eNeuro

Research Article: Methods/New Tools | Novel Tools and Methods

Natural statistics as inference principles of auditory tuning in biological and artificial midbrain networks

https://doi.org/10.1523/ENEURO.0525-20.2021

Cite as: eNeuro 2021; 10.1523/ENEURO.0525-20.2021

Received: 1 December 2020
Revised: 10 March 2021
Accepted: 27 April 2021

Manuscript Title Page

1. Manuscript Title (50 word maximum)

Natural statistics as inference principles of auditory tuning in biological and artificial midbrain networks

2. Abbreviated Title (50 character maximum)

Auditory tuning in biological and artificial brain

3. List all Author Names and Affiliations in order as they would appear in the published article

Sangwook Park, Department of Electrical and Computer Engineering, Johns Hopkins University

Angeles Salles, Department of Psychological and Brain Sciences, Johns Hopkins University

Kathryne Allen, Department of Psychological and Brain Sciences, Johns Hopkins University

Cynthia F. Moss, Department of Psychological and Brain Sciences, Johns Hopkins University

Mounya Elhilali, Department of Electrical and Computer Engineering, Johns Hopkins University

4. Author Contributions:

Designed research SP, ME. Performed research: SP, AS, KA. Analyzed data: SP. Wrote original draft: SP. Revised manuscript: SP, AS, KA, CFM, ME. Funding: CM, ME.

5. Correspondence should be addressed to (include email address)

Mounya Elhilali (E-mail: mounya@jhu.edu)

6. Number of Figures: 14

7. Number of Tables: 1

8. Number of Multimedia: 0

9. Number of words for Abstract: 250

10. Number of words for Significance Statement: 120

11. Number of words for Introduction: 763

12. Number of words for Discussion: 1121

13. Acknowledgements

We thank Dr. Kirsten M Bohn for her assistance with the natural calls database.

14. Conflict of Interest

Authors report no conflict of interest

15. Funding sources

This work was funded by Brain Initiative (NSF-FO 1734744 (2017–2021) and ONR (N000141712736, N000141912014, N000141912689). A. Salles was supported by a Human Frontiers Science Program Long-term Postdoctoral Fellowship (LT000220/2018), and K. Allen was supported by an NIH Institutional Training Grant postdoctoral fellowship (5T32DC000023-35, PIs Kathleen Cullen and Paul Fuchs).

Natural statistics as inference principles of auditory tuning in biological and artificial midbrain networks

Abstract

Bats provide a powerful mammalian model to explore the neural representation of complex sounds, as they rely on hearing to survive in their environment. The inferior colliculus (IC) is a central hub of the auditory system that receives converging projections from the ascending pathway and descending inputs from auditory cortex. In this work, we build an artificial neural network to replicate auditory characteristics in IC neurons of the big brown bat. We first test the hypothesis that spectro-temporal tuning of IC neurons is optimized to represent the natural statistics of conspecific vocalizations. We estimate spectro-temporal receptive fields (STRF) of IC neurons and compare tuning characteristics to statistics of bat calls. The results indicate that the FM tuning of IC neurons is matched with the statistics. Then, we investigate this hypothesis on the network optimized to represent natural sound statistics and to compare its output with biological responses. We also estimate biomimetic STRF's from the artificial network and correlate their characteristics to those of biological neurons. Tuning properties of both biological and artificial neurons reveal strong agreement along both spectral and temporal dimensions, and suggest the presence of nonlinearity, sparsity and complexity constraints that underlie the neural representation in the auditory midbrain. Additionally, the artificial neurons replicate IC neural activities in discrimination of social calls, and provide simulated results for a noise robust discrimination. In this way, the biomimetic network allows us to infer the neural mechanisms by which the bat's IC processes natural sounds used to construct the auditory scene.

Significance Statement

Recent advances in machine learning have led to powerful mathematical mappings of complex data. Applied to brain structures, artificial neural networks can be configured to explore principles underlying neural encoding of complex stimuli. Bats use a rich repertoire of calls to communicate and navigate their world, and the statistics underlying the calls appear to align with tuning selectivity of neurons. We show that artificial neural network with a nonlinear, sparse and deep architecture trained on the statistics of bat communication and echolocation calls results in a close match to neurons from bat's inferior colliculus. This tuning optimized to yield an effective representation of spectro-temporal statistics of bat calls appears to underlie strong selectivity and noise invariance in the inferior colliculus.

1. Introduction

Biological neural circuits are believed to provide an efficient code of the sensory world, which allow us to process complex and dynamic stimulus information from our surroundings. Perception of an auditory scene is created by neural activity filtered through several stages of feed-forward and feedback sensory processing. Sound pressure of an acoustic signal is first transduced into a bio-electrical signal in the cochlea. Subsequently, the bio-electrical signal is relayed through the auditory pathway. The inferior colliculus (IC) is an auditory hub that receives ascending inputs from brainstem nuclei and sends information through the thalamus to the auditory cortex, while it also receives descending inputs from auditory cortex (Casseday et al., 2002). The IC encodes complex auditory features such as frequency sweep rate (Williams and Fuzessery, 2010) and patterning (Gordon and O'Neill, 1998) that are necessary for identification of complex auditory objects and therefore plays a key role in representing these objects in a natural listening environment.

Echolocating bats build a representation of their surroundings by emitting ultrasonic vocalizations and processing the features of returning echoes to compute the location and features of targets and obstacles in the environment. Bats must rapidly process sonar echoes while concurrently parsing environmental noise and calls emitted by conspecifics. In this complex and rapidly changing auditory scene, the bat's brain efficiently encodes acoustic stimuli and allows the animal to accurately track prey, avoid obstacles, and communicate with conspecifics while dynamically navigating a three-dimensional environment. Humans and other animals face similar challenges in the course of their natural acoustic behaviors. With the goal of elucidating principles underlying auditory scene analysis in the midbrain, we examine the relationship between statistics of the rich acoustic repertoire of bat calls and neural response patterns in the bat's IC to explore artificial networks tuned to map natural statistics in these calls and identify emergent properties that match responses in the IC.

Here, we test the hypothesis that the bat's auditory midbrain is optimized to accurately represent the natural statistics in the sounds and echoes that exist in the bat's environment (particularly social and echolocation calls). Past research has suggested that the IC plays a major role in the representation and mapping of communication sounds that give rise to specialized encoding of natural sounds along the ascending auditory system (Aitkin et al., 1994; Suta et al., 2003). An earlier study in the Mexican free tailed bat suggested a possible correspondence between tuning characteristics of individual IC neurons and properties of natural calls from conspecific sounds (Andoni et al., 2007; Brimijoin and O'Neill 2005). In the current study, we corroborate this relationship in a different species and further probe constraints and implications of such optimal encoding of natural sounds on auditory signal processing in a complex scene.

We recorded vocalizations from socially housed bats and analyzed the spectro-temporal statistics of natural sounds (e.g. frequency modulation (FM) velocity, directionality). Using the database of collected statistics, we built an artificial network, which projects sounds onto a latent space that efficiently represents statistics of these natural sounds in a strategy of signal reconstruction (Smith and Lewicki, 2006). This computational model offers a biomimetic architecture whose main operation is to capture the statistics of natural bat calls, without information about the function of biological neurons. We then ask: Does the emergent tuning of this artificial network match properties of biological neurons in the big brown bat inferior colliculus? To answer this question, we also recorded responses to sound stimuli, spectro-temporal ripples, from individual neurons in the IC of big brown bats.

It is known that the spectro-temporal receptive fields (STRF's) suggest a reasonable linear-approximation of neural responses as a transfer function from acoustic stimuli and those are usually used to explore auditory characteristics of the neurons (Andoni et al., 2007; Depireux et al., 2001; Elhilali et al., 2013). We extracted STRF's from IC and artificial neurons and calculated auditory characteristics from these neural response functions (Kowalski et al., 1996; Poon and Yu, 2000). The spectro-temporal tuning characteristics of biological neurons were then compared to both the statistics of natural calls as well as emergent tuning of artificial neurons. By varying the configuration of the artificial network, we employed the theoretical network as springboard to examine possible constraints on the configuration of midbrain networks, and gauge the validity of the hypothesis linking biological encoding in the mammalian midbrain to efficient representation of natural sound statistics. While various artificial neural networks can be optimized to reconstruct an input sound from compressed feature on latent space, finding an architecture that closely emulates the biological network provides insights into the underlying functional role of certain brain nuclei. Here, we examine the relationship between the optimal encoding of natural statistics in bat calls and its role in facilitating robust selectivity across sound classes in the repertoire. The graphical abstract in Fig. 1 shows an overview of the approach taken in this work.

2. Materials and Methods

2.1. Collection of Bat's vocalization

2.1.1. Animals

Big brown bats (*Eptesicus fuscus*) were collected from an exclusion site under a state permit. All experimental procedures were carried out in accordance with a protocol approved by an Institutional Animal Care and Use Committee. A total of approximately 100 bats were housed in our Lab and used for vocal data recordings, and four (2 male, 2 female) bats were used for neurophysiological data collection.

2.1.2. Audio recordings for training the biomimetic network

A bat call library was built from audio recordings of bats housed in a vivarium room where the temperature is kept at 70-80 ^{o}F, and humidity is kept at 30-70 %. This room holds approximately 100 bats in groups of 1-6 separated in mesh cages. The recordings were made for two days using an Avisoft CM16/CMPA ultrasonic microphone and the Avisoft-RECORDER software. Mono audio was recorded at a sampling rate of 300 kHz.

Natural call recordings from big brown bats were processed to extract meaningful segments. An energy-based signal activity detection was performed on the entire database to remove the silences between calls and to split the recordings into segments containing bat calls (Park et al., 2014). As a result, we constructed species specific databases containing 17,713 calls (about 10 min) for big brown bats. This call database was used for training artificial networks. The data was divided into a training set (15,000 randomly selected calls) to learn network parameters and test set (remaining 2,713 calls) for verifying the network.

2.1.3. Social calls for natural sound representation

To investigate discriminability in the artificial network, we used a social call database that includes 26 audio clips for 8 different types of bat calls (Fig. 2). These types include six calls, as defined in (Wright et al., 2013), specifically, Echolocation (Echo), Frequency Modulated Bout (FMB), Upward Frequency Modulated (UFM), Long Frequency Modulated (LFM), Short Frequency Modulated (SFM), and Chevron-Shaped (CS); in addition to two additional calls types, Long-Wave and Hook, which resemble a hook in time-frequency space. All audio clips were up-sampled from 250 kHz to 300 kHz.

2.2. Neurophysiological Inferior Colliculus data

Recordings of neural responses from IC neurons were used to perform two separate analyses: (1) characterize receptive field tuning of IC neurons; and (2) examine discriminability of IC neurons to different con-specific calls. Methods for receptive field analysis are described next in section 2.2.4, while data used for discriminability analysis are described in section 2.2.5.

2.2.1. Receptive field recordings

A head-post was adhered to the skull of bats for head fixation as described in (Macias et al., 2018). The inferior colliculus was located using skull and brain landmarks and a surgical drill was used to make a ≤ 1 mm diameter craniotomy preserving dura. The neurophysiological recordings were performed in a sound-attenuating and electrically shielded chamber (Industrial Acoustics Company, Inc.). Each bat was restrained individually in a custom-made foam mold and the head was fixed by the head-post. Recording sessions were carried out over 3 to 5 consecutive days, each one lasting no more than 4 hours. Water was offered to the bats every 2 hours. No drugs were administered during recordings. During recordings a silver wire for grounding was placed in between muscle and skull about 5mm rostral to the craniotomy site. The 16-channel recording probe (Neuronexus A1x16-5mm-50-177-A16) was inserted into the brain using a micromanipulator. The surface of the brain was registered as 0 μm for depth reference and the probe was advanced in 10 μm steps using a hydraulic microdrive (Stoelting Co.). Recordings were taken at least 100 μm apart. An OmniPlex D Neural Data Acquisition System recording system (Plexon, Inc.) was used to obtain neural responses with 16-bit precision and 40 kHz sampling rate. A transistor-transistor-logic (TTL) pulse for each stimulus presentation was generated with the National Instrument card used for stimulus presentation and was recorded on channel 17 of the analog channels of the acquisition system for synchronization of acoustic stimuli and neural recordings. The stimuli were recorded on channel 18 of the acquisition system to corroborate synchronization.

2.2.2. Moving ripple stimuli

A set of ripple stimuli was generated to estimate STRFs of IC neurons (Kowalski et al., 1996; Depireux et al. 2001; Andoni et al., 2007). Ripples are modulated noise stimuli that are dynamic both in time and frequency. Each ripple can be described as

$$S(t,x) = 1 + \Delta A \times \sin(2\pi(\omega t + \Omega x) + \phi) \qquad (1)$$

where t and x are indices for time and octave scaled frequency. ΔA and ϕ are amplitude and a phase, respectively. And ω and Ω represent modulation rates along temporal (Hz) and spectral (cyc/oct) axes. The temporal and spectral modulation parameters were varied from -176 - 176 Hz in steps of 32 Hz and 0.0 - 1.5 cyc/oct in steps of 0.15 cyc/oct spectrally (Fig. 3A). Each ripple spanned 6.66 octaves from 1.2 kHz to 121 kHz and was 300 ms in duration.

2.2.3. Audio playbacks for neural recordings

Extracellular recordings from the inferior colliculus of awake animals were taken while they passively listened to broadcast of either ripple stimuli, or pure tones at 70 dB. All stimuli were generated at a sampling rate of 250 kHz using a National Instruments card (PXIe 6358) and transmitted with a calibrated custom-made electrostatic ultrasonic loudspeaker connected to an audio amplifier (Krohn-Hite 7500). The loudspeaker was placed at 60 cm (for all ripple and pure tones stimuli)} from the bat's ear. The frequency response of the loudspeaker was compensated by digitally filtering the playback stimuli with the inverse impulse response of the system as described in (Luo and Moss, 2017).

Frequency tuning curves were built by recording neural responses to pure tones of 5 ms duration (with 0.5 ms ramping rise and fall). The tones ranged between 20 and 90 kHz (in 5 kHz steps) and the sound pressure levels ranged from 20 to 70 SPL (10 dB steps). At each recording site first, we played 20 repetitions of the randomized ripple stimulus and then 15 repetitions of each of the randomized pure tones at a different SPL.

