Slow and steady: auditory features for discriminating animal vocalizations

Ronald W. Di Tullio*,[1,2] Linran Wei*,[3] and Vijay Balasubramanian[1,2,4]

[1] *David Rittenhouse Laboratory, Department of Physics and Astronomy, University of Pennsylvania, USA*
[2] *Computational Neuroscience Initiative, University of Pennsylvania, USA*
[3] *David Rittenhouse Laboratory, Department of Physics and Astronomy, University of Pennsylkvania, USA*
[4] *Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA**

We propose that listeners can use temporal regularities – spectro-temporal correlations that change smoothly over time – to discriminate animal vocalizations within and between species. To test this idea, we used Slow Feature Analysis (SFA) to find the most temporally regular components of vocalizations from birds (blue jay, house finch, American yellow warbler, and great blue heron), humans (English speakers), and rhesus macaques. We projected vocalizations into the learned feature space and tested intra-class (same speaker/species) and inter-class (different speakers/species) auditory discrimination by a trained classifier. We found that: 1) Vocalization discrimination was excellent ($> 95\%$) in all cases; 2) Performance depended primarily on the $\sim$10 most temporally regular features; 3) Most vocalizations are dominated by $\sim$10 features with high temporal regularity; and 4) These regular features are highly correlated with the most predictable components of animal sounds.

I. INTRODUCTION

All animals communicate with other individuals of their own and other species. Many communicate through vocalizations – vibrations, generated by intentionally contracting muscles, that propagate through the surrounding medium. These vibrations are received by the auditory systems of other individuals, and are processed to yield percepts ranging from threat, courtship interest and fitness, food quality, and abstract concepts [1–6]. Animal vocalizations vary substantially across species, and even between individuals of a single species. Nevertheless, animals readily learn to discriminate and parse the vocalizations of conspecific and, to a varying extent, allospecific individuals [7–12] How can animals learn to discriminate vocalizations so broadly and effectively despite their substantial variation?

We propose that animal vocalizations are generally learnable because they share underlying structure – temporal regularity, understood here as coincidence (spectro-temporal correlation) and continuity (coherent variation of the correlations over time). This idea is consistent with Gestalt principles of auditory perception [13–16], and suggests computational mechanisms by which these principles could be realized in auditory systems.

To test this idea, we first use Slow Feature Analysis (SFA) [17–21] to learn temporally regular acoustic features from vocalizations produced by birds (American yellow warbler, blue jay, great blue heron, and house finch), humans (English speakers), and macaques. We use SFA, which assumes that information relevant to object identity changes slower than nuisance variability, because we have shown classifiers based on SFA features can discriminate macaque vocalizations against a background chorus [22].

We then demonstrate that a basic feedforward classifier network using features derived from the simplest form of SFA can learn to discriminate vocalizations across species, achieving $> 95\%$ classification accuracy. To determine what is driving classification performance, we distort subsets of slow features and find that the "slowest" (most temporally regular) 10 features have the greatest impact on performance. We confirm that a network trained using this subset alone has performance nearly equal to a network using all features.

We then ask why this particular subset of features is important for classification, and find that, for each animal group (birds, humans, macaques), vocalizations are largely comprised of around 10 very slow features with a larger number of moderate to fast features of decreasing importance to the acoustic structure. Finally, we show that the slow features of vocalizations are also the predictable features. This last result does not follow automatically from the slowly varying nature of acoustic patterns uncovered by SFA, since features that change rapidly but in an orderly manner are also predictable.

Our results suggest the need for new experiments to (a) determine how and where brain circuits analyze and extract temporally continuous features of acoustic patterns, and (b) study how auditory circuits are adapted to the ethology of communication. Such studies will need to treat auditory objects as dynamical entities with temporally changing patterns, rather than as static acoustic signals such as fixed combinations of frequencies.

II. RESULTS

A. Temporal regularity in vocalizations

Broadly, we propose that auditory objects are defined by temporal regularities [22–27] – coincidence, correlation, and temporal continuity of acoustic features (Fig. 1A). To test this idea for animal vocalizations, we used public data (Methods - Auditory Stimuli) for several species – birds (great blue heron, house finch, American

* These authors contributed equally to this work

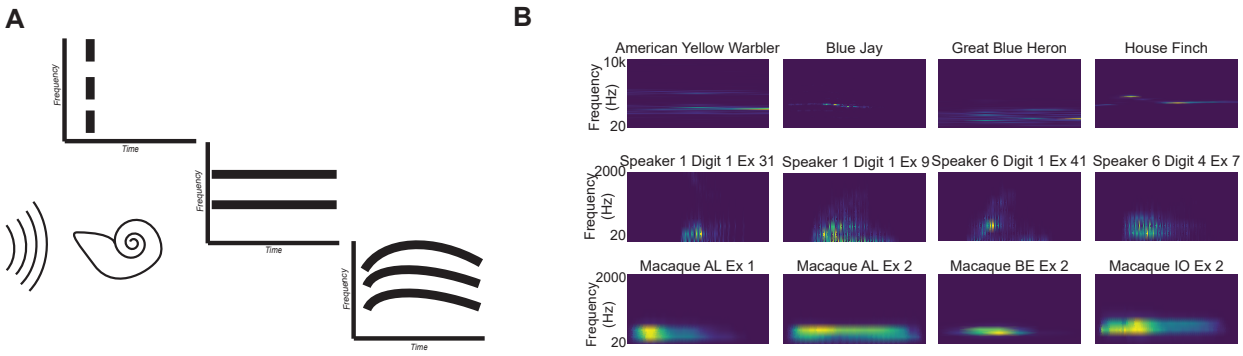

FIG. 1. **Coincidence and continuity – vocalization spectrograms. A.** *Left.* Schematic of auditory stimuli received by the auditory system (cartoon cochlea). *Middle.* Coincident onset of three frequency bands. *Right.* Coincident onset and stable persistence of three frequencies. *Far Right.* Continuously varying correlations between the frequency components of an auditory stimulus. **B.** Example spectrograms for birds, humans, and macaques. Warmer colors indicate more energy and cooler colors indicate less energy at each time-frequency point. Spectrograms have different durations to display details of each vocalization (durations are $\sim$ 50 ms for birds and up to 500 ms for macaques). *Top.* Calls from four bird species. *Middle.* English speakers pronouncing digits 0-5. *Bottom.* Macaques producing *coo* vocalizations. Different examples (Ex) of the same vocalization by a given speaker shown for humans and macaques. Spectrograms calculated via the `spectrogram` function from the `scipy.signal` Python module using the vocalization sampling rate and default parameters (a Tukey window with shape parameter 0.25, time bins of roughly 7.5 ms (due to our sample rates), and no overlaps between time bins)

.

yellow warbler, and blue jay), humans (English speakers saying digits 0-5), and macaques (producing *coo* vocalizations).

