Understanding auditory spectro-temporal receptive fields and their changes with input statistics by efficient coding principles

Lingyun Zhao[1], Li Zhaoping[2]*

1 Department of Biomedical Engineering,

School of Medicine, Tsinghua Univ, Beijing 100084, P.R.China

2. Department of Computer Science, University College London, UK

* Correspondence should be addressed to z.li@ucl.ac.uk

Published in PLoS Computational Biology, 7(8):e1002123. 2011

Abstract

Spectro-temporal receptive fields (STRFs) have been widely used as linear approximations to the signal transform from sound spectrograms to neural responses along the auditory pathway. Their dependence on statistical attributes of the stimuli, such as sound intensity, is usually explained by nonlinear mechanisms and models. Here, we apply an efficient coding principle which has been successfully used to understand receptive fields in early stages of visual processing, in order to provide a computational understanding of the STRFs. According to this principle, STRFs result from an optimal tradeoff between maximizing the sensory information the brain receives, and minimizing the cost of the neural activities required to represent and transmit this information. Both terms depend on the statistical properties of the sensory inputs and the noise that corrupts them. The STRFs should therefore depend on the input power spectrum and the signal-to-noise ratio, which is assumed to increase with input intensity. We analytically derive the optimal STRFs when signal and noise are approximated as Gaussians. Under the constraint that they should be spectro-temporally local, the STRFs are predicted to adapt from being band-pass to low-pass filters as the input intensity reduces, or the input correlation becomes longer range in sound frequency or time. These predictions qualitatively match physiological observations. Our prediction as to how the STRFs should be determined by the input power spectrum could readily be tested, since this spectrum depends on the stimulus ensemble. The potentials and limitations of the efficient coding principle are discussed.

Author Summary

Spectro-temporal receptive fields (STRFs) have been widely used as linear approximations of the signal transform from sound spectrograms to neural responses along the auditory pathway. Their dependence on the ensemble of input stimuli has usually been examined mechanistically as a possibly complex nonlinear process. We propose that the STRFs and their dependence on the input ensemble can be understood by

an efficient coding principle, according to which the responses of the encoding neurons report the maximum amount of information about the sensory input, subject to limits on the neural cost in representing and transmitting information. This proposal is inspired by the success of the same principle in accounting for receptive fields in the early stages of the visual pathway and their adaptation to input statistics. The principle can account for the STRFs that have been observed, and the way they change with sound intensity. Further, it predicts how the STRFs should change with input correlations, an issue that has not been extensively investigated. In sum, our study provides a computational understanding of the neural transformations of auditory inputs, and makes testable predictions for future experiments.

Introduction

In response to acoustic input signals, neurons in the auditory pathway are typically selective to sound frequency f and have particular response latencies. At least ignoring cases with $f < 4$kHz, in which neuronal responses often phase lock to the sound waves, a spectro-temporal receptive field (STRF) is often used to describe the tuning properties of a neuron [1, 2, 3, 4]. This is a two-dimensional function $STRF(f,t)$ that reports the sensitivity of the neuron at response latency t to acoustic inputs of frequency f for a given stimulus ensemble (i.e., given input statistics). More specifically, in a stimulus ensemble, the power $S(f,t)$ of the acoustic input at frequency f at time t fluctuates around an average level denoted by $\bar{S}(f)$. If we let $O(t)$ denote the neuron's response at time t (typically its spike rate), then $STRF(f,t)$ best approximates the linear relationship between $O(t)$ and $S(f,t)$ in this stimulus ensemble as

$$O(t) = \iint STRF(f,\tau)S(f,t-\tau)d\tau df + \text{spontaneous activity} \tag{1}$$

Note that in this paper, we refer to $S(f,t)$ as the input spectrogram, although some authors also include the average input power $\bar{S}(f)$. Though $S(f,t)$ is not a full description of acoustic input, since it ignores features such as the phase of the oscillation in the sound wave, it is the only relevant aspect of the auditory input as far as the STRF is concerned. Note that if we use $O(t)$ to denote the deviation of the neural response from its spontaneous activity level, then both $O(t)$ and $S(f,t)$ have zero mean. We will use this simplification throughout the paper. In studies in which the temporal dimension is omitted, the STRF is called the spectral receptive field (SRF).

Figure 1 cartoons a typical STRF. This has excitatory and inhibitory regions, reflecting its preferred frequency and response latency. For example, if $STRF(f,t)$ peaks at frequency $f = \hat{f}$ and time $t = \hat{t}$, then this neuron prefers frequency $\hat{f}$ and should respond to an input impulse $S(f,t) = \delta\left(f - \hat{f}\right)\delta(t)$ of this frequency with latency $\hat{t}$. We will also refer to $STRF(f,t)$ as the receptive field, the filter kernel, or the transfer function from input to neural responses, as these all convey the same or similar meanings. A neuron's STRF is typically estimated using reverse correlation methods [5, 4].