2.2.4. Analysis of neuronal responses

For the analysis of auditory tuning in response to ripple and pure tone stimuli, responses were sorted offline, then single units were detected using the program 'Wave_clus' (Quiroga et al., 2004). Each individual waveform was inspected and the acceptance threshold for clusters was less than 10% of spikes with < 3 ms inter-spike interval, consistent with the neuronal refractory period. Any sites that showed no response to ripple stimuli were excluded from the spike sorting and further analysis in line with procedures used in other studies (Poon and Yu, 2000; Escabi and Schreiner, 2002; Andoni et al., 2007). After spike sorting, the Euclidian distance error between the mean and variance of number of spikes across trials was computed. Units whose error is less than 1.0 were selected for further analysis, following a Poisson model of spike representation (Corrado et al., 2005; Schwartz et al., 2006). This analysis resulted in 108 single units used for the current study.

Neurophysiological STRFs: At each recording site, ripple stimuli were repeated 10-20 times in a randomized order for each repetition. A PST histogram was calculated from the spike time sequence of each ripple; then histograms were folded into 32-point periods. The strength and phase of the response to each ripple were estimated directly from the fundamental component obtained by applying a 32-point Fast Fourier Transform (FFT) to the period histogram. Magnitude and phase responses to each ripple were combined together into a magnitude matrix $M(\Omega,\omega)$ and a phase matrix $\Phi(\Omega,\omega)$, respectively. To derive a Ripple Transfer Function (RTF), which is a representation of a STRF in the modulation domain, $M(\Omega,\omega)$ and $\Phi(\Omega, \omega)$ were expanded to four quadrants in the modulation domain spanning from -176 Hz and -1.5 cyc/oct to 176 Hz and 1.5 cyc/oct as $M_e(\Omega,\omega) = M_e^*(-\Omega,-\omega) = M(\Omega,\omega)$ and $\Phi_e(\Omega,\omega) = \Phi_e^*(-\Omega,-\omega) = \Phi(\Omega,\omega)$ based on a symmetric property around the origin (Depireux et al., 2001; Andoni et al., 2007). As a result, the RTF was formulated as

$$T(\Omega,\omega) = M_e(\Omega,\omega)e^{j\Phi_e(\Omega,\omega)} \qquad (2)$$

where $j = \sqrt{-1}$. Finally, a STRF was obtained by performing 2D inverse FFT on the RTF as

$$STRF(x,t) = F_{t,-x}^{-1}[T(\Omega,\omega)] \qquad (3)$$

where F^{-1} designates the 2D inverse FFT along each axis in the modulation domain.

2.2.5. Neural discriminability of con-specific calls

In order to examine selectivity of IC neurons to calls from the bat's natural repertoire, we re-used neural data previously collected in an earlier study (Salles et al., 2020), where we collected neuronal responses to Echolocation calls (Echo) versus Frequency-Modulated Bout (FMB) social calls. The study followed the same methodology for data collection as described here. 'Wave_clus' was used to detect and classify single units from the recordings. The spikes responding to either FMB or Echo were counted in windows of 25 ms duration, starting 5ms after stimulus onset. Some units with an average of less than five spikes over 20 times recordings were excluded because they were considered as a non-responsive unit to the stimulus. Multi-unit activity was determined from inter-spike intervals with < 3ms that were inconsistent with neuronal refractory period; and units with greater than 10% of spikes with < 3ms inter-spike interval were excluded from analysis. As a result, total 575 units were finally obtained and their responses are used in the present work to contrast neural discriminability between Echo and FMB calls with artificial neurons.

2.3. Responses in artificial neurons

2.3.1. Artificial network front-end processing

To develop a biomimetic architecture, a biologically-inspired auditory spectrogram is used as input for the network (Shamma, 1985a; Shamma, 1985b; Yang et al., 1992; Wang and Shamma, 1994). The auditory spectrogram incorporates four processing stages that emulate peripheral processing in the mammalian system: cochlear filtering, auditory-nerve transduction, hair cell responses, and lateral inhibition (Chi et al., 2005). Briefly, an incoming acoustic waveform is analyzed along a bank of constant-Q filters spanning a logarithmic scale. Then, each frequency channel undergoes a high-pass, nonlinear compression and low-pass filtering followed by lateral inhibition across frequency, following the implementation available in the NSL toolbox (Chi and Shamma, 2005) with the following settings: The frame length was set to 0.2 ms without overlap, and each octave was represented with 24 channels (i.e. 128 channels over 5.33 octaves). Octave-scaled center-frequencies were represented as $f_c = 440 \times 2^{((c-32)/24+\gamma)}$ where f_c is a center frequency of the c^{th} channel, and γ is a constant factor of octave shift ($\gamma = 4.38$). Inputs to the artificial network were sampled as square patches of the spectrogram spanning 128 frequency channels (i.e. 5.33 octaves) and 160 time-samples (i.e. 32 ms).

2.3.2. Structure of artificial network

An artificial neuron, i.e. node mimicking a biological neuron, is mathematically modelled by a linear combination of pre-node outputs and a non-linear activation function. An artificial network is constructed by connecting a large number of nodes to each other. Using nonlinear activation functions enables the network to perform nonlinear computations on feedforward propagation. For this study, we favored a generative architecture using an autoencoder composed of an encoder, which compresses original data into a compact code; and a decoder, which reconstructs the original signal from that code (Baldi, 2012; Doersch, 2016). The intuition is to directly test our hypothesis that the network would infer a statistical model of the training dataset of natural calls, and if successful should allow a faithful reconstruction of the inputs.

The proposed architecture is shown in Fig. 4. First an encoder stage ***E*** is composed of convolutional layers, pooling layers, and a fully connected layer. A latent vector represents compressed features learned from the input data. A decoder stage ***D*** composed of reverse operations using transposed convolutions, reconstructs the input features from a latent vector. A sampling stage, interposed between the encoder and decoder, emulates neural activity yielding sparse binary activations.

Using the same general building block composed of convolution and pooling layers, this study investigates various configurations of the network by varying: (1) *depth*, which is the number of blocks. In Fig. 4, the black-flow shows a double stacking structure as an example. A deeper network can be constructed by stacking more blocks, on the other hand, a shallow network can be created by removing a block; (2) *nonlinearity*, by varying the slope of nonlinear activation function employed; and (3) *sparsity*, by controlling the density of sampling in the latent space.

The encoder architecture ***E*** follows a convolutional neural network (CNN) framework in order to reduce the number of trainable parameters, hence controlling for over-fitting issues and generalizability to unseen data (Dietterich, 1995). The convolutional layers compute output feature maps using 2D convolutions between input feature maps and several filters as

$$I_o^l[f,t,k] = \sum_{\xi,\tau,m} I_i^l[\xi,\tau,m]\, f^l[\xi - f, \tau - t, m, k] \tag{4}$$

where f, t, l, k and m are indices for spectral, temporal, layer, channel of output feature map, and channel of input feature map respectively. I_i, I_o, and f^l are feature maps for input and output, and convolutional filter applied in the l^{th} layer, respectively. Multi-scale filters are employed in each convolutional layer to balance broad span (in time and frequency) vs. localized analyses. Then, output feature maps concatenate filter outputs using multi-sized filters (Fig. 5A) (Szegedy et al., 2015). Specifics of both filter composition and dimensions of intermediate feature maps are summarized in Table 1. Neural activation by an acoustic feature is emulated by applying a nonlinear function after convolution as

$$I_a^l[f,t,k] = \max(I_o^l[f,t,k],\ \alpha \times I_o^l[f,t,k]) \tag{5}$$

where α is a constant within an interval [0, 1] (Maas et al., 2013). Next, pooling layers compress the output from the previous convolutional layer by extracting a maximum among some values enclosed by a non-overlapping window (i.e. max-pooling) I^l (Scherer et al., 2010). As a result, the width and height of the output are reduced by half. At the top of the encoder, a fully connected layer is applied for mapping into a latent space, which involves natural statistics requiring to reconstruct original input, as $v_c = W_l^T \times flatten(I^L)$ where I^L is a feature map in the last pooling layer, W is weight matrix in the fully connected layer, and $flatten(.)$ is a reshape function from a 3D tensor to a vector.

In the middle stage, a binary code vector v_b is generated by performing a Bernoulli sampling process. A sigmoid function is applied to the latent vector to calculate prior probabilities. Thus, the output of the middle stage is represented as $v_b = Bernoulli(\sigma(v_c))$ where $\sigma(.)$ is a sigmoid function.

The decoder ***D*** is composed of a fully connected layer and transposed convolution layers. In the fully connected layer, a latent vector is expanded into an initial space as $\hat{v} = W_l \times v_b$, and the vector $\hat{v}$ is reshaped to a 3D tensor as a set of initial feature maps as $\hat{I}^l = reshape(\hat{v})$. From initial feature maps, a transposed convolution using multi-scale filters is sequentially performed until the output has the same dimensions as the input patch (Shelhamer et al., 2017; Radford et al., 2015). Convolutional filters used in the encoder are applied for transposed convolution after transposing input channel from output channel dimension as $\dot{f}^l[f,t,k,m]$. A transposed convolution using multi-scale filters is performed in three steps (Fig. 5B). First, the input feature map $\hat{I}^l$ is split into submaps, $[\hat{I}_1^l, \hat{I}_2^l, \ldots, \hat{I}_N^l]$, as many as the number of filters. Second, transposed convolution is individually performed for each pair of submap and filter. Finally, a set of output feature maps is obtained by averaging the results of the second step.

2.3.3. Training artificial network

The network was trained using the cost function:

$$L = \frac{1}{2}\sum_n[(x_n - D(E(x_n)))^2 + \lambda(\rho - \sum_i \sigma(v_{c_i}))^2] \qquad (6)$$

where x_n is an input patch with respect to the n^{th} index, $E(.)$ represents an encoder function while $D(.)$ is for a decoder, and ρ means the average number of active nodes. The first term represents the mean square error between an input patch and its reconstruction by the autoencoder. The sparse constraint prevents overfitting as well as emulates sparsity of active neurons in the brain. Let Y be a random variable representing the number of active nodes by the Bernoulli process. Then, the distribution known as the Poisson binomial distribution is denoted as

$$\Pr(Y = \rho) = \sum_A[\prod_{i\in A}\sigma(v_{c_i})\prod_{j\in A^c}(1 - \sigma\left(v_{c_j}\right))] \qquad (7)$$

where A is a set whose elements are possible combination for choosing ρ nodes from N nodes. This distribution can be approximated by $Binomial(N, \mu/N)$ where $\mu = \sum_i \sigma(v_{c_i})$ (Choi and Xia, 2002). The network training was implemented using TensorFlow (Abadi et al., 2016). AdamOptimization was applied for an optimizer with $1.0e-4$ learning rate. And, λ was set to $1.0e-4$. For more details, readers can find the implementation on http://www.github.com/JHU-LCAP/BioSonar-IC-model /.

Comparisons between the biological neurons and artificial neurons were performed to infer the network configuration that best matches the characteristics of IC neurons (as explained next). The best configuration composed of a triple stacking network, a parameter of nonlinearity $\alpha = 0.2$ in (5), and 10% sparsity constraint in (6).

2.3.4. Biomimetic STRFs

Once trained, the network was interrogated following the same procedure as biological neurons. The same ripple stimuli were given as input to the network and activity of the nodes before applying the sigmoid activation and the Bernoulli sampling, v_c in Fig. 4 was characterized. Each ripple was transformed into an auditory spectrogram (as described earlier). A sequence of input patches for each ripple were then composed by applying a sliding window (window length: 160 frames) in every 2 ms (sliding step: 10 frames) (Fig. 3B). Input patches in the sequence were consecutively fed into the pre-trained encoder, then a latent vector v_c was obtained every 2 ms. The same procedure for extracting biological STRF's was followed (See. Section 2.2.4). To find the magnitude m and phase ϕ of the responses, we performed a 32-point Fast Fourier Transform (FFT) and derived the magnitude and unwrapped phase of the fundamental component (Fig. 3C). By repeating this procedure for all ripples, the magnitude and phase were collected in a matrix $M(\Omega, \omega)$ and a $\Phi(\Omega, \omega)$, respectively (Fig. 3D). These modulation responses were then converted into time-frequency STRF profiles by performing a 2D inverse FFT on the RTF (Fig. 6). Note that, in this study, all network architectures employed a total 100 artificial neurons (spanning a 100-dimensional latent space) so that 100-biomimetic STRFs were used for analysis.

2.4. Analysis of auditory characteristics

2.4.1. Natural statistics and Auditory characteristics

Frequency Modulation (FM) velocity (statistics of bat calls): To characterize conspecific vocalizations, we calculated FM velocities of each call segment in our database. Since moving ripples were used as bases components of the Fourier modulation domain (Singh and Theunissen, 2003), we derived auditory spectrograms of each call, then performed a 2D FFT after mean subtraction to remove constant components. $T_c(\Omega, \omega) = F_{f,t}[S(f,t) - \bar{S}]$ where $F_{f,t}[.]$ is the 2D FFT, S is an auditory spectrogram of a bat call, and $\bar{S}$ is its mean over the time and frequency axes. A velocity line was estimated by performing a line fitting on the magnitude of 2D FFT result. Finally, the FM velocity of a bat call was acquired by calculating the slope of the velocity line.

Best Velocity (BV): We defined a best velocity as the center of mass with respect to response power in a magnitude plot. To estimate the center of mass, we performed a Gaussian surface fitting on the 1st quadrant of magnitude plot. After normalization as $\bar{M}_e = M_e/[\sum_{\Omega,\omega} M_e \Delta\Omega\Delta\omega]$where $\Delta\omega$ and $\Delta\Omega$ are respectively step size of temporal and spectral modulation rate, the fitting was performed to estimate mean vector and covariance matrix, by minimizing a square mean error function as $Err = \frac{1}{2}\sum_{\Omega,\omega}(ln(\bar{M}_e) - ln(G_{\mu,\Sigma}))^2$, where $G_{\mu,\Sigma}$ is a Gaussian distribution with mean vector μ and covariance matrix Σ. By performing the Least Square Error (LSE) estimator iteratively (Kay, 1993), we derived the Gaussian mean vector and covariance matrix. Best Velocity was defined as the slope of the mean vector (Fig. 7).

Orientation (Ori): To characterize velocity selectivity, we defined orientation as the angle between a line connecting the origin to the center of mass and a dominant eigenvector of the Gaussian covariance matrix. Note that the dominant eigenvector indicates the dominant direction of magnitude spread at the center of mass (Fig. 7) (Andoni et al., 2007).

