMC Neuroscience.
20. Ramachandran, P., B. Zoph, and Q.V. Le, *Searching for activation functions.* arXiv preprint arXiv:1710.05941, 2017.

21. Pillow, J.W., et al., *Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model.* Journal of Neuroscience, 2005. **25**(47): p. 11003-11013.
22. Pillow, J.W., et al., *Spatio-temporal correlations and visual signalling in a complete neuronal population.* Nature, 2008. **454**(7207): p. 995-999.
23. Paninski, L., *Estimation of entropy and mutual information.* Neural computation, 2003. **15**(6): p. 1191-1253.
24. Sharpee, T., N.C. Rust, and W. Bialek, *Analyzing neural responses to natural signals: maximally informative dimensions.* Neural computation, 2004. **16**(2): p. 223-250.
25. Moskovitz, T.H., N.A. Roy, and J.W. Pillow, *A comparison of deep learning and linear-nonlinear cascade approaches to neural encoding.* BioRxiv, 2018: p. 463422.
26. Rezaei, M.R., et al. *A Comparison Study of Point-Process Filter and Deep Learning Performance in Estimating Rat Position Using an Ensemble of Place Cells*. in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2018. IEEE.
27. Amidi, Y., et al., *Parameter Estimation in Multiple Dynamic Synaptic Coupling Model Using Bayesian Point Process State-Space Modeling Framework.* Neural Computation, 2021. **33**(5): p. 1269-1299.
28. Fard, R.S., et al., *Analysis of Distributed Neural Synchrony through State-Space Coherence Analysis.* bioRxiv, 2020.
29. Morris, C. and H.J.B.j. Lecar, *Voltage oscillations in the barnacle giant muscle fiber.* 1981. **35**(1): p. 193-213.
30. Ratté, S., et al., *Impact of neuronal properties on network coding: roles of spike initiation dynamics and robust synchrony transfer.* 2013. **78**(5): p. 758-772.
31. Pirschel, F. and J. Kretzberg, *Multiplexed population coding of stimulus properties by leech mechanosensory cells.* Journal of Neuroscience, 2016. **36**(13): p. 3636-3647.
32. Destexhe, A., M. Rudolph, and D.J.N.r.n. Paré, *The high-conductance state of neocortical neurons in vivo.* 2003. **4**(9): p. 739-751.
33. Prescott, S.A., Y. De Koninck, and T.J. Sejnowski, *Biophysical basis for three distinct dynamical mechanisms of action potential initiation.* PLoS Comput Biol, 2008. **4**(10): p. e1000198.

Appendix A: The augmented LNL as a Poisson GLM

We assume the number of spikes in time are discrete events. By dividing the time horizon of the experiment, $(0, T]$, into K (K is a large number) subintervals $(t_{k-1}, t_k]_{k=1}^{K}$ and consider N_k as number of the events in the time interval $(t_{k-1}, t_k]$. We can model the spike observation process with a point process by inhomogeneous Poisson distribution and parameter λ_k as:

$$P(N_k|s_k) = \frac{1}{N_k!}\,(\Delta\lambda_k)^{N_k} e^{-\Delta\lambda_k} \tag{A1}$$

Where Δ is the width of the time-bins. By assuming that number of spikes are conditionally independents in time, we can write the whole observation process as

$$P(N_{1:K}|s_{1:K}) = \prod_{k=1}^{K} \frac{1}{N_k!}\,(\Delta\lambda_k)^{N_k} e^{-\Delta\lambda_k} \tag{A2}$$