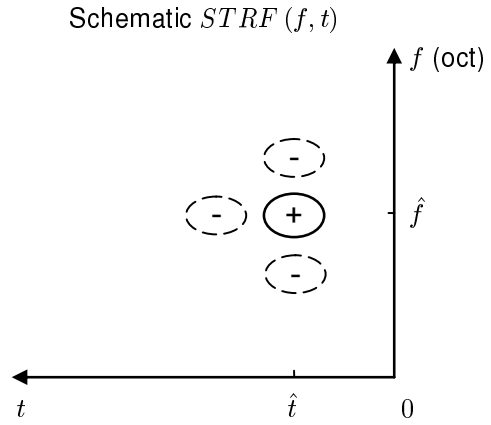

Figure 1: A schematic example of a typical spectro-temporal receptive field, plotted with a reversed abscissa. This STRF has one excitatory and three inhibitory regions, prefers frequency $\hat{f}$, and evokes response at a typical latency $\hat{t}$. Since the response at time $t = 0$ is $O(t = 0) = \iint STRF\,(f,\tau)\,S\,(f,-\tau)\,d\tau df$, an input stimulus $S\,(f,t) = STRF\,(f,-t)$ exactly as depicted in this plot is most likely to elicit a large response $O\,(t = 0)$ at time $t = 0$, or indeed a spike.

However, there are extensive nonlinearities in the signal transformation along the auditory pathway. Indeed, the STRF formulation of neural responses, though linear in spectral power, is already a second-order nonlinear function of the auditory sound wave. There are two kinds of nonlinearities when inputs are represented as spectrograms. The simplier one is a static nonlinearity $f_{nonlinear}(O(t))$, which when applied to the linear approximation $O(t)$ of equation (1) enables better predictions of the neural responses[6, 7]. This static nonlinearity however does not alter the spectro-temporal selectivity of the neuron seen in the linear STRF. This paper is interested in the more complex nonlinearity that the STRFs are dependent on the stimulus ensemble used to estimate them[1, 5, 8, 9]. For example, the STRFs are wider when the input intensity is weaker[10], or when the stimuli are animal vocalizations rather than noise[11]. The STRF (or SRF) also becomes more band-pass when sound intensity increases. The dependence of the STRFs on the stimulus ensemble holds, for example, for type IV neurons in the cochlear nucleus of cats[12, 13], the inferior colliculus (IC) of the frog[8] and the gerbil [7], and field L region of the songbird (which is analogous to mammalian auditory cortex) [14]. (The dependence on sound intensity also holds for the linear relationship between the auditory nerve responses and input sound waves[5]). Nonlinearities in the auditory system become progressively stronger further from the periphery.

Despite the nonlinearities, the concept of the STRF is still widely used, not only because it provides a meaningful description of the spectro-temporal selectivity of the neurons in a given stimulus ensemble, but also because it can predict neural responses to novel stimuli reasonably well, as long as the stimuli are drawn from the same stimulus ensemble as that used to estimate the STRF in the first place. Reasonable predictions from the STRFs have been obtained for the responses of auditory nerves(see [15]) and auditory midbrain neurons[6, 7, 16] (also see[2]). They have also been obtained for responses of the auditory cortical

neurons when the stimulus ensemble is composed of biologically more meaningful static or dynamic ripples (broadband sound with sinusoidally modulated spectral envelopes and their linear combinations [17, 18, 19]). If the linear neural filter is augmented to include the filtering performed by the head and ears, it is also possible to predict the preferred locations of sound sources of auditory cortical neurons based on the linear neural filter for input spectrograms[20]. Meanwhile, linear STRF models fail to capture many complex phenomena, particularly in the auditory cortex, and nonlinearities are not limited to being just static or monotonic. It has been suggested that some auditory cortical neurons process auditory objects in a highly non-linear manner, by selectively responding to a weak object component while ignoring loud components that occupy the same region in frequency space in auditory mixtures of these object components [21], and some prefer low over high spectral contrast sounds [22]. Strong nonlinearities in the auditory processes have long since motivated nonlinear models of auditory responses (e.g.,[5, 12, 23]).

This paper aims to understand from a computational, rather than a mechanistic, perspective why the auditory encoding transform should depend on the stimulus ensemble in the ways observed. More specifically, the paper focuses on cases in which STRFs can reasonably capture neural responses, and aims to identify and understand the computational goal of the STRFs for a given stimulus ensemble – finding a metric according to which the STRFs are optimal for the ensemble. This would provide a rationale for how the physiologically measured STRFs should depend on or adapt to the stimulus ensemble. This paper does not address what linear or nonlinear mechanisms could build the optimal STRFs, or whether or how nonlinear auditory processes enable the adaptation of the STRFs to the stimulus ensemble. Existing computational models of auditory neurons, including ones with the notion that cochlear hair cells perform independent component analysis to provide an efficient code for inputs using spikes in the auditory nerves[24, 25], cannot explain the observed dependence of the STRFs on the stimulus ensemble (see Discussion for more details).