We first analyzed the acoustic structure of these vocalizations in terms of their spectrograms, that is, the Fourier power spectra in small temporal windows [15, 22, 23, 25, 28–30] (Fig. 1B). By inspection, although the sounds produced by each species differ in detail, all of them manifest frequency correlations that change in a regular manner over time. Similar structure is revealed by processing the sounds through a model cochlea to arrive at a "cochleagram" (Methods - Simple Cochlear Model for more details), which likewise produces a time-varying frequency power spectrum.

To quantify these observations we calculated the average modulation power spectrum (MPS, 2d Fourier power spectrum of the cochleagram, Fig. 2A) for vocalizations within animal groups (birds, humans, macaques). The MPS displays acoustic energy in combined temporal and frequency bands. The vocalizations we examine are dominated by components with low temporal modulation (Fig. 2B, upper panel). Averaging across spectral modulation for each group (Fig. 2B, lower panel) makes this concentration even clearer. We next asked how these temporal regularities could be exploited to support discrimination between vocalizations.

B. Discriminating vocalizations

To test auditory learning we train animals to discriminate between vocalizations after exposure to examples. Following this logic, we asked if a network classifier could discriminate between pairs of vocalization on the basis of learned, temporally regular features.

First, we constructed inter- and intra-class vocalization pairs for each animal group. For birds, intra-class pairs were from the same species while inter-class vocalizations were from different species; for humans intra-class pairs were different or identical digits said by an individual and inter-class pairs were the same digit said by different speakers; for macaques we used a similar definition as humans – intra-class pairs were different examples of *coos* from a speaker and inter-class pairs were *coos* from different macaques (Methods - Inter- and Intra- Class Discrimination Task). We constructed 100 examples of each kind of pair for each animal group (300 inter-class and 300 intra-class pairs in total). Next we used a simple cochlear model (Methods - Simple Cochlear Model) to analyze and generate a cochleagram for each vocalization. We concatenated the cochleagrams for each vocalization in a pair and applied Slow Feature Analysis (SFA).

SFA solves an optimization problem to represent inputs in terms of a hierarchy of signal features with the least temporal variation. We chose a form of SFA in which the features are orthogonal linear combinations of the input data (Methods – Slow Feature Analysis). SFA can be applied with non-linear features – [22] analyzes macaque calls using quadratic SFA – but we chose linear features which are computationally tractable both for brains and computers, and easily interpretable. Linear SFA returns as many features as there are input components (42 in our case), allowing us to decompose the signal into a sum of features. We initially used the 20 slowest features, because, if slow features efficiently capture distinguishing structure in vocalizations, good clas-

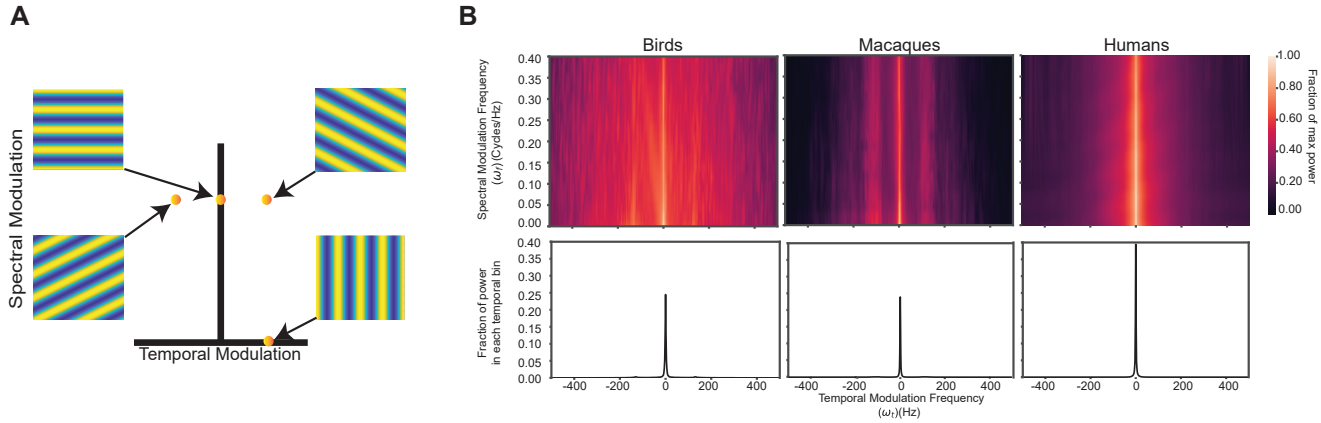

FIG. 2. **Coincidence and continuity – modulation spectra. A.** A spectrogram or cochleagram can be decomposed as a weighted sum of spectral- and temporal-modulation rates. The spectral-modulation reflects how rapidly the spectrogram changes over frequencies at an instant of time. The temporal-modulation reflects how fast the spectrum changes over time. If a spectrogram had only spectral modulation but no temporal modulation, it would be a point on the ordinate of the spectral-temporal modulation space, with a spectrogram as shown at the upper left. If a spectrogram had only temporal modulation but no spectral modulation, it would be a point on the abscissa of the spectral-temporal modulation space, with a spectrogram as shown on the lower right. The other two spectrograms are examples with both spectral- and temporal-modulation. Color represents energy at each time-frequency point. **B (Upper).** Average spectrotemporal modulation spectra for each animal group. **B (Lower).** Average spectral power as a function of temporal modulation. Modulation spectra are calculated following [23, 28]. 200 vocalizations were used to generate each plot. These vocalizations comprised the pairs were used in all subsequent analyses and figures unless stated otherwise (Methods - Auditory Stimuli).

sification should be achieved with a subset of the features.

The time-varying, transient nature of acoustic objects suggests that the auditory system must perform dynamically changing computations [13, 16, 26, 27, 31, 32]. Indeed, stream segregation, in which mixed inputs are parsed into sources and objects based on attentional goals, is at the heart of auditory perception [13] and seems to require such computations [11, 15, 16, 26, 33–35]. Previous work has shown that SFA can be implemented in an online fashion that dynamically calculates the slowest features of time-varying input signals [18, 36]. Nevertheless, for computational simplicity, we used a batch version of SFA, but to mirror, at least partially, the dynamism of the auditory system and online SFA, we reapplied the algorithm to each vocalization pair individually. Thus the algorithm approximated the dynamic temporal feature extraction that we expect in audition.