Inseparability (Ins): Singular value decomposition (SVD) is applied to each STRF for calculating inseparability (Depireux et al., 2001). This approach decomposes the STRF into a linear combination of rank-1 matrices; in other words, $STRF = \sum_i \lambda_i \boldsymbol{u}_i \boldsymbol{v}_i^H$, where $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are respectively left- and right-eigenvectors (column vector) corresponding to a singular value λ_i, and H means Hermitian transpose (Strang, 2009). Based on this definition, a STRF is called separable if the STRF can be approximated by summation of just a few matrices otherwise it is inseparable. We measured inseparability of a STRF calculated as $Ins = 1 - \lambda_1^2/\sum_i \lambda_i^2$, where $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots$. Note that the inseparability is bounded within the interval [0,1] where the Ins is equal to 0 for separable STRFs otherwise it goes to 1.

Direction Selectivity Index (DSI): To investigate direction selectivity of STRF's, we compared total power in the 1st and the 2nd quadrant of the RTF. If a STRF favors downward-moving ripples, total power in the 1st quadrant of magnitude plot is larger than the other since the 1st quadrant is composed of responses evoked by downward-sweeping ripples. From this perspective, a DSI was defined as $DSI = (P_2 - P_1)/(P_2 + P_1)$ where P_i is a power in the i^{th} quadrant of RTF, and it is calculated by $P_i = \sum_{(\Omega,\omega)\in Q_i} |T(\Omega, \omega)|$ where Q_i means the i^{th} quadrant. Since the power on each quadrant is a non-negative value, the DSI is bounded within the interval [-1,1] where downward/upward selectivity is represented to negative/positive DSI while 0 represents no selectivity in the direction. DSI for natural vocalizations was derived using the Fourier representation described to derive FM velocity.

Best Frequency (BF): To investigate frequency selectivity of STRFs, we defined a BF as the frequency of the maximum peak of absolute STRF, $|STRF|$ over the entire time and frequency spans. Best frequency (spectral peak) of natural calls was computed by finding the peak frequency of the average spectrum.

2.4.2. A bootstrap for statistical comparison}

We performed a bootstrap analysis to evaluate similarity between distributions of characteristics (e.g. FM velocity, BV) comparing natural calls, IC neurons, and artificial neurons. The procedure selects random 30 samples from natural calls in each iteration with replacement. For IC neurons, random samples from each of the 4 bats are used in each iteration to maintain a balanced representation across bats. In case of artificial neurons, we trained 10 independent-networks (using different initialization procedures) and combined the neurons from each network into a complete set that was then sampled during the bootstrap procedure. For each comparison and each bootstrap repetition, the distance between means was noted. 1000 repetitions were used to generate a distribution of mean distances $d_{(\mu,\sigma)}$ where μ and σ are the mean and standard deviation. The *p-value* for accepting null hypothesis was calculated as $p = 1 - 2\int_0^{|\varepsilon|} d_{(0,\sigma)}(x)dx$ where $d_{(0,\sigma)}$ is a zero-mean Gaussian distribution with same variance σ, and ε was a real number satisfying $d_{(0,\sigma)}(\varepsilon) = d_{(\mu,\sigma)}(\varepsilon)$.

2.5. **Natural sound representation with artificial neurons**

2.5.1. Analysis of response selectivity in artificial neurons

We explored response selectivity to bat calls in biomimetic neurons. To replicate the study performed on IC neurons (Salles et al., 2020), FMB and Echo calls in the sound database were used to measure responses on artificial neurons. Each audio clip was fed into the network after converting to auditory spectrogram (See, Section 2.3.1), then we obtained activation probabilities for 100-nodes due to the stimulus as $\sigma(v_c)$ in Fig. 4. We averaged the activation probabilities over the same type of calls, and placed the results for 100-nodes onto a 2D scatter plot. Since the IC neurons are categorized into 3-groups such FMB selective, Echo selective, and Non-selective (Salles et al., 2020), we performed k-means clustering (k=3) on the principal axis by the principal component analysis (PCA).

2.5.2. Social call representation with artificial neurons

We explored bat's call representation with the biomimetic network. In order to perform stochastic analysis, we made 10-copies for each audio clip in the natural sound database (See. Section 2.1.3) by a data augmentation based on temporal shift so that 260 audio clips were ready for the response analysis on artificial network. After converting the audio clips for 8 types of bat call to auditory spectrogram (See, Section 2.3.1), the spectrograms were fed into the network to obtain the network's responses, the v_c in Fig. 4. Then, we estimated the Gaussian distributions for the responses to each call type, and measured a distance between two distributions by using the Jensen-Shannon divergence (JSD) as $JSD(P\|Q) = (KLD(P\|M) + KLD(Q\|M))/2$ where P and Q represent two target distributions, KLD is the Kullback-Leibler divergence (KLD), and $M = (P + Q)/2$ (Endres and Schindelin, 2003). Unlike the KLD, the JSD is bounded within the interval [0,1] where 0 means that two distributions are equal. Finally, we quantified a discriminability across the classes by averaging JSDs of all cases choosing 2 of 8. In evaluation, we calculated the averaging JSDs with 10-models for each configuration that were trained on different initial values and summarized the mean and standard deviation of the 10 results. Additionally, we explored the noise effect on the sound representation with simulated audios produced by adding Gaussian random noise to each of the 260 audio clips depending on signal to noise ratio (SNR).

To compare with neural data, we performed this analysis between FMB and Echo responses. In the same manner, we calculated JSD based on the networks. We adopted neural data used in the previous study (Salles et al., 2020). Among 575 neurons, we chose 351 neurons which were recorded with same version of stimuli, and constructed 351-dimensional vector to represent response pattern across the neurons by concatenating the number of spikes on each neuron. Once the vector is projected onto 100 dimensional space based on PCA, we estimated Gaussian distributions for FMB and Echo responses. Then, we calculated JSD between the distributions.

3. Results

3.1. Database of natural big-brown bat calls

Acoustic recordings of bat calls emitted while socially housed in the laboratory yielded a data set of natural calls containing a wide range of vocalization types. Fig. 8 shows the time-frequency representation of several types of vocalizations in the database. The bat vocalizations include isolated (non-overlapping) calls representing communication (Fig. 8A-C) or echolocating (Fig. 8D-G) sounds as well as overlapping calls from two distinct bats (Fig. 8G-H). Best Velocity (BV) values reflect the broad range of FM energies in these social communication calls (BV = 18 oct/s, 43 oct/s and 140 oct/s for Fig. 8A, B and C respectively). Echolocation calls show even higher FM energies with shorter signals (BV = 274 oct/s and 333 oct/s for Fig. 8D and E respectively). In Fig. 8G and H, we note presence of multiple calls though the statistics derived from that segment are largely influenced by the dominant call (BV = 381 oct/s and 410 oct/s for Fig. 8G and H respectively). The natural complexity in the animal’s auditory environment was maintained in this study and no supervised curation of this data set was performed beyond removal of silence segments (see *Methods*). We also note presence of ambient background in all recordings as a result of the cage environment and recording setup used to collect this data.

3.2. Auditory characteristics of biological STRF's

To explore auditory characteristics of big brown bat midbrain, we calculated STRF’s from neural recordings of IC neurons. Fig. 9A highlights examples from 6 neurons, revealing a downward sweep selectivity, with excitation and inhibition represented as red and blue areas, respectively. The best frequency (BF) is also shown as red dashed line indicating the maximum peak, positive or negative, of the STRF. We evaluated auditory characteristics across all neural recordings with respect to BV, DSI (direction selectivity), orientation, and inseparability (Fig. 10, yellow histograms). Using a bootstrap procedure, we compared the auditory characteristics of IC neurons to properties of natural calls (Fig. 10, gray background regions for standard deviation). The analysis revealed that the distribution of BVs in IC neurons is statistically equivalent to that of natural calls (μ=-1.63, σ=17.13, *p-value*=0.9622, Fig. 10A). A match was also observed for direction selectivity μ =-0.01, σ=0.03, *p-value*=0.8789, Fig. 10B). This result is consistent with the hypothesis that IC neurons have consistent tuning to the statistics of conspecific vocalizations (Andoni et al., 2007). We noted that the majority of IC neurons (93.6%) favored downward sweeps (Fig. 10B) (Gittelman et al., 2009) while their orientation is centered around 0 deg. Most IC neurons yield higher than rank-one STRF’s (average inseparability index 0.49 ± 0.09).

The distribution of frequency tuning (BF) of IC neurons tended to fall between 10 and 30 kHz. Particularly, BF's of 87% of neurons are below 30 kHz (Fig. 11B). In contrast, spectral peaks observed in the vocalization database revealed a higher spectral peak (37.17 ± 5.62) as shown in Fig. 11A. This profile is likely driven by the strength of the first harmonic component in vocalization which tends to be stronger than other components. As seen from the examples in Fig. 8, most vocalizations contain multiple harmonic peaks with higher energy in the first component resulting in a difference between the BF of IC neurons and spectral peaks of the calls database (μ=-12.58, σ=1.82, *p-value*=0).

3.3. Auditory characteristics of artificial STRF's

Using natural calls, an artificial network was trained to best represent the statistics of the vocalization. Characteristics of model neurons were analyzed in the same way as biological neurons using spectro-temporal receptive fields. The distribution of model characteristics is shown in Fig. 10, overlaid in blue. Compared to natural calls, model neurons reveal a statistically matching distribution with respect to BV (bootstrap μ=-2.62, σ=19.85, *p-value*=0.9473) and DSI (bootstrap μ=-0.005, σ=0.01, *p-value*=0.9382). Model neurons also match the spectral peak of natural calls (bootstrap μ=-0.49, σ=4.75, *p-value*=0.9592, Fig. 11C). These results are not surprising given that the model was trained to mimic the statistics of these calls. Still, the model was not specifically configured to match specific directionality or velocity patterns but rather represent the time-frequency profile of the calls as a whole.

In parallel, the comparison between model and biological neurons reveal remarkable agreement. A bootstrap procedure was performed to compare all auditory characteristics of these STRF's, and results are shown in inset panels in Fig. 10. We note that characteristics of biomimetic neurons match the properties of IC neurons including BVs (μ=2.92, σ=16.61, *p-value*=0.9300), DSI (μ=0.01, σ=0.07, *p-value*=0.9358), orientation (μ=1.34, σ=3.26, *p-value*=0.8370), and inseparability (μ=-0.02, σ=0.04, *p-value*=0.8079). The BFs of artificial neurons are statistically different from IC neurons (bootstrap μ=12.10, σ=4.44, *p-value*=0.1731), even though there is substantial overlap at the range of 0-40 kHz. The BFs of artificial neurons are more broadly distributed over the entire frequency range with about 14% of artificial neurons having high BF (above 60 kHz) (Ferragamo et al., 1997).

3.4. Architecture of the biomimetic network

While results reported so far focus on the 'best' biomimetic network, we also investigated how changing the architecture of the model affects the tuning parameters of artificial neurons. We systematically varied the model in terms of structural complexity (the number of stacking blocks), sparsity of the latent space and non-linearity of the activation function. Fig. 12A shows the mean and standard deviation of characteristics of model neurons across 10-network validations for each pair of complexity and sparsity ($\alpha = 0.2$). The mean FM velocities and orientation in the natural calls database are represented by a black line on each panel; while the grey regions represent 95% confidence intervals for each mean. The results show that a very shallow model (mono-stacking) results in a greatly biased negative orientation, as well slower BV estimates. By increasing the model depth, there is an increased match between the model's spectro-temporal configuration (represented by BV and orientation) and that of natural statistics. Furthermore, extremely low or high sparsity values also result in over or under-estimating statistics of natural calls; with 10% sparsity results in a great match with average statistics of the natural calls.

Using the triple stacking network with 10% sparsity, we investigated the effect of the model non-linearity on the same auditory characteristics of model neurons (Fig. 12B). Setting the non-linearity parameter to 1.0 results in a fully linear processing which clearly produces in a mismatch between the model and call characteristics. By increasing the degree of non-linearity (decreasing alpha), we note a closer match between the two.

It should be noted that across all the different configurations of the model, all architectures were able to converge (i.e. minimize the reconstruction error between the spectrogram of a given call sound and its reconstruction using the model's latent space). Fig. 12C shows the average reconstruction error over the 10 models for each parameter set. While all models successfully converge to reconstruct natural calls and encode statistics of in the database, only a few configurations result in a reasonable match to the spectro-temporal characteristics of model neurons. As a matter of fact, the model was not constrained to match these properties in its latent space; it is merely trained to represent the call spectrograms as faithfully as possible. This requirement has multiple plausible solutions, and only certain configurations result in a close match with velocity and orientation characteristics of natural calls.

3.5. Natural call representation with the biomimetic network

So far, the results suggest that a deep nonlinear architecture with high sparsity to achieve an optimal representation of the statistics of natural bat vocalizations is capable to replicate auditory characteristics of the bat's midbrain. We next examined the implications of this mapping to facilitate discrimination of the large variety in the call repertoire. A study revealed that tuning characteristics of bat IC neurons differentially encode different sound categories in the bat vocalizations, specifically echolocation calls and food-claiming FMB (frequency-modulated bout) social calls (Salles et al., 2020). We examined whether the artificial network, trained simply to emulate natural statistics in the bat repertoire (without knowledge of different sound classes) also yields distinct activations of these different groups. Fig. 13A top replicates the response selectivity of biological IC neurons, showing a scatter plot of average activation probabilities for each neuron in response to FMB calls (x-axis) vs. Echolocation (echo) calls (y-axis), projected on the principal axis by PCA. The figure inset shows the original neural responses before data projection. Fig. 13B-D depict a similar analysis of call selectivity for the mono, double and triple artificial network, respectively. Note that each panel from B to D was produced by one network of 10-models for example. The top panels show a scatter. Across the 3 network configurations, we note that the mono stacking model induces mostly non-selective activation across its neural population (Fig. 13B) while the double stacking model yields biased responses in favor of Echo calls (Fig. 13C). The triple stacking model reveals a more balanced activation from Echolocation and FMB social call types (Fig. 13D) that closely matches biological selectivity.