In other hand, we can model the λ_k in a way that captures effects of two features of the stimulus as linear combination of their effects in λ_k (discussed in the Result section), as

$$\lambda_k = \sum_{i=1}^{D} \omega_i f_i(\mu_i * s_k) \tag{A3}$$

where, $\theta = \{\omega_1, \dots, \omega_D\}$ is our parameters set and D is the dimension of the feature map. Finally by using maximum-likelihood estimation we can tune the model parameters

$$\hat{\theta} = arg\ log\ P(N_{1:K}|s_{1:K}) \rightarrow$$

$$log\ P(N_{1:K}|s_{1:K}) = \sum_{k=1}^{K} log\,\frac{1}{N_k!} + \sum_{k=1}^{K} N_k\ log\ \Delta + N_k\left(\sum_{i=1}^{D} \omega_i f_i(\mu_i * s_{1:k})\right)$$

$$-\Delta \sum_{i=1}^{D} \omega_i f_i(\mu_i * s_{1:k}) \tag{A4}$$

Based on Jensen's inequality and considering that $log\ log\ x$ is a concave function, we have:

$$log\ P(N_{1:K}|s_{1:K}) \geq C + \sum_{k=1}^{K} N_k \left(\sum_{i=1}^{D}\left(\omega_i f_i(\mu_i * s_{1:k})\right) - \Delta\omega_i f_i(\mu_i * s_{1:k})\right) =$$

$$Q\ ,\ C = \sum_{k=1}^{K} log\ \frac{1}{N_k!} + \sum_{k=1}^{K} N_k\ log\ \Delta + N_k \tag{A5}$$

where Q is a lower bound for the Log-likelihood. So, we can find the θs by maximizing the Q over them as

$$\widehat{\theta_{ML}} = \frac{\partial log\ P(N_{1:K}|s_{1:K})}{\partial\theta} = 0 \rightarrow$$

$$\frac{\partial log\ P(N_{1:K}|s_{1:K})}{\partial\theta} = 0 + \sum_{k=1}^{K}\left(\frac{N_k}{\omega_i} - \Delta f_i(\mu_i * s_{1:k})\right) = \sum_{k=1}^{K} N_k - \frac{1}{\omega_i}\sum_{k=1}^{K} \Delta f_i(\mu_i *$$

$$s_{1:k}) = 0 \rightarrow \widehat{\omega_i} = \frac{\sum_{k=1}^{K} N_k}{\sum_{k=1}^{K} \Delta f_i(\mu_i * s_{1:K})} \tag{A6}$$

By estimating the model parameters, we reach a model for encoding the firing rate of the multiplexed spikes.

Appendix B: Modeling nonlinearity (f(x)) by ex-quadratic functions

The estimators do dimensionality reduction task and map high dimensional sensory stimulus to a lower dimensional linear feature map $= K^T s$, K is the basis of the feature map space. Based on the definition of GLM, mentioned above, we still need to find optimum model for $f(x)$, $f: R^d \rightarrow R$. By considering motivation in [4], a reasonable way is using exponential general quadratic function:

$$f = exp\left(\frac{1}{2}x^T Cx + b^T x + a\right) \quad \text{(B1)}$$

where C is a symmetric matrix, b is a vector, and a is a scalar. So, now we can use maximum-likelihood to optimize the parameters set, $\{C, b, a\}$. To do that we need to maximize the log-likelihood of observing spike given all spikes and the parameters set ($L = log\ log\ P(r_{1:N}|s_{1:N}, C, b, a)$). By assuming that spikes firing in time are independent then we can rewrite it as:

$$L = \frac{1}{n_{sp}}\sum_i log\ \ P(r_i|s_i, C, b, a) \quad \text{(B2)}$$

where n_{sp} is total number of spikes, so our objective is to maximize L by finding best parameters set:

$$\{\hat{C}, \hat{b}, \hat{a}\} = argmax_{\{C,b,a\}}\ L \quad \text{(B3)}$$

By following the optimization steps in [5] and assuming that the stimulus are drawn from $x \sim N\ (0, \Phi)$; the maximum-likelihood estimation of the parameters are:

$$\hat{C} = \Phi^{-1} - \Lambda^{-1}, \hat{b} = \Lambda^{-1}\mu \quad \text{(B4.a)}$$

$$\hat{a} = log\left(\frac{nsp}{N}|\Phi\Lambda^{-1}|^{0.5}\right) - \frac{1}{2}\mu^T\Phi^{-1}\Lambda^{-1}\mu \quad \text{(B4.b)}$$