We then projected each vocalization pair onto its 20 slowest SFA features and used this projection to train a simple multilayer perceptron (MLP, Methods - Multilayer Perceptron, Fig. 3A). The MLP had an input layer for the SFA features used, a compressive five node hidden layer, and a single node classification layer. Each hidden layer had a rectifying linear (ReLu) nonlinearity and the output node used a soft-max nonlinearity. We chose this network to have a simple architecture that could be realized in the brain, with basic, neurally-implementable non-linearities, moderate compression of inputs, and a probabilistic decision about object identity. Animals in noisy environments appear to parse vocalizations from small segments that occur within "dips in the noise" [37–44]. To approximate this ability, we trained our network on randomly selected portions of the vocalizations in each pair and trained it to identify the left out portions (details in Methods - Multilayer Perceptron). We repeated this process five times and averaged the percent correct across repeats. This paradigm allowed us to assess the quality of the SFA representation, and its ability to support vocalization discrimination in noisy environments.

Our classifier achieved high performance for all animal groups and both inter-class and intra-class discrimination. The results were well above chance (Fig. 3B *left*): Birds Intra-class 99.87% (99.71-99.96%); Humans Inter-class 98.39% (97.40- 99.3%); Humans Intra-class 97.70% (96.23-98.87%); Macaques Inter-class 99.94% (99.90-99.96%); Macaques Intra-class: 99.94% (99.90-99.96%) – Median and Interquartile range indicated. The greatest variance in performance occurred for human vocalizations. These findings show that the vocalizations can be discriminated on the basis of their slow features.

C. The slowest features are best

We expected that the slowest features should be most important for vocalization discrimination. To test this prediction, we again projected vocalizations pairs onto their learned features and trained the network. Then during testing, just prior to classification, we randomly shuffled the feature labels in groups of five from the slowest to the fastest of the first 20 features. This shuffling procedure interfered with the mapping between features

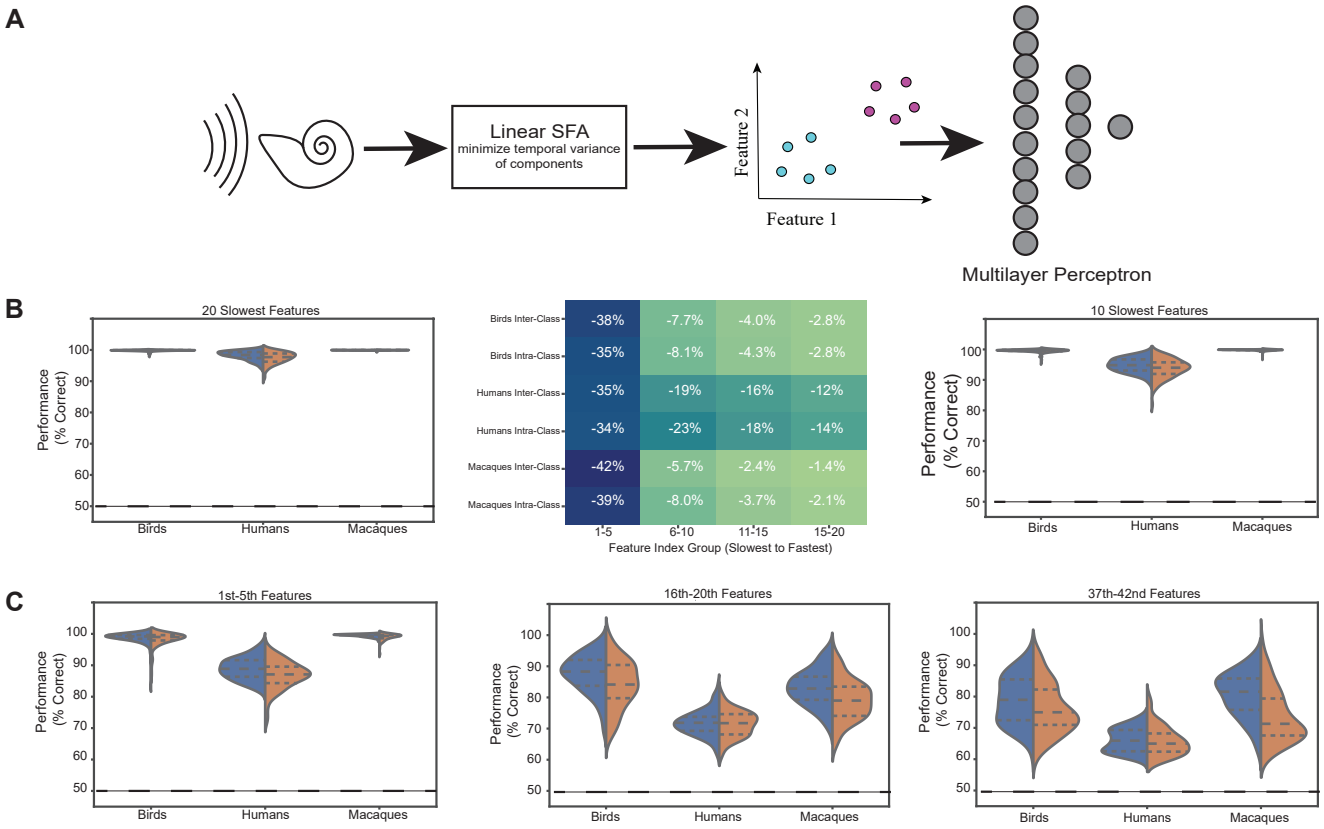

FIG. 3. **Analysis procedure and discrimination results. A.** We passed auditory stimuli through a simplified cochlear model (a spectral filter followed by a temporal filter, Methods - Simple Cochlear Model). We applied linear SFA to the output, and chose a subset of learned features, ordered by slowness to generate a feature space. We projected test stimuli into this feature space and trained a simple multilayer perceptron (input layer, 5 node hidden layer, output node) on the projected signals using five-fold, stratified shuffle split cross-validation (Methods - Multilayer Perceptron) to separate data points from two vocalizations. For each vocalization pair, we chose different snippets as training and test stimiuli (Methods - Discriminating Between Vocalizations). **B.** *Left:* Violin plots = % correct for inter-class (blue) and intra-class (red) classification with 20 slowest features on 100 vocalization pairs. Birds: intra-class stimuli from the same species; inter-class stimuli from different species. Humans: intra-class stimuli = same or different digits from one speaker; inter-class stimuli = same digit from different speakers. Macaques: intra-class stimuli = different *coo*s from an individual; inter-class stimuli = *coo*s from different individuals. Median (central dashed line) and interquartile range (upper and lower dashed lines) indicated. Chance performance is 50%. *Middle:* Heat map of performance decrease caused by shuffling the feature set on the x-axis before classification. Darker blues = greater performance decrease. Rows = animal group and classification types. *Right:* Same as left panel, but for classification with the 10 slowest features. **C.** Same as left panel of **B** but with subsets of five progressively faster features.

and vocalization identity learned by the network during training. We then tested the network and measured the percent decrease in classification performance as a function of which subset of features were shuffled. We saw that the five slowest features had the greatest impact on performance followed by the fifth-tenth slowest features. Fig. 3B *Middle* shows the average *decrease* in performance for each subset of features. Human vocalizations show the largest decrease in performance when any subset of features is shuffled, consistent with the higher variability in classification performance for this group.