We extend the analysis of call selectivity in the artificial network to other classes of calls in the bat repertoire. We evaluated discriminability across 8 types of calls using the Jensen-Shannon divergences (JSDs) (Endres and Schindelin, 2003). Fig. 14 shows the results for various network depths, linearity and sparsity for the calls in the database (clean) as well as with additional simulated additive noise with decreasing signal-to-noise (SNR) ratios. The triple stacking model (with high sparsity and nonlinear activation) produces the most discriminable responses, as well as more robust discrimination even in presence of noise. Shallower architectures are clearly affected by presence of resulting in reduced discriminability. Linear activations and low sparsity appear to also affect discriminability and robustness to noise albeit not at the same rate. These results suggest that the optimal representation of call statistics likely plays a role in facilitating the identification of different sound classes even in presence of noise. Similarly, a study with guinea pigs has shown the robust discrimination in the responses to communication sounds (Souffi et al., 2020). Such hypothesis aligns with earlier reports (Chechik et al., 2006) but remains to be validated in the IC of the big brown bat. As reference, we computed JSD for the two echo and FMB call classes (Fig. 14) for both artificial and biological neurons. Both measures reveal a close agreement and high discriminability that far surpasses selectivity from shallower architectures.

4. Discussion

The biomimetic artificial network provides a nonlinear response model of neural selectivity

To examine the tuning of auditory neurons, each cell can be considered as a system with a mapping function F that represents a relation between stimulus s and neural response r i.e. $r = F(s)$. While characterizing the full system function may be theoretically or experimentally nearly impossible, linearized models using STRF are often used to build a computational response model as $r(t) = \int s(t,f) * h(t,f) df$ where t and f is respectively time and frequency index, $s(t,f)$ is spectro-temporal representation for a stimulus, * is a convolution operator, and $h(t,f)$ represents a spectro-temporal receptive fields (STRF's) (Depireux et al., 2001; Elhilali et al., 2013; Machens et al., 2004; Fritz et al., 2003). This model is often applied with reasonable success to predict neural responses to other sound classes including conspecific vocalizations or other natural sounds. Although the linear model is a reasonable approximation for mimicking neural responses in the brain, it is limited in its ability to inform nonlinear transformations that are usually observed in between stimulus and response (Escabi and Schreiner, 2002; Theunissen et al., 2000). One of the main advantages of including nonlinear activations in a feed-forward propagation in the proposed neural network is that it implicitly incorporates the effects of these nonlinear mappings in the propagation of activity throughout the network. Still, we are able to evaluate the linearized portion of the response (via STRF's of artificial neurons) without explicitly incorporating the nonlinear terms in the STRF model itself. This black-box approach to incorporate complexities of neural mapping via deep neural networks opens the possibility to more intricate readouts of the representation of artificial networks. We anticipate that such biomimetic artificial network can be used to build a system mimicking the bat’s ability for object shape recognition using its bio-sonar.

Midbrain responses are optimized to represent the statistics of natural calls in a bat's soundscape

In this study, we explored the hypothesis that the bat’s IC neurons are tuned to represent the FM velocity and spectro-temporal structure of conspecific vocalizations. Evidence in support of this Sender-Receiver matching has been previously reported in the pallid bat (Fuzessery, et al., 2006) and Mexican free-tailed bat (Andoni et al., 2007), as well as other species such as zebra finches (Woolley et al., 2005) (also see (Woolley and Portfors, 2013). Here, we report similar findings in the big brown bat, and establish a close correspondence between acoustic characteristics of natural calls and tuning of spectro-temporal receptive fields of IC neurons of the big brown bat. Going beyond this relationship, an artificial network trained independently on these natural calls reveals tuning properties that not only conform with spectro-temporal features of the calls (which they were trained on), but also unveils IC-like tuning structure and complexity (e.g. separability) that the model was not specifically trained on (Fig. 10). This result hints that the midbrain architecture gives rise to tuning configurations that leverage the spectro-temporal richness of the bat's repertoire to not only represent these features with high fidelity but also enable selective responses to discriminate between classes of natural calls.

The artificial network used in the current study shows that the neural encoding of an incoming stimulus gives rise to a response across neural populations that enables it to faithfully reconstruct this stimulus, revealing a high fidelity mapping without loss of information. While not explicitly happening in the brain, this stimulus reconstruction from the internal latent space is the basis for training the artificial network which yields emergent tuning that matches the biology. It is important to note that tuning properties of artificial neurons were derived using moving ripples which invoke the principle of signal decomposition by separating each conspecific call into a sum of ripples with different orientations, rates and phases, in line with the Fourier theory of signal representation. While the network was never trained on these ripples, its response to each ripple spectral motion pattern both in terms of magnitude and phase (both needed for STRF reconstruction) suggest a quantitative correspondence with the downward-sweeping signals that are prominent in the bat repertoire. It is also important to note that not all known coding properties of the bat midbrain are represented in STRFs (Brimijoin and O’Neill, 2010) and that future steps to test time varying response properties (such as an adaptation) would further validate the ability of this network to replicate the biology of the bat IC (Lesica and Grothe, 2008; Rabinowitz et al., 2013; Lohse et al., 2020).

A deep architecture with sparsity is best suited to model the statistics of natural calls

Varying the architecture of the network led to different latent spaces to represent the characteristics of the database of natural calls. Specifically, changing the complexity of the network (in terms of depth), sparsity and nonlinearity converged on different solutions for representing conspecific sounds. Under all configurations, the networks were able to reconstruct the input spectrogram with minimal error indicating that its latent space is sufficiently informative about the statistics in the training database (Fig. 12C). Nonetheless, only a specific configuration with high sparsity, nonlinearity and sufficient depth is able to replicate biological tuning properties, giving insights into coding principles underlying configuration of IC networks probed in this study. Naturally, while this investigation cannot rule out other configurations that would also reveal a strong match to biology, it can eliminate parameters that converge on solutions that are far from the biology (e.g. shallow networks, linear models). It is worth noting that we were unable to train a quadruple stacking network to represent statistics in the database so we are unable to comment on the extend to which an even deeper network may correlate with biological tuning. The output of a 4th block could be missing spectro-temporal features due to over-compression. This is an issue that could explored using large input patches or modifying the pooling step.

Tuning to conspecific natural sounds may underlie selective and robust encoding of auditory objects

We note that directional selectivity to FM sweeps in biological and artificial neurons, results in high discriminability between different classes of calls. Specifically, these results support the notion that by having neural sub-population tuned to different subsets of spectro-temporal statistics, the network is able to encode and differentially respond to different vocalizations and social or echolocation calls. This discriminability is enhanced in the triple sparse and nonlinear network that best matches biological tuning and much reduced in other network configurations despite the fact that these other models were also successfully trained to represent the same natural statistics in the bat call repertoire. This variability may stem from correlated behavior across the neural population which was previously shown to play an important role in enhanced discriminability of vocalizations in the auditory midbrain (Schneider and Woolley, 2010). This encoding selectivity remains fairly stable in presence of stationary ambient noise suggesting that the high dimensional mapping encoding incoming natural calls results in a noise invariant representation that is believed to start emerging at the level of the inferior colliculus and further strengthen in auditory cortex (see, Willmore et al., 2014).

References

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al.. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems; 2016. Available from: http://arxiv.org/abs/1603.04467.

Aitkin L, Tran L, Syka J. The responses of neurons in subdivisions of the inferior colliculus of cats to tonal, noise and vocal stimuli. Experimental Brain Research. 1994;98(1):53–64.

Andoni S, Li N, Pollak GD. Spectrotemporal Receptive Fields in the Inferior Colliculus Revealing Selectivity for Spectral Motion in Conspecific Vocalizations. Journal of Neuroscience. 2007;27(18):4882–4893.

Baldi P. Autoencoders, Unsupervised Learning, and Deep Architectures. In: ICML workshop on unsupervised and transfer learning; 2012. p. 37–49.

Brimijoin WO, O'Neill WE. On the prediction of sweep rate and directional selectivity for FM sounds from two-tone interactions in the inferior colliculus. Hearing Research. 2005;210(1-2):63–79.

Brimijoin WO, O'Neill WE. Patterned tone sequences reveal non-linear interactions in auditory spectrotemporal receptive fields in the inferior colliculus. Hearing Research. 2010;267(1-2):96–110.

Casseday JH, Fremouw T, Covey E. The Inferior Colliculus: A Hub for the Central Auditory System. In: Integrative Functions in the Mammalian Auditory Pathway. New York: Springer; 2002. p. 238–318.

Chechik G, Anderson MJ, Bar-Yosef O, Young ED, Tishby N, Nelken I. Reduction of Information Redundancy in the Ascending Auditory Pathway. Neuron. 2006;51(3):359–368.

Chi T, Ru P, Shamma SA. Multiresolution spectrotemporal analysis of complex sounds. Journal of the Acoustical Society of America. 2005;118(2):887–906.

Chi T, Shamma S. NSL Matlab Toolbox; 2005. Available from: http://www.isr.umd.edu/~speech/nsltools.tar.gz.

Choi KP, Xia A. Approximating the number of successes in independent trials : Binomial versus Poisson. The annals of applied probability. 2002;12(4):1139–1148.

Corrado GS, Sugrue LP, Seung HS, Newsome WT. Linear-Nonlinear-Poisson Models of Primate Choice Dynamics. Journal of the Experimental Analysis of Behavior. 2005;84(3):581–617.

Depireux DA, Simon JZ, Klein DJ, Shamma SA. Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. Journal of neurophysiology. 2001;85(3):1220–1234.

Dietterich T. Overfitting and undercomputing in machine learning. ACM Computing Surveys. 1995;27(3):326–327.

Doersch C. Tutorial on Variational Autoencoders; 2016. Available from: http://arxiv.org/abs/1606.05908.

Elhilali M, Shamma SA, Simon JZ, Fritz JB. A Linear Systems View to the Concept of STRF. Handbook of Modern Techniques in Auditory Cortex. Nova Science Pub Inc; 2013. p. 33–60.

Endres DM, Schindelin JE. A new metric for probability distributions. IEEE Transactions on Information Theory. 2003;49(7):1858–1860.

Escabi MA, Schreiner CE. Nonlinear Spectrotemporal Sound Analysis by Neurons in the Auditory Midbrain. Journal of Neuroscience. 2002;22(10):4114–4131.

Ferragamo MJ, Haresign T, Simmons JA. Frequency tuning, latencies, and responses to frequency-modulated sweeps in the inferior colliculus of the echolocating bat, Eptesicus fuscus. Journal of Comparative Physiology - A Sensory, Neural, and Behavioral Physiology. 1997;182(1):65–79.

Fritz J, Shamma S, Elhilali M, Klein D. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. Nature Neuroscience. 2003;6(11):1216–1223.

Fuzessery ZM, Richardson MD, Coburn MS. Neural Mechanisms Underlying Selectivity for the Rate and Direction of Frequency-Modulated Sweeps in the Inferior Colliculus of the Pallid Bat. Journal of Neurophysiology. 2006;96(3):1320–1336.

Gittelman JX, Li N, Pollak GD. Mechanisms underlying directional selectivity for frequency-modulated sweeps in the inferior colliculus revealed by in vivo whole-cell recordings. Journal of Neuroscience. 2009;29(41):13030–13041.

Gordon M, O'Neill WE. Temporal processing across frequency channels by FM selective auditory neurons can account for FM rate selectivity. Hearing Research. 1998;122(1-2):97–108.

Kay SM. Fundamentals of statistical signal processing. Englewood Cliffs, N.J.: Prentice-Hall PTR; 1993.

Kowalski N, Depireux DA, Shamma SA. Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra. Journal of Neurophysiology. 1996;76(5):3503–3523.

Lesica NA, Grothe B. Efficient temporal processing of naturalistic sounds. PLoS ONE. 2008;3(2).

Lohse M, Bajo VM, King AJ, Willmore BDB. Neural circuits underlying auditory contrast gain control and their perceptual implications. Nature Communications. 2020;11(1):1–13.

Luo J, Moss CF. Echolocating bats rely on audiovocal feedback adapt sonar signal design. Proceedings of the National Academy of Sciences of the United States of America. 2017;114(41):10978–10983.

Park J, Kim W, Han DK, Ko H. Voice activity detection in noisy environments based on double-combined Fourier transform and line fitting. The Scientific World Journal. 2014;2014.

Poon PWF, Yu PP. Spectro-temporal receptive fields of midbrain auditory neurons in the rat obtained with frequency modulated stimulation. Neuroscience Letters. 2000;289(1):9–12.

Quiroga RQ, Nadasdy Z, Ben-Shaul Y. Unsupervised Spike Detection and Sorting with Wavelets and Superparamagnetic Clustering. Neural Computation. 2004;16(8):1661–1687.

Rabinowitz NC, Willmore BDB, King AJ, Schnupp JWH. Constructing Noise-Invariant Representations of Sound in the Auditory Pathway. PLoS Biology. 2013;11(11).

Radford A, Metz L, Chintala S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks; 2015. Available from: http://arxiv.org/abs/1511.06434.

Maas AL, Hannun AY, Ng AY. Rectifier nonlinearities improve neural network acoustic models. In: in ICML Workshop on Deep Learning for Audio, Speech and Language Processing. vol. 28; 2013. p. 1–6.

Machens CK, Wehr MS, Zador AM. Linearity of cortical receptive fields measured with natural sounds. The Journal of neuroscience. 2004;24(5):1089–1100.

Macıas S, Luo J, Moss CF. Natural echolocation sequences evoke echo-delay selectivity in the auditory midbrain of the FM bat, eptesicus fuscus. Journal of Neurophysiology. 2018;120(3):1323–1339.

Salles A, Park S, Sundar H, Macias S, Elhilali M, Moss CF. Neural Response Selectivity to Natural Sounds in the Bat Midbrain. Neuroscience. 2020;434:200–211.

Scherer D, Muller A, Behnke S. Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition. In: International Conference on Artificial Neural Networks (ICANN); 2010. p. 92–101.

Schneider DM, Woolley SMN. Discrimination of Communication Vocalizations by Single Neurons and Groups of Neurons in the Auditory Midbrain. Journal of Neurophysiology. 2010;103(6):3248–3265.

Schwartz O, Pillow JW, Rust NC, Simoncelli EP. Spike-triggered neural characterization. Journal of Vision. 2006;6(4):484–507.

Shamma SA. Speech processing in the auditory system I: The representation of speech sounds in the responses of the auditory nerve. The Journal of the Acoustical Society of America. 1985a;78(5):1612–1621.

Shamma SA. Speech processing in the auditory system II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve. The Journal of the Acoustical Society of America. 1985b;78(5):1622–1632.

Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017;39(4):640–651.

Singh N, Theunissen F. Modulation spectra of natural sounds and ethological theories of auditory processing. Journal of the Acoustical Society of America. 2003;106:3394–3411.