These results suggested that a network trained on the ten slowest features would achieve performance similar to the complete network in the previous section. Indeed, the reduced feature set led to similar performance (Fig. 3B *Right*): Birds Inter-class 99.69% (99.53- 99.85%) ; Birds Intra-class 99.65% (99.39- 99.89%); Humans Inter-class 94.83% (93.08- 96.72%); Humans Intra-class 94.00% (91.19- 95.75%); Macaques Inter-class 99.86% (99.78- 99.92%); Macaques Intra-class 99.85% (99.73- 99.92%) – Median and Interquartile range indicated.. The slight increase in variance is expected because we are using fewer features and hence fewer parameters [45, 46]. Finally, we tested the change in performance when progressively faster groups of five features were used (Fig. 3C). Performance decreased and variability increased monotonically with feature speed. Taken together, these results show that the slowest features of vocalizations are the most useful for discrimination.

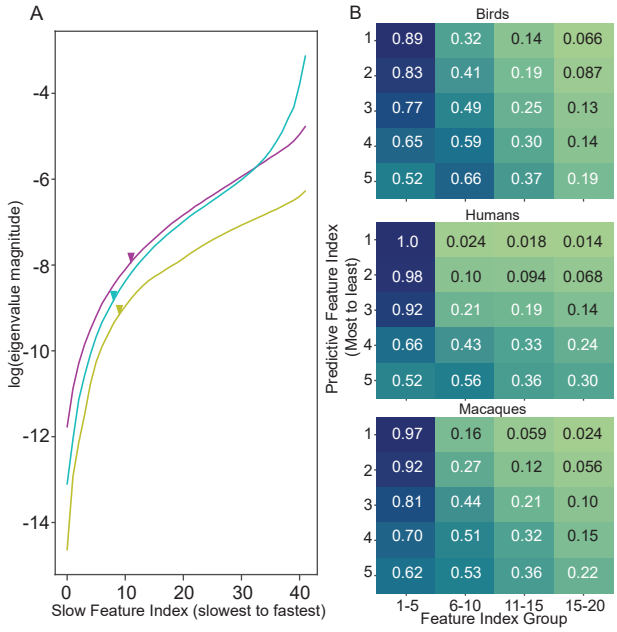

FIG. 4. **Slowness and predictability. A.** Plot of the logarithm of the eigenvalues (y-axis) associated with each slow feature index (x-axis) for each animal group. A smaller eigenvalue indicates a slower feature. Magenta = birds; Yellow = humans; Cyan = macaques. **B.** Canonical Correlation Analysis (CCA) heatmaps (Methods - PFA to SFA Comparison) comparing slow (x-axis) and predictive features (y-axis) for birds, humans and macaques. Smaller indices indicate slower/more predictive features and darker blues indicate higher correlation. Cells show the CCA score (range 0-1 = zero to perfect correlation).

D. Vocalizations consist of few, very slow features

In view of the above, we asked whether vocalizations are mostly comprised of a small number of very slow features or if slowness is evenly distributed across features. To test, we formulated SFA as a generalized eigenvalue problem (see Methods - Slow Feature Analysis and Methods - Eigenvalue Analysis), extracted the distribution of eigenvalues for each vocalization, and averaged within each animal group. Fig. 4A shows a sharp decrease in the slope of the eigenvalue distribution for all animal groups at around the tenth slowest feature. An inflection point analysis using the KNEEDLE algorithm [47] confirms this observation for each animal group (Birds: 11th, Humans: 9th, Macaques: 8th; colored triangles in Fig. 4A). Noting that smaller eigenvalues indicate slower features, we see that vocalizations are comprised of a small number of very slow features. This explains the concentration of power at low temporal modulation in Figs. 1, 2 and the sufficiency of about 10 slowest SFA features for vocalization discrimination 2B.

E. Predictability equals slowness

An alternative hypothesis we could have followed is that animals learn to discriminate vocalizations by extracting the *predictable* components of sounds. Predictability is not the same as continuity or smoothness, because highly discontinuous signals can be completely predictable from their pasts if they are deterministically generated. Thus we tested the relationship between the predictable and slow features of vocalizations.

We decomposed individual vocalizations using Predictable Feature Analysis (PFA)[48, 49] as well as SFA, and used Canonical Correlation Analysis (CCA) [50–53] to compare PFA and SFA features. Briefly, CCA compares multidimensional data sets by finding projections maximizing correlation between the sets (Methods - PFA SFA Comparison). We compared sets of five SFA features, from the slowest five to the 16th-20th slowest, to each of the five most predictable features, effectively reducing CCA to linear regression of PFA features with SFA feature subsets [53]. The first five SFA features correlated highly with the three most predictable PFA features (Fig. 4B). The fourth and fifth most predictable feature correlated highly with both the SFA features 1-5 and 6-10. By construction, subsequent SFA and PFA features are orthogonal to all preceding features. Thus, we expect faster SFA features to correlate more with less predictable features.

These results show convincingly that learning the slow features of vocalizations simultaneously learns their predictable features and, conversely, that the predictable features of vocalizations are temporally slow.

III. DISCUSSION

In this article we used Slow Feature Analysis to isolate the most temporally regular components of vocalization pairs from different species of birds, humans, and macaques, projected these pairs into the SFA feature space, and trained a simple multilayer perceptron to classify the projected data points. This simple network achieved excellent performance (median $> 95\%$) for all animal groups with just the ten slowest features. We found further that using faster features systematically degraded classification, that vocalizations appear to be mostly comprised of ~ 10 slowest features, and that the slowest features and most predictable features span each other. All told these results support our proposal that animal vocalizations can be generally learned and discriminated because they share underlying temporal regularities – coincidence (spectro-temporal correlation) and continuity (coherent variation of correlations over time).

At a behavioral level, our results suggest several avenues for research. First, some animals attempt to "counter" or mask vocalizations of others. For example, male cowbirds sometimes sing over competing males and females can produce "chatter" to block unwanted male

songs [3, 54]. Perhaps such chatter calls or singing-over behaviors specifically target slow song features, affecting perception by conspecific individuals. Many researchers have probed how changes to auditory stimuli affect auditory object perception, but experiments have generally used artificial stimuli because the temporal statistics of natural stimuli have not been systematically parameterized [13, 26, 55–57]. Our results suggest that manipulating slower features of vocalizations, learned through SFA analysis, will have stronger behavioral effects than manipulating faster features, thus providing a systematic basis for design of experiments. More broadly, our results suggest that behaviorally relevant sounds have critical acoustic structure that auditory neuroscientists should analyze and account for in experimental stimuli, adding to a general sense in the field that ethology and neurosicence should interact more closely [34, 57–68]. In particular, we should recognize that many auditory stimuli that drive behavior are dynamic objects determined by how they change in time, and that the neural circuits processing them are likely poorly summarized by static, time-averaged descriptions like spectro-temporal receptive fields or Convolutional Neural Networks.