Smith EC, Lewicki MS. Efficient auditory coding. Nature. 2006;439:978–982.

Souffi S, Lorenzi C, Varnet L, Huetz C, Edeline JM. Noise-Sensitive but More Precise Subcortical Representations Coexist with Robust Cortical Encoding of Natural Vocalizations. Journal of Neuroscience. 2020;40(27):5228–5246.

Strang G. Introduction to linear algebra. 4th ed. Wellesley-Cambridge Press; 2009.

Suta D, Kvasnak E, Popelar J, Syka J. Representation of Species-Specific Vocalizations in the Inferior Colliculus of the Guinea Pig. Journal of Neurophysiology. 2003;90(6):3794–3808.

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 07-12-June; 2015. p. 1–9.

Theunissen FE, Sen K, Doupe AJ. Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. Journal of Neuroscience. 2000;20(6):2315–2331.

Wang K, Shamma SA. Self-normalization and noise-robustness in early auditory representations. IEEE Transactions on Speech and Audio Process. 1994;2:421–435.

Williams AJ, Fuzessery ZM. Facilitatory mechanisms shape selectivity for the rate and direction of FM sweeps in the inferior colliculus of the pallid bat. Journal of Neurophysiology. 2010;104(3):1456–1471.

Willmore BDB, Cooke JE, King AJ. Hearing in noisy environments: noise invariance and contrast gain control. The Journal of Physiology. 2014;592(16):3371–3381.

Woolley SMN, Fremouw TE, Hsu A, Theunissen FE. Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. Nature Neurosci. 2005;8(10):1371–1379.

Woolley SMN, Portfors CV. Conserved mechanisms of vocalization coding in mammalian and songbird auditory midbrain. Hearing Research. 2013;305(1):45–56.

Wright GS, Chiu C, Xian W, Wilkinson GS, Moss CF. Social calls of flying big brown bats (Eptesicus fuscus). Frontiers in Physiology. 2013;4 AUG(August):1–9.

Yang X, Wang K, Shamma SA. Auditory representations of acoustic signals. IEEE Transactions on Information Theory. 1992;38(2):824–839.

Figures and Tables

Fig. 1 Overview of study foci. A database of natural calls from a colony of big brown bats is collected and analyzed for its auditory characteristics. Shown in the figure is a distribution of FM velocities. Right: Tuning characteristics of biological neurons from the big brown bat inferior colliculus are derived using Spectro-Temporal Receptive Field (STRF) method, and properties of biological neurons are derived (e.g. best velocity, BV). Shown in the figure is a brain slice identifying the location of the IC in the big brown bat (from Salles et al., 2020). Left: Computational models with various configurations are examined and emergent tuning properties of artificial networks are derived to compare against statistics of natural calls as well as biological neurons.

Fig. 2 Example spectrograms of 8 bat calls in social call database. **A**, Echolocation (Echo). **B**, Frequency modulated bout (FMB). **C**, Upward frequency modulated (UFM). **D**, Long frequency modulated (LFM). **E**, Short frequency modulated (SFM). **F**, Chevron shaped (CS). **G**, Hook. **H**, Long-wave.}

Fig. 3 Ripple transfer function extraction. **A**, A subset of ripple stimuli. **B**, Input patch sequence configuration from a ripple stimulus to a code vector for characterizing the network. **C**, Examples of responses on each node (each element of the code vector), and definition of magnitude m and phase ϕ in a ripple response. **D**, Magnitude and phase plots on one of nodes.}

Fig. 4 Convolutional layered autoencoder structure for biomimetic network. The flow denoted in black shows a double stacking structure as a standard example. Based on this structure, a deeper structure can be constructed by stacking more modules, on the other hand, a shallow structure is created by removing a module on the top of the standard example.

Fig. 5 Operations using multi-scale filters. **A**, convolution using multi-scale filters. **B**, transposed convolution using multi-scale filters.

Fig. 6 STRF calculation. **A**, expanded magnitude and phase matrices which are matching to Fig. 2D. **B**, 2-Dimensional (above) and 3-Dimensional (bottom) representation of the STRF that is obtained by performing 64 by 64 interpolation and Gaussian smoothing sequentially. In 2D representation, red area represents excitation regions while blue represents inhibition regions.

Fig. 7 Descriptions for best velocity and orientation. **A**, magnitude plot. **B**, the result of Gaussian surface fitting. The red ellipse represents Gaussian mean vector μ (center) and covariance matrix Σ (rotation), the best velocity is defined by a slope of Gaussian mean vector and the orientation error is defined by an angle difference between mean vector and covariance rotation.

Fig. 8 Example spectrograms of several types of calls monophonic cases for social communication (A-C), echolocation (D-F), and polyphonic cases (G-H). Note the differences in frequency content, duration, and sweep velocity. Note that peak frequency is represented onto each panel as the black dashed line.

Fig. 9 Examples for biological STRF and biomimetic STRF. **A**, biological STRFs obtained from bat's IC neuron. **B**, biomimetic STRFs obtained from a triple stacking network with 10% sparsity. Note that red and blue area show excitation and inhibition regions, respectively.

Fig. 10 Histogram of biological and biomimetic STRFs according to auditory characteristics. **A**, Best velocity. **B**, Direction selectivity index. **C**, Orientation (The zero-mean is marked as the red line). **D**, Inseparability.

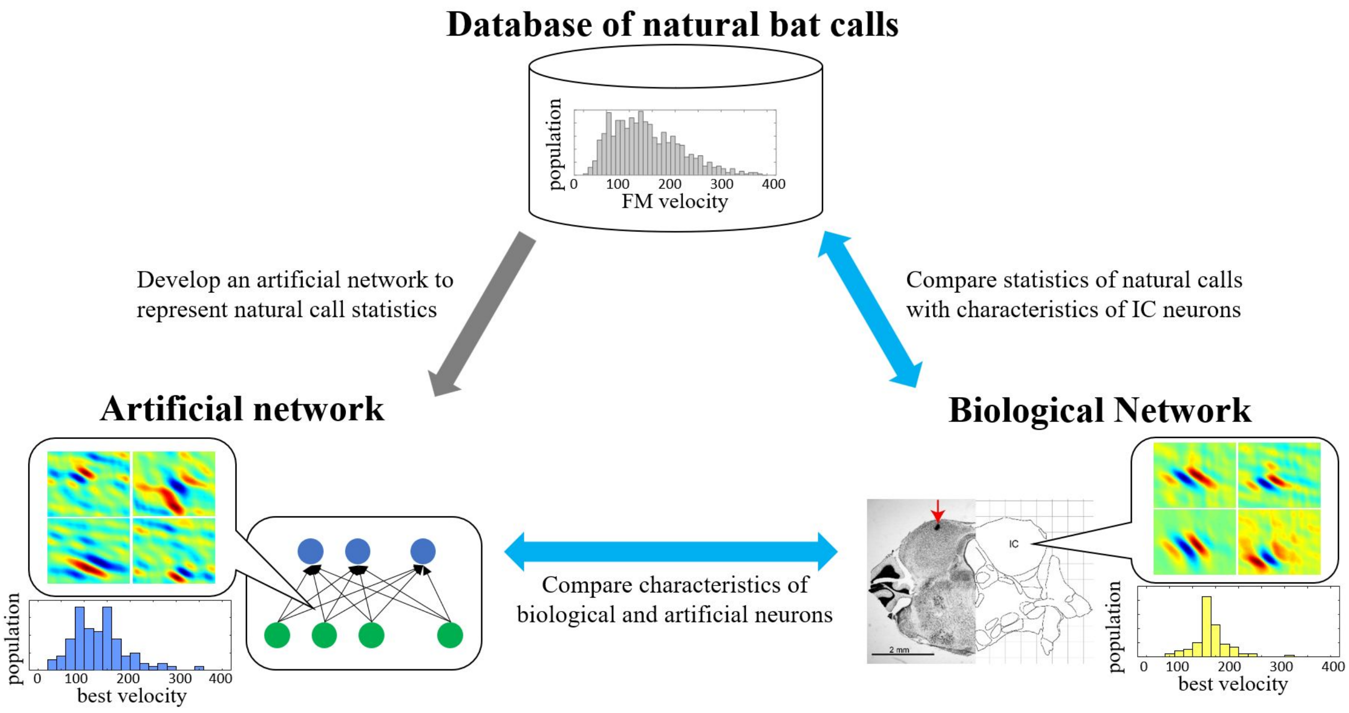
Fig. 11 Analysis of best frequency in dataset, IC neurons, and artificial neurons. **A**, histogram of peak frequencies in natural calls, the background grey line represents averaging spectrum envelop of natural calls. **B**, BFs on IC neurons. **C**, BFs on artificial neurons.

Fig. 12 Best velocity and Orientation of biomimetic STRF's depending on network's configuration. **A**, for the number of stacking modules and sparsity (0.2 lReLu). **B**, for nonlinearity (Triple stacking model with 10% sparsity). **C**, average reconstruction error over the 10 networks for each parameter set, blue and red dashed lines are mean of the errors for all configurations at the beginning of training and the end of the training, respectively.

Fig. 13 Selectivity to FMB vs. Echolocation call for; **A**, IC neurons (Salles et al., 2020), spike frequency was calculated by dividing the number of spikes by total number of spikes starting 5 ms after stimulus onset. The horizontal axis on the bottom was circularly shifted with zero-centered non-selective neurons. **B**, mono-stacking model. **C**, double-stacking model. **D**, triple-stacking model.}

Fig. 14 Natural sound representation by biomimetic network in different SNR conditions

Table 1. Description of network parameters, midlevel feature maps, and input.



Database of natural bat calls

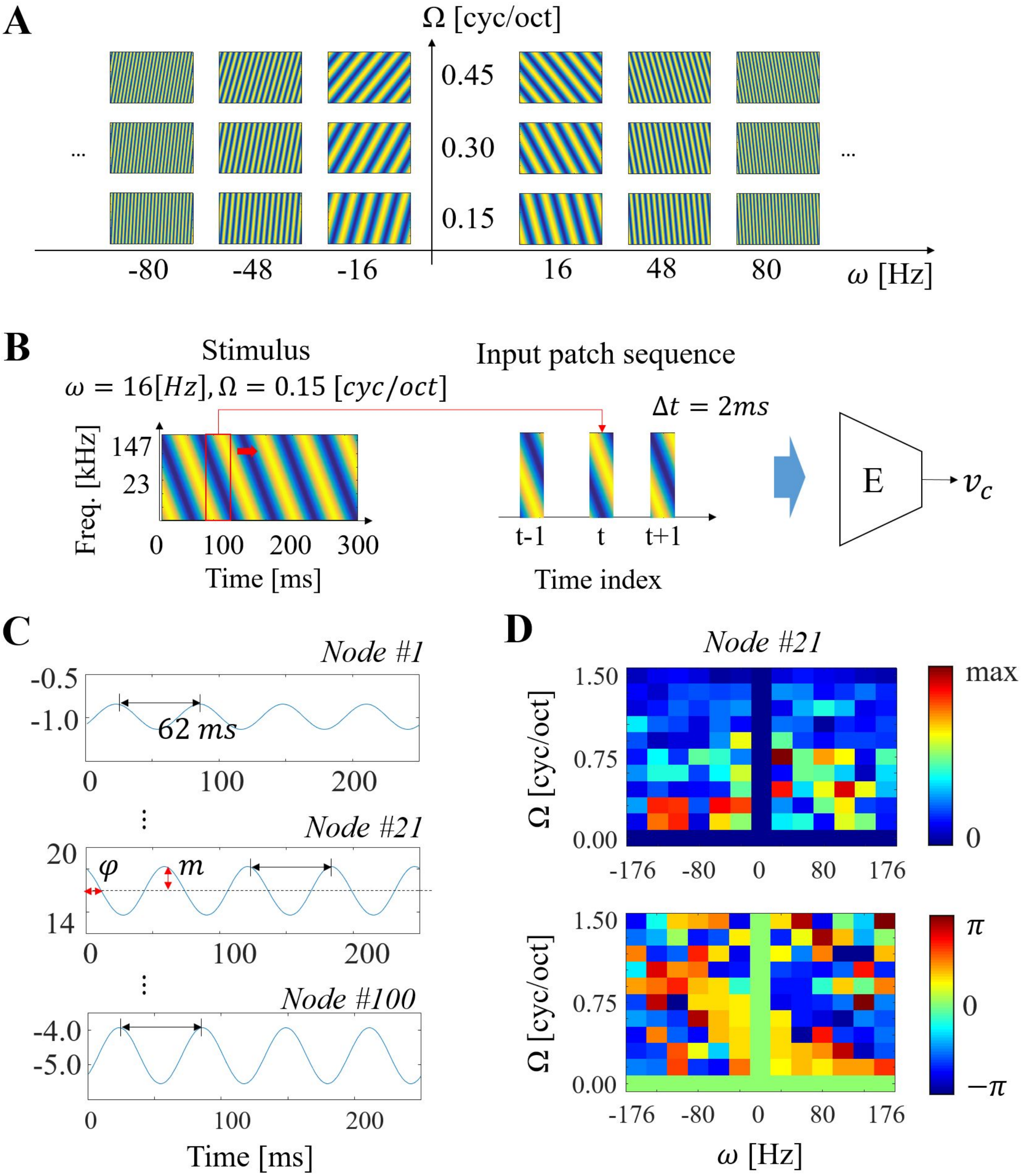
Develop an artificial network to represent natural call statistics