At a mechanistic level, our results suggest that circuits that tune rapidly to temporal modulation structure in sounds could perform key auditory tasks like vocalization discrimination. There is some evidence for such circuits in the auditory system [25, 27, 29, 30, 32, 33] but their dynamics and function are not clearly understood. As discussed in [22], a population of cells responding to a range of temporal modulations could implement SFA-like computations [18, 36]. It would be interesting to determine whether or not such a population exists, and where along the auditory pathway it is located. Because SFA is so computationally simple, we could also directly explore whether neural activity correlates with slow features of a stimulus. To do so, we could calculate the slow features, project the stimulus into the SFA feature space, and use the feature magnitudes as regressors to predict firing rates of neurons recorded during stimulus presentation. If the SFA regressors capture significant neural variance, these cells, or perhaps the circuit pathway leading to them, may encode SFA features.

At a theoretical level, our results inform the use of frameworks like efficient coding theory to understand the functional organization of neural circuits. The broad idea of efficient coding is that neural circuits are costly, and were selected in a manner that produces the greatest benefit for the investment [69]. This approach has explained many aspects of early vision, including nonlinearities in the fly visual system [70], center-surround receptive fields of neurons in the vertebrate retina [71–77], spike timing statistics [78], the preponderance of OFF cells over ON cells [79, 80], the mosaic organization of ganglion cells [81, 82], the scarcity of blue cones and the large variability in numbers of red and green cones in humans [83], selection of predictive information by ganglion cells [84, 85], and the expression of ion channels in insect photoreceptors [86]. Similar analyses suggest that the olfactory periphery [87–89] is also adapted to the statistical structure of the environment to use limited resources efficiently to represent odor information. In visual cortex, adaptation to natural stimulus statistics seems to explain the functional structure of edge detectors in V1 [90], and texture analyzers in V2 [91–93]. Even the grid cells of the entorhinal cortex appear to be efficiently organized to represent space [94]. There are surprisingly few comparable studies of auditory circuits [95–98], but our studies reveal rich temporal structure in ethologically relevant stimuli. Some parts of the auditory pathway may be adapted to efficiently process these structures.

IV. METHODS

A. Auditory Stimuli

We used publicly available vocalization datasets for humans, macaques, and birds (blue jay, house finch, American yellow warbler, and great blue heron). The human vocalizations included six English speakers pronouncing digits zero through five 50 times each (a total of 50*6 = 300 pronunciations per person) [99]. Macaque vocalizations were all *coos* [1, 22, 28]. The data set included 8 individuals with approximately 1000 coos from each individual.[100] Bird vocalizations were not labeled by individual, but were grouped by species, and included approximately 250 song examples from each species [101], Each animal group was recorded at a different sampling rate (humans at 48 kHz, birds at 44.1 kHz, and macaques at 32 kHz, respectively). These differences were accounted for in our cochlear model by including the appropriate sampling frequency in the gammatone filter bank (see below). We selected vocalizations randomly from these dataset to create vocalization pairs. If vocalizations were accidentally paired with themselves, that pair was eliminated from further analysis. This was rare (occurring $< 5\%$ of the time) and resulted in rejection of 29 of the 600 total pairs. See Inter- and Intra- Class Discrimination Task below for more details.

B. Simple Cochlear Model

We implemented a simple cochlear model to approximate the structure of inputs to the auditory system. The model included two filter stages: 42 gammatone spectral filters followed by a temporal filter following previous work [22, 96, 102, 103]. Each of the 42 filter outputs (42 spectral $\times$ 1 temporal filters) was normalized to zero mean and unit variance. The gammatone filters were implemented using the `GammatoneFilterbank` function from python module `pyfilterbank.gammatone`. The gammatone filters had center frequencies between 22.9 Hz to 20208 Hz, covering the frequency range of

all vocalizations studied. The temporal filter was implemented as a difference of two kernels of the form

$$g(n) = a\, n^m\, e^{-bn}\,, \tag{1}$$

where n is in units of samples and a, b, and m are shape and rate parameters. The temporal filter was created by taking the difference between filters g_1 ($a = 1.5$, $b = 0.04$, and $m = 2$) and g_2 ($a = 1$, $b = 0.036$, and $m = 2$). These parameters mimicked aspects of cochlear temporal processing [103, 104]. Each filter output was normalized to zero mean and unit standard deviation.

C. Slow Feature Analysis.

The conceptual underpinning of Slow Feature Analysis (SFA) is the "slowness principle" [17–19], which hypothesizes that higher-order information in a stimulus (e.g., stimulus identity) changes more slowly than nuisance fluctuations. Linear SFA is defined by the equations

$$\begin{aligned} y_{(m,t)} &= \mathbf{w}_m \mathbf{x}_t \\ \mathbf{w}_m &= \text{argmin}_{\mathbf{w}_m} \left(\langle (y_{(m,t)} - y_{(m,t-1)})^2 \rangle \right) \end{aligned}$$

where $\mathbf{x}_t$ is an auditory stimulus (or set of stimuli) at each time t, $\mathbf{w}$ is the set of slow features, $y_{(m,t)}$ is the projection of the stimulus into the feature space defined by $\mathbf{w}$, and $\langle\rangle$ is the expected value. "Slowness" is realized by finding the particular set of features that minimizes the average squared temporal difference of each m component of y_m. Because the temporal difference $y_{(m,t)} - y_{(m,t-1)}$ is the discrete analog of a derivative, the minimization in the second line selects features to generate a projected output that changes most slowly from moment to moment. More complex versions of SFA replace the first equation with a nonlinear feature extractor.

The optimization problem above is subject to three constraints, which ensure non-trivial solutions:

$$\begin{aligned} \langle y_{(m,t)} \rangle &= 0 : \text{zero mean} \\ \langle y_{(m,t)} y_{(m,t)} \rangle &= 1 : \text{unit variance} \\ \forall m < j : \langle y_{(m,t)} y_{(j,t)} \rangle &= 0 : \text{decorrelation} \end{aligned}$$

The second and third constraints ensure that the SFA features are linearly independent and order them by slowness, similarly to the ordering induced in principal component analysis (PCA). If the number of features is taken to be smaller than input dimension SFA performs dimensionality reduction.

Also similarly to PCA, SFA can be formulated as a generalized eigenvalue problem. For PCA, we seek eigenvectors of the data correlation matrix with the largest eigenvalues as these capture the most data variance. For SFA, we instead seek eigenvectors of the temporal difference correlation matrix (i.e., eigenvectors of $\Delta X \Delta X^T$ where $\Delta X = X_{t+1} - X_t$) with the *smallest* eigenvalues. These capture directions of least temporal variance, i.e. where the signal changes slowest. For a comparison of PCA and linear SFA see [105].