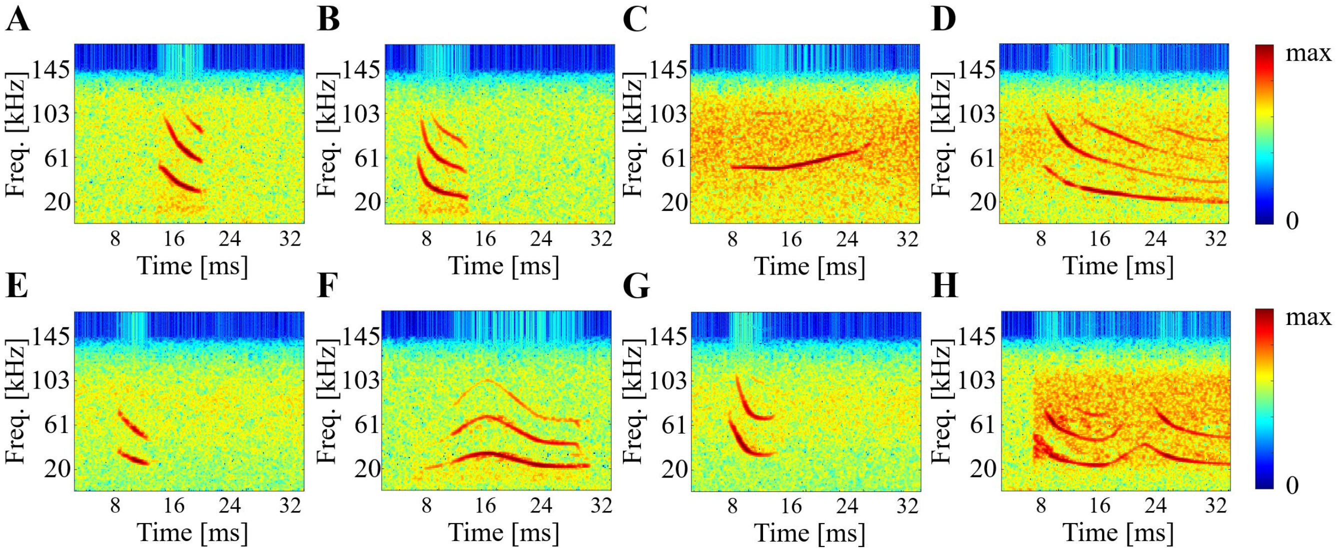
Compare statistics of natural calls with characteristics of IC neurons

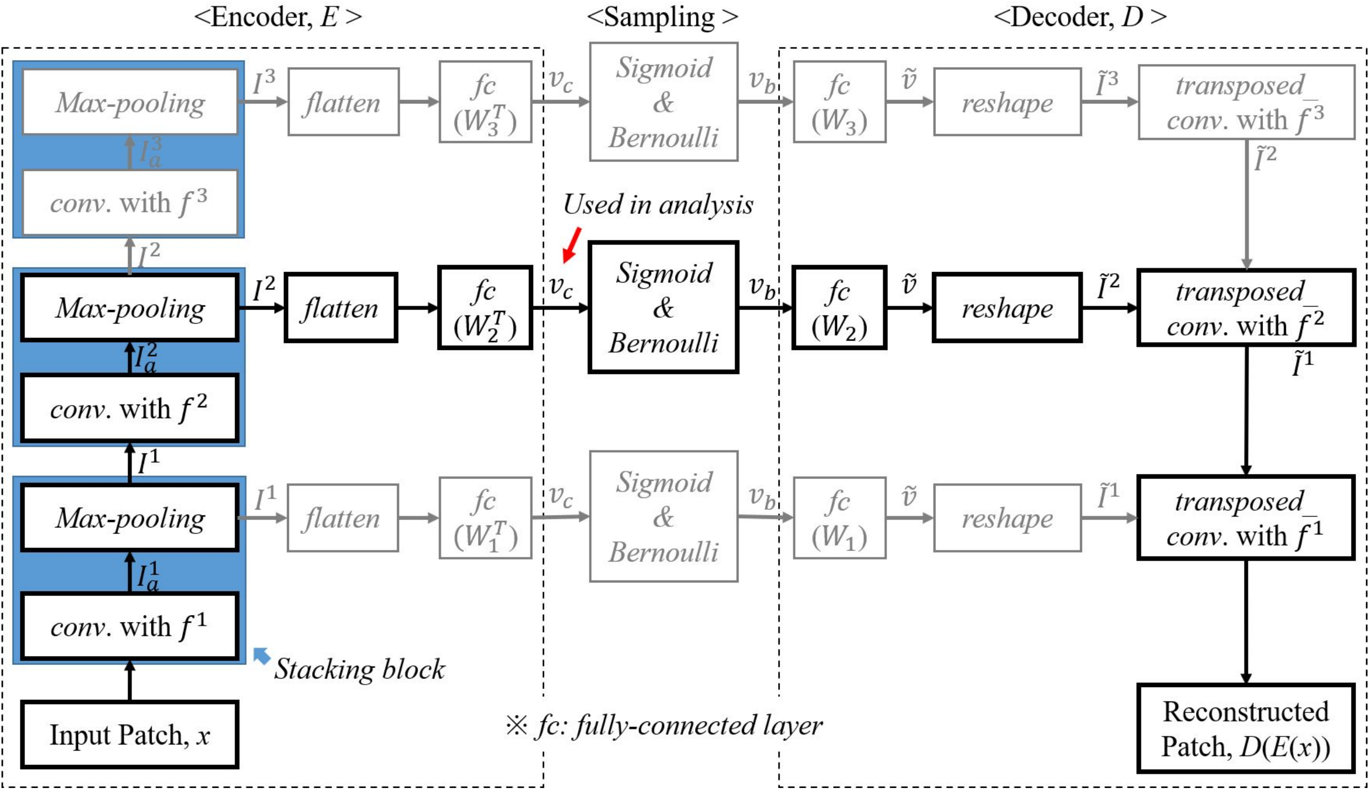
Artificial network

Biological Network

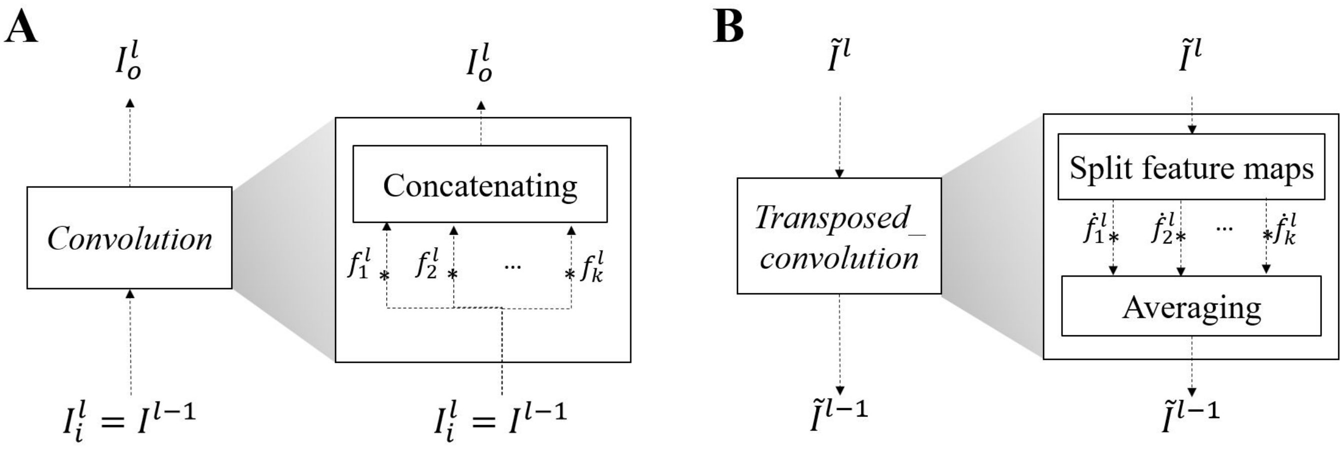
Compare characteristics of biological and artificial neurons

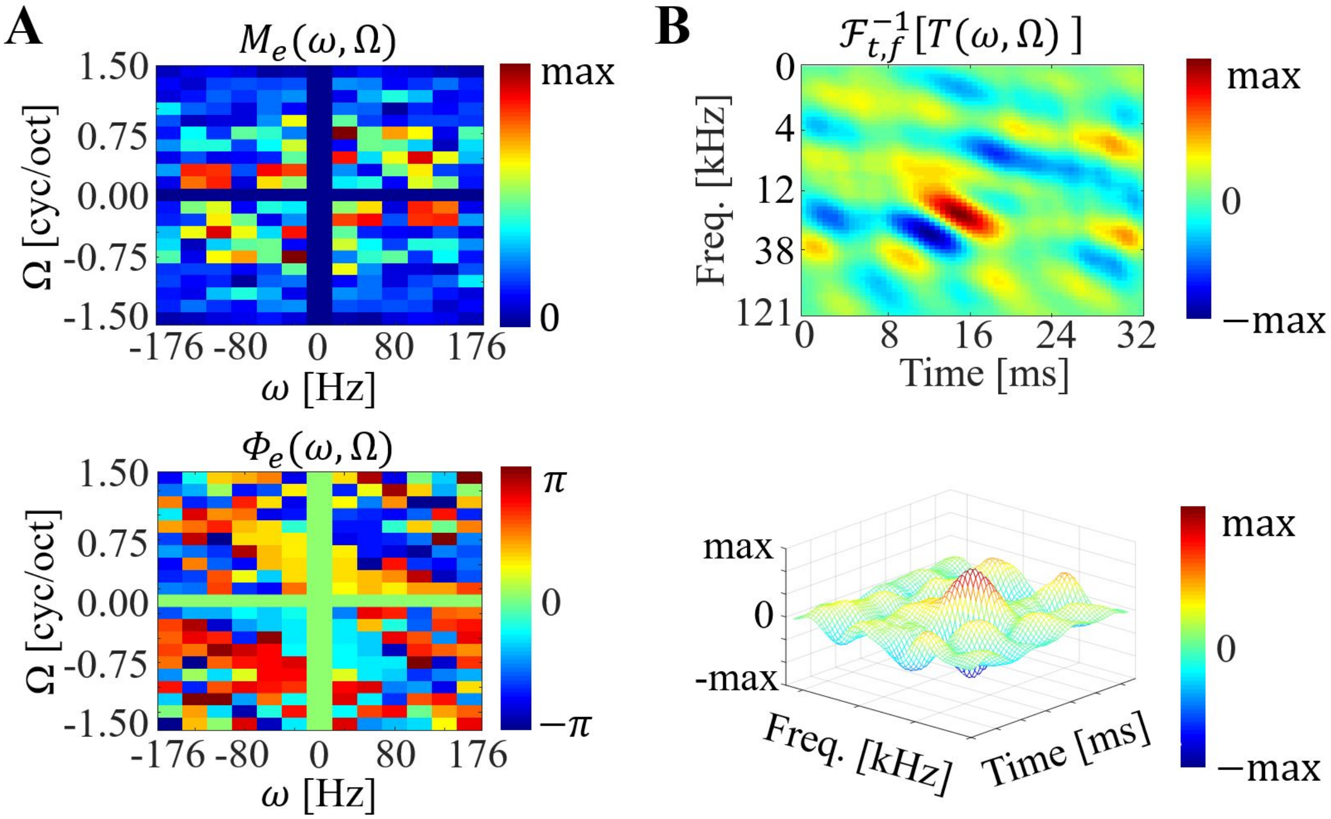

A
Ω [cyc/oct]
0.45
0.30
0.15
...
...
-80
-48
-16
16
48
80
ω [Hz]
B
Stimulus
$\omega = 16[Hz], \Omega = 0.15\ [cyc/oct]$
Input patch sequence
$\Delta t = 2ms$
Freq. [kHz]
147
23
0
100
200
300
Time [ms]
t-1
t
t+1
Time index
E
v_c
C
Node #1
-0.5
-1.0
62 ms
0
100
200
Node #21
20
14
φ
m
0
100
200
Node #100
-4.0
-5.0
0
100
200
Time [ms]
D
Node #21
Ω [cyc/oct]
1.50
0.75
0.00
-176
-80
0
80
176
max
0
Ω [cyc/oct]
1.50
0.75
0.00
-176
-80
0
80
176
π
0
−π
ω [Hz]

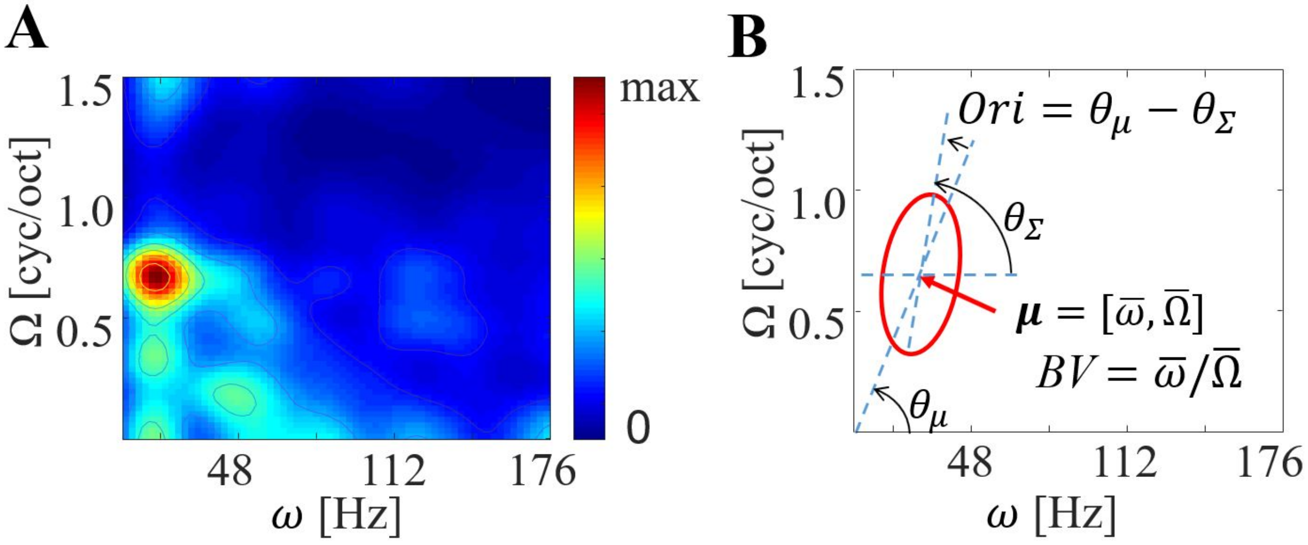

A
B
C
D
E
F
G
H
Freq. [kHz]
145
103
61
20
Time [ms]
8
16
24
32
max
0

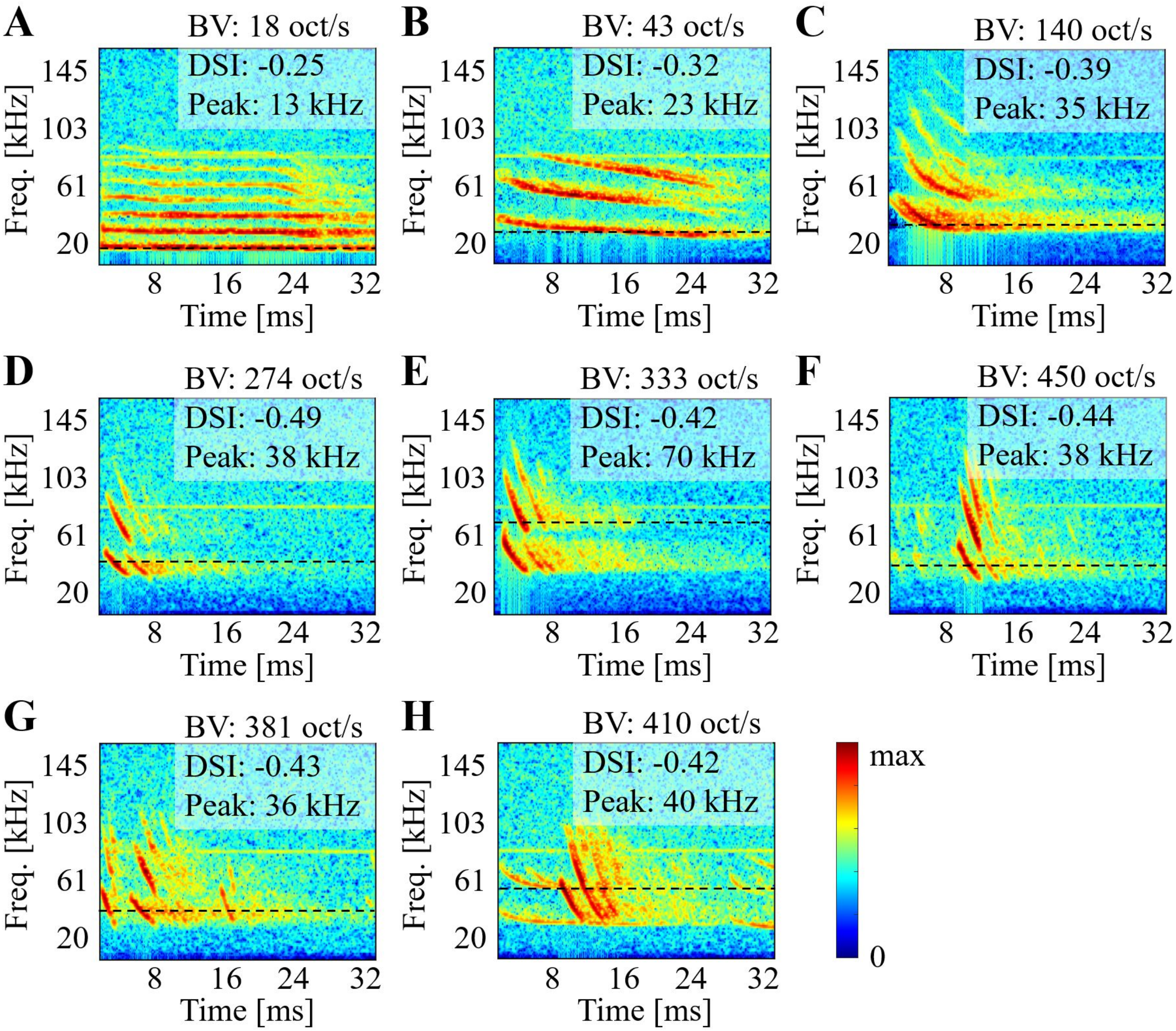


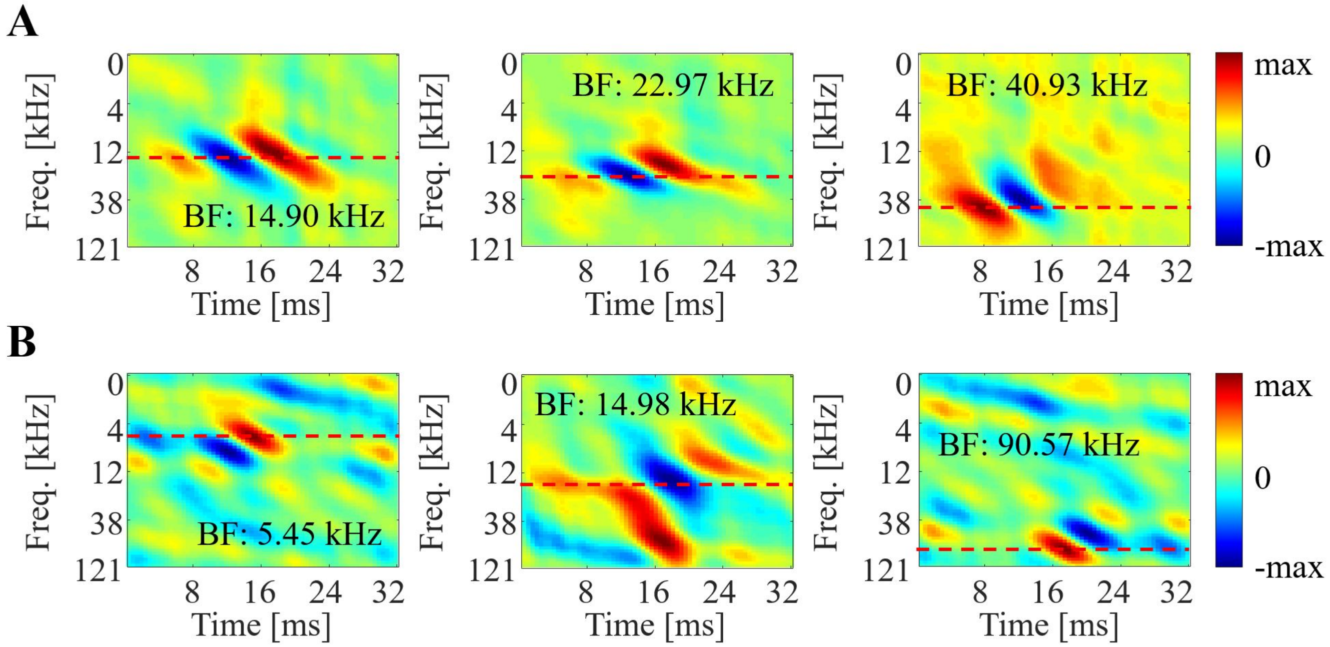
<Encoder, E>
<Sampling>
<Decoder, D>
Max-pooling
I^3
I_a^3
conv. with f^3
I^2
flatten
fc (W_3^T)
v_c
Sigmoid & *Bernoulli*
v_b
fc (W_3)
$\tilde{v}$
reshape
$\tilde{I}^3$
transposed_ conv. with $\tilde{f}^3$
$\tilde{I}^2$
Used in analysis
Max-pooling
I^2
I_a^2
conv. with f^2
I^1
flatten
fc (W_2^T)
v_c
Sigmoid & *Bernoulli*
v_b
fc (W_2)
$\tilde{v}$
reshape
$\tilde{I}^2$
transposed_ conv. with $\tilde{f}^2$
$\tilde{I}^1$
Max-pooling
I^1
I_a^1
conv. with f^1
flatten
fc (W_1^T)
v_c
Sigmoid & *Bernoulli*
v_b
fc (W_1)
$\tilde{v}$
reshape
$\tilde{I}^1$
transposed_ conv. with $\tilde{f}^1$
Stacking block
Input Patch, x
※ *fc: fully-connected layer*
Reconstructed Patch, $D(E(x))$

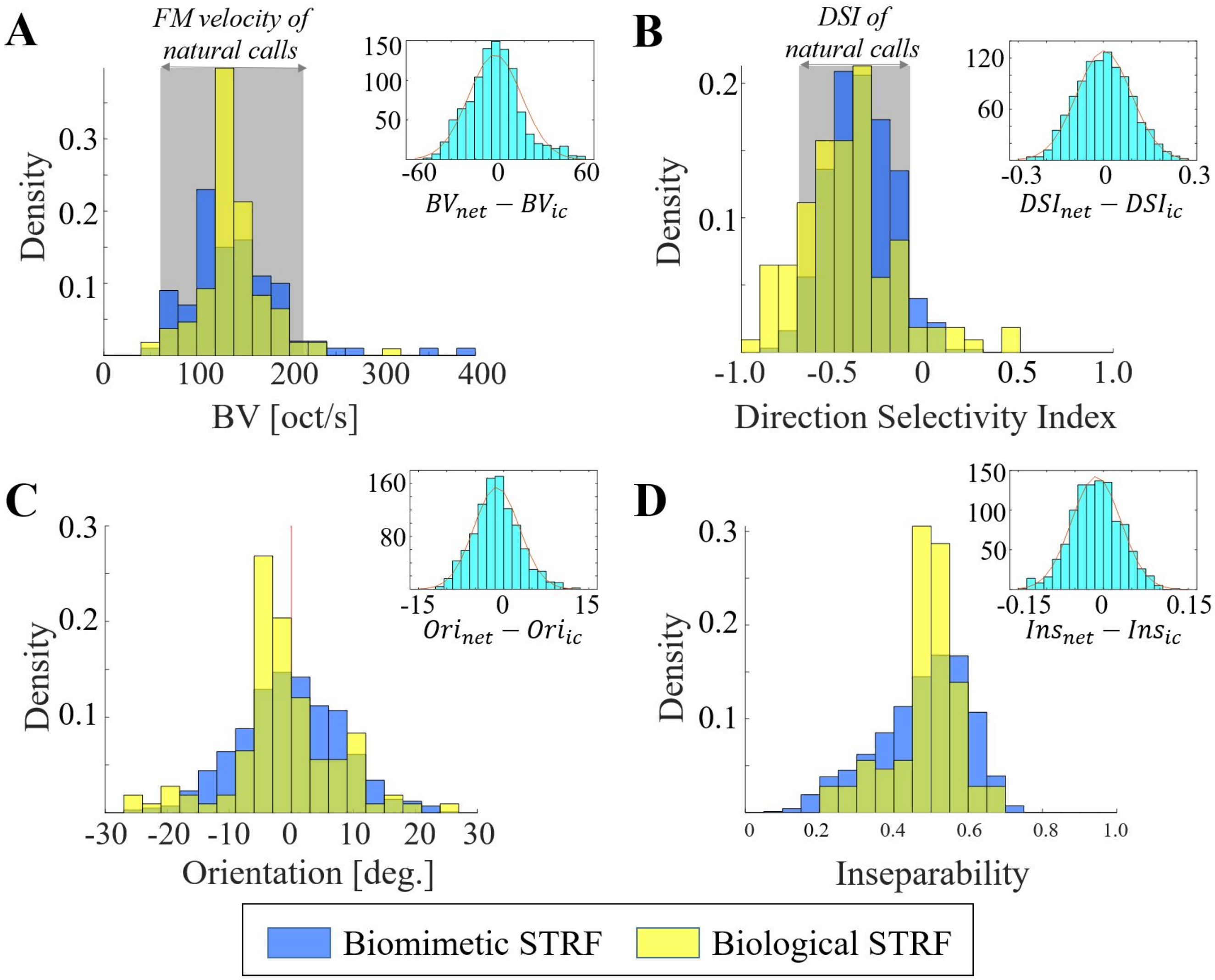

A
I_o^l
Convolution
$I_i^l = I^{l-1}$
I_o^l
Concatenating
f_1^l *
f_2^l *
...
* f_k^l
$I_i^l = I^{l-1}$
B
$\tilde{I}^l$
Transposed_
convolution
$\tilde{I}^{l-1}$
$\tilde{I}^l$
Split feature maps
$\dot{f}_1^l$ *
$\dot{f}_2^l$ *
...
* $\dot{f}_k^l$
Averaging
$\tilde{I}^{l-1}$

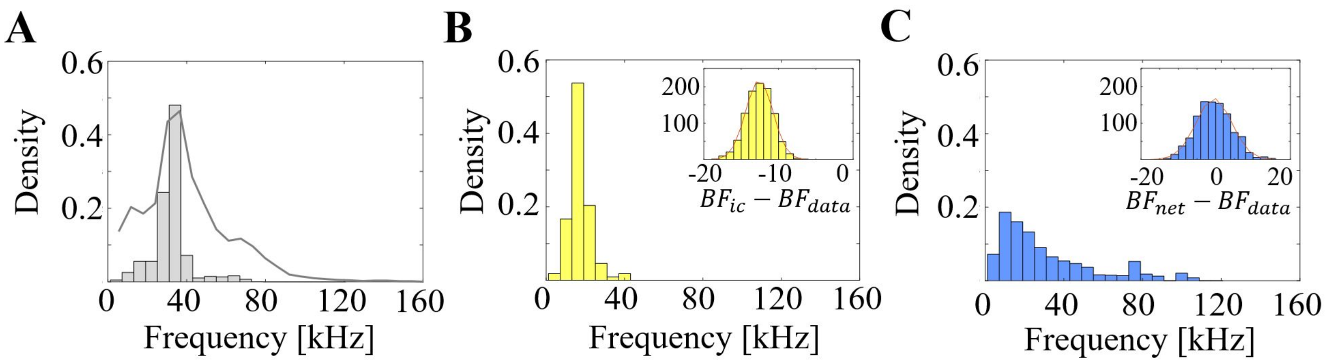

A
$M_e(\omega,\Omega)$
1.50
0.75
0.00
-0.75
-1.50
Ω [cyc/oct]
-176 -80 0 80 176
ω [Hz]
max
0
$\Phi_e(\omega,\Omega)$
1.50
0.75
0.00
-0.75
-1.50
Ω [cyc/oct]
-176 -80 0 80 176
ω [Hz]
π
0
$-\pi$
B
$\mathcal{F}^{-1}_{t,f}[T(\omega,\Omega)]$
0
4
12
38
121
Freq. [kHz]
0 8 16 24 32
Time [ms]
max
0
−max
max
0
-max
Freq. [kHz]
Time [ms]
max
0
−max