D. Inter- and Intra- Class Discrimination Task

To determine whether slow features were sufficient for auditory discrimination, we created a battery of test pairs at random for each animal group. We considered inter-class and intra-class pairs reflecting two broad kinds of discrimination/classification problems: discriminating between different clusters (inter-class) vs discriminating between different examples within a cluster. For humans, we defined inter-class pairs as different speakers saying the same digit and intra-class pairs as the same speaker saying different digits or giving different pronunciations of the same digit. For macaques, we used a similar definition: inter-class pairs were different individuals producing *coo*s while intra-class pairs were different *coo* examples from an individual. For birds, we defined inter-class pairs as vocalizations from different species and intra-class pairs as vocalizations from the same species. Inter- and intra-class vocalizations pairs were analyzed the same way: each member of the pair was passed through the cochlear model to produce a cochleagram, i.e., a matrix showing the power in each frequency group at each time. The cochleagram for each vocalization in the pair was then horizontally concatenated together and analyzed by linear SFA. Next, the vocalization pair was projected into a subset of the feature space found by SFA. This projection results in a cloud of data points corresponding to the SFA projection of the cochleagram at a moment of time. Subsets of these points were then used to train a multilayer perceptron (MLP) to discriminate one vocalization from the other in a pair (see below). The SFA algorithm and the MLP were reset for each vocalization pair.

E. Multilayer Perceptron

We constructed a simple multilayer perceptron (MLP) to discriminate between vocalization pairs. This MLP consisted of an input layer with as many nodes as the number of SFA features used, a five node hidden layer, a single node output layer. Each hidden layer node used a rectifying linear (ReLu) nonlinearity and the output node used a soft-max nonlinearity. The vocalizations within each pair were randomly assigned the label 0 or 1 and data points belong to that vocalization (cochleagrams frames projected onto the SFA feature space) were labeled accordingly. Thus the output node estimates the probability that input vector belongs to vocalization 1. Inputs with probability greater than 50% were assigned to vocalization 1 otherwise they were assigned to vocalization 0. The network was fit using the ADAM solver [106]. To prevent overfitting, we used five-fold, stratified shuffle split cross-validation with a 50-50 train-to-test split of the data. Briefly, during each repetition the algorithm randomly shuffled the data and sampled 50% of the data from the vocalization pair making sure to maintain the original ratio of data between the two vo-

calizations. We then sub-sampled the data to ensure that there were equally numbers of data points in the training set. These samples were used for training and and the network was tested on the remaining data, which we also subs-sampled. Sub-sampling ensures that chance performance is 50% rather than the 1-the marginal probability of the shorter vocalizations. This procedure was repeated five times and samples were taken with replacement (i.e., the same sample could appear in multiple repetitions). See the documentation for `StratifiedShuffleSplit` in the `model selection` module of `sklearn` for more details. The average score across the repetitions was reported. We used percent of test data classified correctly as our performance metric.

F. Eigenvalue Analysis

The Slow Feature Analysis (SFA) optimization problem is often posed as a generalized eigenvalue problem [17–19, 105]. When posed this way, we can interpret each slow feature as an eigenvector with an eigenvalue proportional to its slowness, so that smaller eigenvalues indicate slower features. We can construct distributions of these eigenvalues by performing SFA on many vocalizations, recording their eigenvalues, and averaging across vocalizations. We performed this analysis for each animal group. We calculated the eigenvalues for each intra-class vocalization pair and then averaged across all vocalization for each animal group. A given vocalization can appears multiple times in the calculation, but this statistically unlikely given the number of vocalizations (see Auditory Stimuli). The calculation generated the three distributions plotted in 4. Since SFA requires that each slow feature be orthogonal to all other features, the distribution is monotonically increasing. That is, each subsequent slow feature will be faster than or as slow as the previous slow features and, equivalently, each eigenvalue will be greater than or equal to the previous eigenvalues. We analyzed the rate of change of slowness across the features by finding the first inflection point on the monotonically increasing eigenvalue distribution of each animal group. This inflection point represents the greatest second derivative along the distribution, after which the eigenvalue, and hence slowness, varies less over the SFA features. We found this inflection point using the KNEEDLE algorithim [47].

G. PFA to SFA comparison

Predictable Feature Analysis (PFA) [48, 49] seeks the components of a signal that permit best prediction of its future by a pre-specified model. We utilize a linear auto-regressive model which is easy to fit and mathematical tractable. We follow the optimization procedure in [49] to find the predictive features, via a generalized eigenvalue problem. The eigenvectors produced by this procedure are the predictive features. While linear SFA generates a set of slow features which can be sorted by slowness, PFA with a linear auto-regressive model generates a set of predictive features which can be sorted by predictability according to the magnitude of each predictable feature's eigenvalue. We can then compare the slowest and most predictive features.

As discussed in the main text, predictability and slowness are related but not equivalent [48, 49]. A signal could change rapidly but be highly predictable (e.g., a binary signal that changes from 0 to 1 every time step is very fast but also very predictable). Thus we asked whether the predictable components of vocalizations were similar to their slow components. We performed this comparison by using Canonical Correlation Analysis (CCA) [50, 51]. CCA can be viewed as an extension of univariate correlation for comparing multidimensional data sets. There are many variations of CCA, but in its most basic form this method finds the left and right singular vectors of the correlation matrix between two data sets that has the greatest singular value. These singular vectors generate the projections of the two data sets that have the greatest correlation. In our case, the calculation is simpler since we consider each PFA feature individually. When one of the data sets is univariate, CCA reduces to linear regression with the higher dimensionality vectors used as regressors to fit the univariate data. Thus our CCA correlations are effectively the R-squared value when SFA is used to predict the PFA signal. For more details on forms of CCA see [53], and for rigorous deviation of basic CCA see [52]

H. Programming languages and code availability

Code was written in Python 3.8.19 and is available via the Balsubramanian lab `Github` repository:`https://github.com/BalasubramanianLab-CNI`. Links to necessary modules are included in the `ReadMe`, or are provided in the repository.

Acknowledgments: VB and RWD were supported in part by NIH grant R01EB026945 *Coincidence and continuity: uncovering the neural basis of auditory object perception.* LW was supported in part by the Velay Fellowship from the University of Pennsylvania. VB thanks the Aspen Center for Physics (NSF grant PHY-2210452) for hospitality as this work was completed.

[1] A. A. Ghazanfar and L. R. Santos, Nat. Rev. Neurosci. **5**, 603 (2004).

[2] A. L. O'Loghlen and S. I. Rothstein, Animal Behaviour

79, 1285 (2010).
[3] S. E. Maguire, M. F. Schmidt, and D. J. White, PLOS one **8**, e63239 (2013).
[4] J. A. Hinojosa, E. M. Moreno, and P. Ferré, Language, Cognition and Neuroscience **35**, 813 (2020).
[5] P. Gabrić, (2021).
[6] F. M. Citron, The Routledge handbook of second language acquisition and neurolinguistics , 148 (2023).
[7] G. Klump, E. Kretzschmar, and E. Curio, Behavioral Ecology and Sociobiology **18**, 317 (1986).
[8] H. Heffner and R. Heffner, Behaviour **24**, 81 (1992).
[9] K. J. Jones and W. L. Hill, Ethology **107**, 717 (2001).
[10] R. M. Seyfarth and D. L. Cheney, Annals of the New York Academy of Sciences **1000**, 32 (2003).
[11] C. T. Miller and Y. E. Cohen, Primate neuroethology , 237 (2010).
[12] S. A. MacLean and D. N. Bonter, PLoS One **8**, e82384 (2013).
[13] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound* (The MIT Press, 1990).
[14] C. J. Darwin, Trends Cogn. Sci. **1**, 327 (1997).
[15] T. D. Griffiths and J. D. Warren, Nat. Rev. Neurosci. **5**, 887 (2004).
[16] R. W. DiTullio and Y. E. Cohen, Sensory Individuals: Unimodal and Multimodal Perspectives , 163 (2023).
[17] L. Wiskott and T. J. Sejnowski, Neural computation **14**, 715 (2002).
[18] H. Sprekeler, C. Michaelis, and L. Wiskott, PLoS Computational Biology **3**, e112 (2007).
[19] H. Sprekeler and L. Wiskott, Available at SSRN 3076122 (2008).
[20] H. Sprekeler, Neural computation **23**, 3287 (2011).
[21] F. Schönfeld and L. Wiskott, Frontiers in computational neuroscience **9**, 51 (2015).
[22] R. W. DiTullio, C. Parthiban, E. Piasini, P. Chaudhari, V. Balasubramanian, and Y. E. Cohen, Frontiers in Computational Neuroscience **17**, 1150300 (2023).
[23] N. C. Singh and F. E. Theunissen, J. Acoust. Soc. Am. **114**, 3394 (2003).
[24] A. Hsu, S. M. Woolley, T. E. Fremouw, and F. E. Theunissen, Journal of Neuroscience **24**, 9201 (2004).
[25] Y. E. Cohen, F. Theunissen, B. E. Russ, and P. Gill, J. Neurophysiol. **97**, 1470 (2007).
[26] S. A. Shamma, M. Elhilali, and C. Micheyl, Trends in neurosciences **34**, 114 (2011).
[27] L. Krishnan, M. Elhilali, and S. Shamma, PLoS Comput. Biol. **10**, e1003985 (2014).
[28] Y. E. Cohen, F. Theunissen, B. E. Russ, and P. Gill, J. Neurophysiol. **97**, 1470 (2007).
[29] K. L. Christison-Lagay, S. Bennur, J. Blackwell, J. H. Lee, T. Schroeder, and Y. E. Cohen, Hear. Res. **312**, 128 (2014).
[30] K. L. Christison-Lagay and Y. E. Cohen, Front. Neurosci. **12**, 601 (2018).
[31] J. H. McDermott, Current Biology **19**, R1024 (2009).
[32] M. Rezaeizadeh and S. Shamma, Cereb Cortex Commun **2**, tgab060 (2021).
[33] S. Woolley, T. E. Fremouw, A. Hsu, and F. E. Theunissen, Nature neuroscience **8**, 1371 (2005).
[34] F. E. Theunissen and J. E. Elie, Nature Reviews Neuroscience **15**, 355 (2014).
[35] J. K. Bizley and Y. E. Cohen, Nat. Rev. Neurosci. **14**, 693 (2013).
[36] G. Bellec, M. Galtier, R. Brette, and P. Yger, Journal of computational neuroscience **40**, 317 (2016).
[37] S. Buus, The Journal of the Acoustical Society of America **78**, 1958 (1985).
[38] R. W. Peters, B. C. Moore, and T. Baer, The Journal of the Acoustical Society of America **103**, 577 (1998).
[39] J. L. Verhey, D. Pressnitzer, and I. M. Winter, Experimental Brain Research **153**, 405 (2003).
[40] S. Amoser and F. Ladich, Journal of Experimental Biology **208**, 3533 (2005).
[41] T. Houtgast and J. M. Festen, International Journal of Audiology **47**, 287 (2008).
[42] R. J. Dooling and S. H. Blumenrath, in *Animal communication and noise* (Springer, 2013) pp. 229–250.
[43] M. Dylla, A. Hrnicek, C. Rice, and R. Ramachandran, Journal of the Association for Research in Otolaryngology **14**, 547 (2013).
[44] F. Rocchi and R. Ramachandran, Journal of Neurophysiology **120**, 2819 (2018).
[45] L. Wasserman, *All of statistics: a concise course in statistical inference* (Springer Science & Business Media, 2013).
[46] G. James, D. Witten, T. Hastie, R. Tibshirani, *et al.*, *An introduction to statistical learning*, Vol. 112 (Springer, 2013).
[47] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, in *2011 31st international conference on distributed computing systems workshops* (IEEE, 2011) pp. 166–171.
[48] F. Creutzig and H. Sprekeler, Neural Computation **20**, 1026 (2008).
[49] S. Richthofer and L. Wiskott, in *2015 IEEE 14th international conference on machine learning and applications (ICMLA)* (IEEE, 2015) pp. 190–196.
[50] H. Hotelling, in *Breakthroughs in statistics: methodology and distribution* (Springer, 1992) pp. 162–190.
[51] H. J. Kim, N. Adluru, B. B. Bendlin, S. C. Johnson, B. C. Vemuri, and V. Singh, in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13* (Springer, 2014) pp. 251–267.
[52] V. Uurtio, J. M. Monteiro, J. Kandola, J. Shawe-Taylor, D. Fernandez-Reyes, and J. Rousu, ACM Computing Surveys (CSUR) **50**, 1 (2017).
[53] X. Zhuang, Z. Yang, and D. Cordes, Human brain mapping **41**, 3807 (2020).
[54] H. L. Anderson, A. Perkes, J. S. Gottfried, H. B. Davies, D. J. White, and M. F. Schmidt, Animal behaviour **172**, 155 (2021).
[55] N. Kowalski, D. A. Depireux, and S. A. Shamma, J. Neurophysiol. **76**, 3524 (1996).
[56] M. A. Escabı and C. E. Schreiner, Journal of Neuroscience **22**, 4114 (2002).
[57] T. Sharpee, N. C. Rust, and W. Bialek, Neural computation **16**, 223 (2004).
[58] J. C. Fentress, The development and integration of behaviour , 77 (1991).
[59] A. Gomez-Marin, J. J. Paton, A. R. Kampff, R. M. Costa, and Z. F. Mainen, Nature neuroscience **17**, 1455 (2014).
[60] D. Grigorescu, EDUCATION AND PSYCHOLOGY CHALLENGES CHALLENGES , 293 (2019).
[61] E. J. Dennis, A. El Hady, A. Michaiel, A. Clemens, D. R. G. Tervo, J. Voigts, and S. R. Datta, Journal of Neuroscience **41**, 911 (2021).
[62] D. Mobbs, T. Wise, N. Suthana, N. Guzmán,