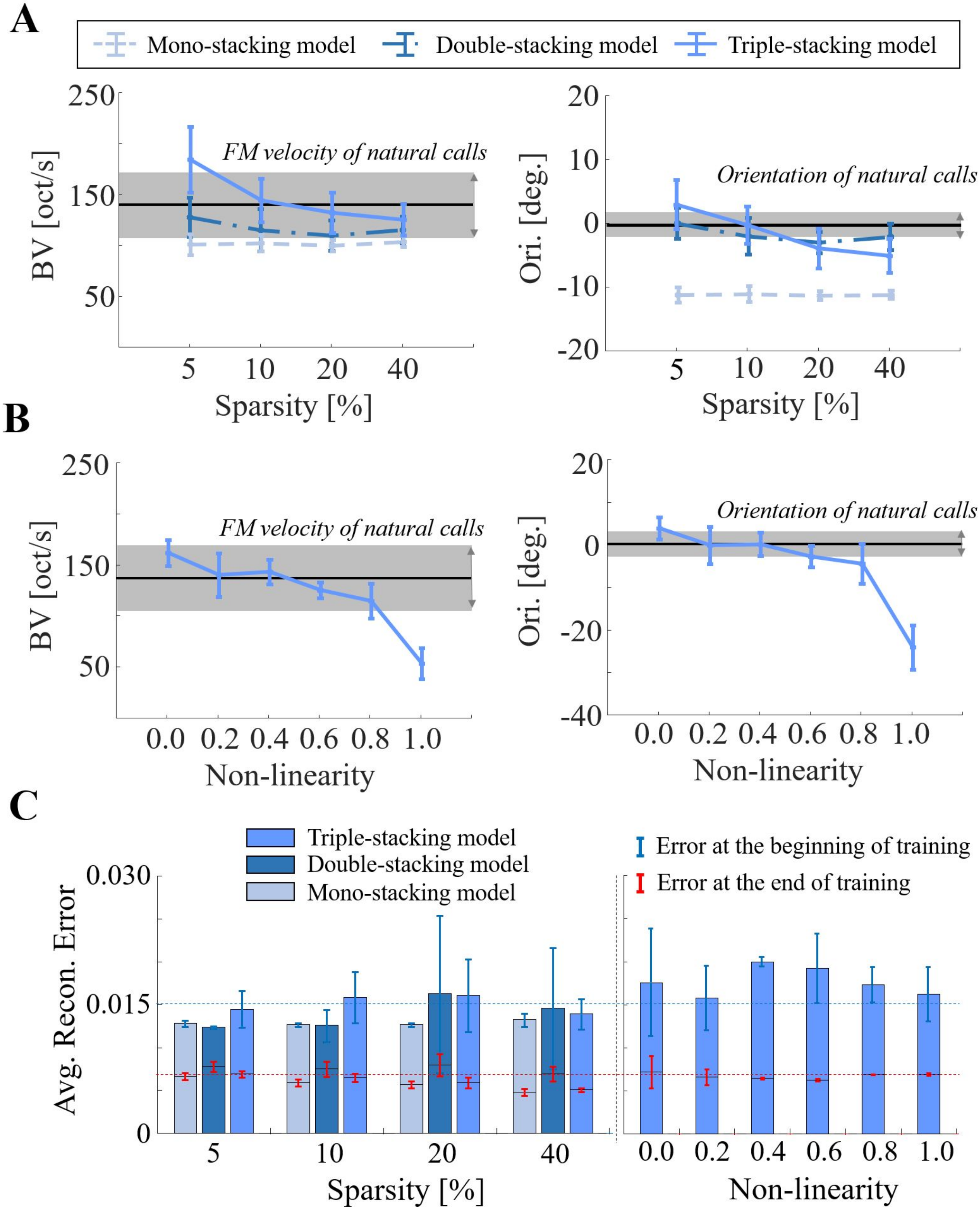

A
1.5
1.0
0.5
Ω [cyc/oct]
48
112
176
ω [Hz]
max
0
B
1.5
1.0
0.5
Ω [cyc/oct]
$Ori = \theta_\mu - \theta_\Sigma$
θ_Σ
$\boldsymbol{\mu} = [\overline{\omega}, \overline{\Omega}]$
$BV = \overline{\omega}/\overline{\Omega}$
θ_μ
48
112
176
ω [Hz]

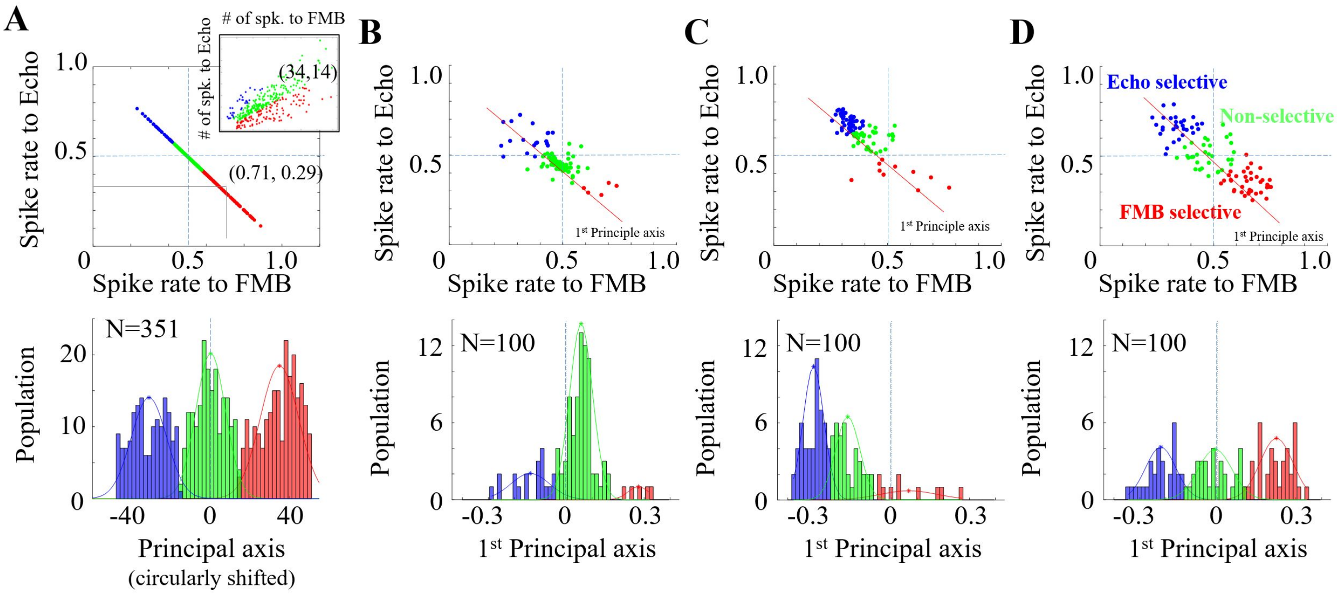

A
BV: 18 oct/s
DSI: -0.25
Peak: 13 kHz
B
BV: 43 oct/s
DSI: -0.32
Peak: 23 kHz
C
BV: 140 oct/s
DSI: -0.39
Peak: 35 kHz
D
BV: 274 oct/s
DSI: -0.49
Peak: 38 kHz
E
BV: 333 oct/s
DSI: -0.42
Peak: 70 kHz
F
BV: 450 oct/s
DSI: -0.44
Peak: 38 kHz
G
BV: 381 oct/s
DSI: -0.43
Peak: 36 kHz
H
BV: 410 oct/s
DSI: -0.42
Peak: 40 kHz
Freq. [kHz]
145
103
61
20
Time [ms]
8
16
24
32
max
0

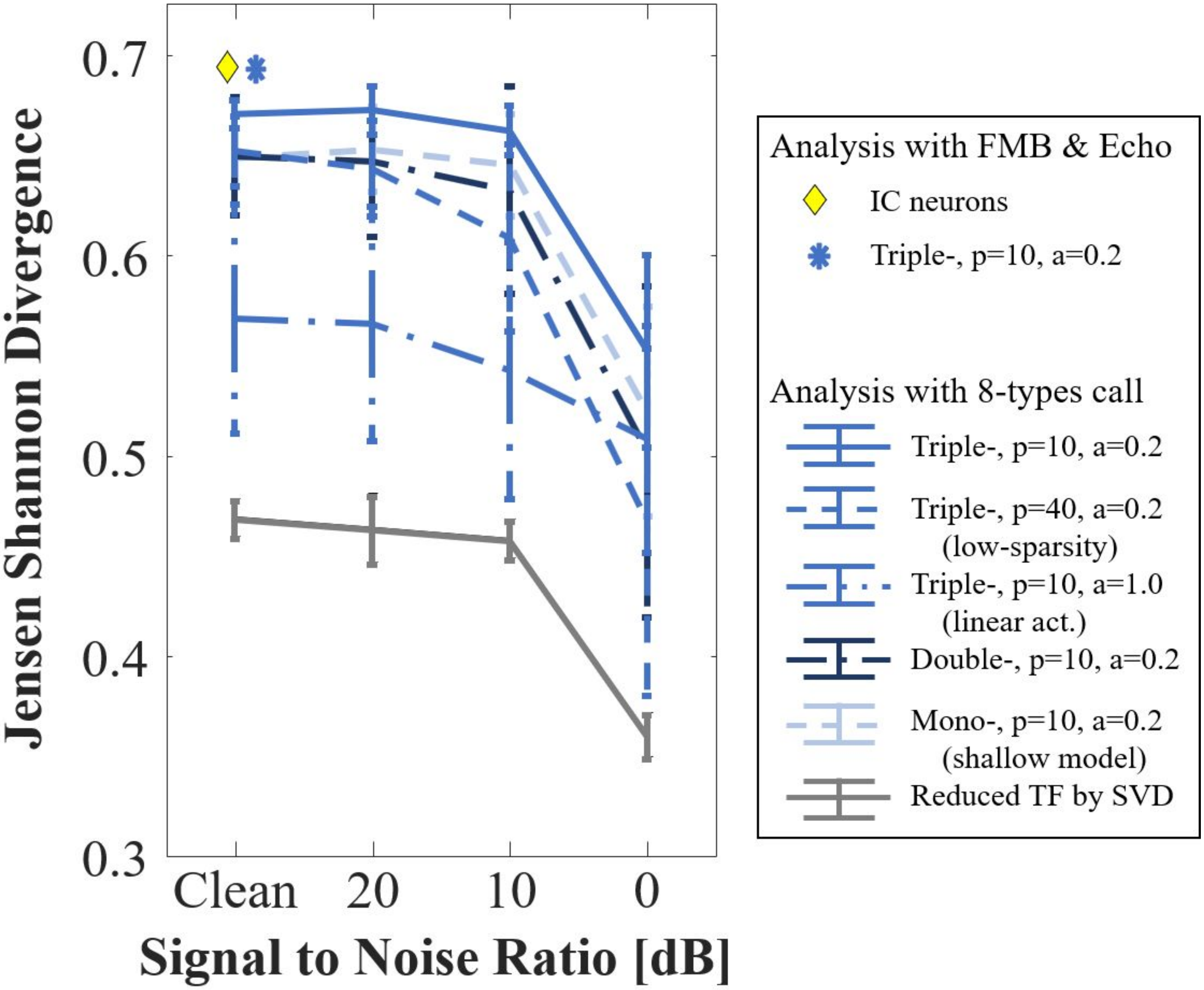

A
B
Freq. [kHz]
0
4
12
38
121
Time [ms]
8
16
24
32
BF: 14.90 kHz
BF: 22.97 kHz
BF: 40.93 kHz
BF: 5.45 kHz
BF: 14.98 kHz
BF: 90.57 kHz
max
0
-max


A
FM velocity of natural calls
Density
0.3
0.2
0.1
0
100
200
300
400
BV [oct/s]
150
100
50
-60
0
60
$BV_{net} - BV_{ic}$
B
DSI of natural calls
Density
0.2
0.1
-1.0
-0.5
0
0.5
1.0
Direction Selectivity Index
120
60
-0.3
0
0.3
$DSI_{net} - DSI_{ic}$
C
0.3
0.2
0.1
Density
-30
-20
-10
0
10
20
30
Orientation [deg.]
160
80
-15
0
15
$Ori_{net} - Ori_{ic}$
D
0.3
0.2
0.1
Density
0
0.2
0.4
0.6
0.8
1.0
Inseparability
150
100
50
-0.15
0
0.15
$Ins_{net} - Ins_{ic}$
Biomimetic STRF
Biological STRF


A
0.6
0.4
0.2
Density
0
40
80
120
160
Frequency [kHz]
B
0.6
0.4
0.2
Density
0
40
80
120
160
Frequency [kHz]
200
100
-20
-10
0
$BF_{ic} - BF_{data}$
C
0.6
0.4
0.2
Density
0
40
80
120
160
Frequency [kHz]
200
100
-20
0
20
$BF_{net} - BF_{data}$


A
Mono-stacking model
Double-stacking model
Triple-stacking model
BV [oct/s]
250
150
50
FM velocity of natural calls
5
10
20
40
Sparsity [%]
Ori. [deg.]
20
10
0
-10
-20
Orientation of natural calls
5
10
20
40
Sparsity [%]
B
BV [oct/s]
250
150
50
FM velocity of natural calls
0.0
0.2
0.4
0.6
0.8
1.0
Non-linearity
Ori. [deg.]
20
0
-20
-40
Orientation of natural calls
0.0
0.2
0.4
0.6
0.8
1.0
Non-linearity
C
Triple-stacking model
Double-stacking model
Mono-stacking model
Error at the beginning of training
Error at the end of training
Avg. Recon. Error
0.030
0.015
0
5
10
20
40
Sparsity [%]
0.0
0.2
0.4
0.6
0.8
1.0
Non-linearity


A
# of spk. to FMB
# of spk. to Echo
(34,14)
1.0
0.5
0
0.5
1.0
Spike rate to Echo
Spike rate to FMB
(0.71, 0.29)
N=351
20
10
0
Population
-40
0
40
Principal axis
(circularly shifted)
B
Spike rate to Echo
1.0
0.5
0
0.5
1.0
1st Principle axis
Spike rate to FMB
N=100
12
6
0
Population
-0.3
0
0.3
1st Principal axis
C
Spike rate to Echo
1.0
0.5
0
0.5
1.0
1st Principle axis
Spike rate to FMB
N=100
12
6
0
Population
-0.3
0
0.3
1st Principal axis
D
Spike rate to Echo
Echo selective
Non-selective
FMB selective
1.0
0.5
0
0.5
1.0
1st Principle axis
Spike rate to FMB
N=100
12
6
0
Population
-0.3
0
0.3
1st Principal axis


0.7
0.6
0.5
0.4
0.3
Jensen Shannon Divergence
Clean
20
10
0
Signal to Noise Ratio [dB]
Analysis with FMB & Echo
IC neurons
Triple-, p=10, a=0.2
Analysis with 8-types call
Triple-, p=10, a=0.2
Triple-, p=40, a=0.2
(low-sparsity)
Triple-, p=10, a=1.0
(linear act.)
Double-, p=10, a=0.2
Mono-, p=10, a=0.2
(shallow model)
Reduced TF by SVD