N. Kriegeskorte, and J. Z. Leibo, Neuron **109**, 2224 (2021).
[63] C. Testard, S. Tremblay, and M. Platt, Current opinion in neurobiology **68**, 76 (2021).
[64] C. T. Miller, D. Gire, K. Hoke, A. C. Huk, D. Kelley, D. A. Leopold, M. C. Smear, F. Theunissen, M. Yartsev, and C. M. Niell, Current Biology **32**, R482 (2022).
[65] J. Bordes, L. Miranda, B. Müller-Myhsok, and M. V. Schmidt, Neuroscience & Biobehavioral Reviews , 105243 (2023).
[66] R. d'Isa and R. Gerlai, Frontiers in Behavioral Neuroscience **16**, 1090248 (2023).
[67] S. Tremblay, C. Testard, R. W. DiTullio, J. Inchauspé, and M. Petrides, Nature Neuroscience **26**, 295 (2023).
[68] C. Testard, S. Tremblay, F. Parodi, R. W. DiTullio, A. Acevedo-Ithier, K. L. Gardiner, K. Kording, and M. L. Platt, Nature **628**, 381 (2024).
[69] P. Sterling and S. Laughlin, *Principles of neural design* (MIT press, 2015).
[70] S. Laughlin, Zeitschrift für Naturforschung c **36**, 910 (1981).
[71] J. J. Atick and A. N. Redlich, Neural computation **2**, 308 (1990).
[72] J. H. van Hateren, Journal of Comparative Physiology A **171**, 157 (1992).
[73] B. T. Vincent and R. J. Baddeley, Vision Research **43**, 1285 (2003).
[74] X. Kuang, M. Poletti, J. D. Victor, and M. Rucci, Current Biology **22**, 510 (2012).
[75] X. Pitkow and M. Meister, Nature neuroscience **15**, 628 (2012).
[76] K. D. Simmons, J. S. Prentice, G. Tkačik, J. Homann, H. K. Yee, S. E. Palmer, P. C. Nelson, and V. Balasubramanian, PLoS Computational Biology **9**, e1003344 (2013).
[77] D. Gupta, W. Mlynarski, A. Sumser, O. Symonova, J. Svaton, and M. Joesch, bioRxiv 10.1101/2022.01.11.475815 (2022).
[78] A. L. Fairhall, G. D. Lewen, W. Bialek, and R. R. de Ruyter van Steveninck, Nature **412**, 787 (2001).
[79] C. P. Ratliff, B. G. Borghuis, Y.-H. Kao, P. Sterling, and V. Balasubramanian, Proceedings of the National Academy of Sciences **107**, 17368 (2010).
[80] J. Gjorgjieva, H. Sompolinsky, and M. Meister, Journal of Neuroscience **34**, 12127 (2014).
[81] B. G. Borghuis, C. P. Ratliff, R. G. Smith, P. Sterling, and V. Balasubramanian, Journal of Neuroscience **28**, 3178 (2008).
[82] Y. S. Liu, C. F. Stevens, and T. O. Sharpee, Proceedings of the National Academy of Sciences **106**, 16499 (2009).
[83] P. Garrigan, C. P. Ratliff, J. M. Klein, P. Sterling, D. H. Brainard, and V. Balasubramanian, PLoS Computational Biology **6**, e1000677 (2010).
[84] S. E. Palmer, O. Marre, M. J. Berry, and W. Bialek, Proceedings of the National Academy of Sciences **112**, 6908 (2015).
[85] J. M. Salisbury and S. E. Palmer, Journal of Statistical Physics **162**, 1309 (2016).
[86] M. Weckström and S. B. Laughlin, Trends in Neurosciences **18**, 17 (1995).
[87] T. Teşileanu, S. Cocco, R. Monasson, and V. Balasubramanian, Elife **8**, e39279 (2019).
[88] V. Singh, M. Tchernookov, and V. Balasubramanian, Physical Review E **104**, 024415 (2021).
[89] K. Krishnamurthy, A. M. Hermundstad, T. Mora, A. M. Walczak, and V. Balasubramanian, Frontiers in Computational Neuroscience **16**, 917786 (2022).
[90] B. A. Olshausen and D. J. Field, Nature **381**, 607 (1996).
[91] G. Tkačik, J. S. Prentice, J. D. Victor, and V. Balasubramanian, Proceedings of the National Academy of Sciences **107**, 18149 (2010).
[92] A. M. Hermundstad, J. J. Briguglio, M. M. Conte, J. D. Victor, V. Balasubramanian, and G. Tkačik, Elife **3**, e03722 (2014).
[93] T. Tesileanu, M. M. Conte, J. J. Briguglio, A. M. Hermundstad, J. D. Victor, and V. Balasubramanian, ELife **9**, e54347 (2020).
[94] X.-X. Wei, J. Prentice, and V. Balasubramanian, Elife **4**, e08362 (2015).
[95] O. Schwartz and E. P. Simoncelli, Nature neuroscience **4**, 819 (2001).
[96] M. S. Lewicki, Nature neuroscience **5**, 356 (2002).
[97] E. C. Smith and M. S. Lewicki, Nature **439**, 978 (2006).
[98] N. L. Carlson, V. L. Ming, and M. R. DeWeese, PLoS Computational Biology **8**, e1002594 (2012).
[99] S. Becker, J. Vielhaben, M. Ackermann, K.-R. Müller, S. Lapuschkin, and W. Samek, Journal of the Franklin Institute **361**, 418 (2024).
[100] M. Fukushima, A. M. Doyle, M. P. Mullarkey, M. Mishkin, and B. B. Averbeck, Royal Society open science **2**, 150432 (2015).
[101] Z. Zhao, S.-h. Zhang, Z.-y. Xu, K. Bellisario, N.-h. Dai, H. Omrani, and B. C. Pijanowski, Ecological Informatics **39**, 99 (2017).
[102] X. Zhang, M. G. Heinz, I. C. Bruce, and L. H. Carney, J. Acoust. Soc. Am. **109**, 648 (2001).
[103] S. Tabibi, A. Kegel, W. K. Lai, and N. Dillier, Journal of neuroscience methods **277**, 63 (2017).
[104] R. F. Lyon, A. G. Katsiamis, and E. M. Drakakis, in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems* (IEEE, 2010) pp. 3809–3812.
[105] R. Turner and M. Sahani, Neural computation **19**, 1022 (2007).
[106] D. P. Kingma and J. Ba, arXiv preprint arXiv:1412.6980 (2014).