Restricting attention to the temporal properties of STRF, Lesica and Grothe[26] observed that the temporal filter in STRF adapted to the level of ambient noise in the input environment. In particular, the temporal receptive field in the STRF changed from being bandpass to being low pass with the increase of ambient noise. They argued using a simple model that such adaptation in the STRF enables more efficient coding of the input information.

This study applies the principles of efficient coding to understand the auditory STRF and its variations with sound intensities and other input characteristics. It generalizes the work of Lesica and Grothe[26] to understand the temporal and spectral filtering characteristics of STRF adaptation to changes in noise, signal and correlations in input statistics. Explicitly, the principle of efficient coding states that the neural receptive fields should enable the neural responses to transmit as much sensory information as possible to the central nervous system, subject to the limitation in neural cost in representing and transmitting information. This principle has been proposed[27] and successfully applied to the visual system to understand the receptive fields in the early visual pathway[28, 29, 30, 31, 32, 33] (see review[34]). We will borrow heavily techniques and intuitions from vision to derive and explain the results in this paper.

To make initial progress, it is necessary to start with some simplifying assumptions. First, we assume that

the statistical characteristics of the stimulus ensemble do not change more rapidly than the speed at which the sensory encoding adapts, so that the stimulus ensemble can be approximated as being stationary as far as optimal encoding is concerned. Knowing when this assumption does not hold tells us when the encoding is not optimal, e.g., when one sees poorly for a brief moment before the visual encoding adapts to a sudden change from a dark room to a bright garden. Second, for mathematical convenience, we assume that the linear STRF model as in equation (1) can approximate adapted auditory neural responses reasonably well. As we know from above, this assumption often does not hold, particularly for auditory cortical neurons. This paper leaves the extension of the optimal encoding to nonlinear cases for future studies. Third, to derive a closed-form, analytical, solution to the optimal STRF, we assume that the input statistics in the stimulus ensemble can be approximated as being Gaussian, with higher order correlations in the input contributing only negligibly to the inefficiency of the representation in the original sensory inputs. Although it is known that the natural auditory inputs are far from Gaussian[35], as for the case of vision, the discrepancy may have only a limited impact on the input inefficiency, as measured by the amount of information redundancy in the original sensory input [36, 37, 38].

To understand how sensory inputs should be recoded to increase coding efficiency, we start with visual encoding to draw insights and made analogies with auditory encoding. In vision, large amounts of raw data about the visual world are transduced by photoreceptors. However, the optic nerve, which transmits the input data to the visual cortex via thalamus, can only accommodate a dramatically smaller data rate. It has thus been proposed that early visual processes use an efficient coding strategy to encode as much information as possible given the limited bandwidth [27, 34], in other words, to recode the data such that the redundancy in the data is reduced and consequently the data can be transmitted by the limited bandwidth. Compression (while preserving most information) is possible since images are very redundant [39, 40, 41, 42], e.g., with strong correlations between visual inputs at nearby points in time and space. Removing such correlations can cut down the data rate substantially [34].

One way to remove the correlations is to transform the raw input S into a different representation O in neural responses that would then have a much smaller data rate than S, yet preserving essential input information. This transform is often approximated by the visual receptive field, analogous to the auditory STRFs. For instance, the (spatial) center-surround receptive fields of the retinal ganglion cells help remove spatial redundancy [30, 31, 43]. They do this by making the ganglion cells preferentially respond to spatial contrast in the input, and so eliminating responses to visual locations whose input is redundant with that of their neighbors. Consequently, the responses of retinal ganglion cells are much less correlated than those of the photoreceptors, making their representation much more efficient. One facet of this efficient encoding hypothesis is that the optimal receptive field transform should depend on the statistical properties, such as the correlation structure and intensity, of the input. This dependence has been used to explain adaptation, to changes in input statistics, of visual receptive field characteristics, such as the sizes of center-surround regions and the color tuning of retinal neurons, or the ocular dominance properties of striate cortical neurons [32, 34, 44, 45, 46, 47]. In the auditory system, information redundancy is also reduced along the auditory pathway[48]. Although this redundancy reduction was only investigated in the neural responses to sensory

inputs rather than in the coding (STRF) transform leading to the neural responses, it suggested that coding efficiency is one of the goals of early auditory processes.

More formally, the efficient coding scheme is depicted in Figure 2A. The input contains sensory signal S and noise N (e.g., input sampling noise). The net input $S + N$ is encoded by a linear transfer function K into output.

$$O = K(S + N) + N_o \tag{2}$$

which also contains additional noise N_o introduced in the encoding process. When the input has multiple channels, e.g., many different photoreceptors or hair cells, $S = (S_1, S_2, ..., S_j, ...)$ is a vector with many components, as indeed is N. Output O is a vector representing the neural population responses from many neurons. For output neuron i, we have $O_i = \sum_j K_{ij}(S_j + N_j) + N_{o,i}$. Therefore K is a matrix, and its i^{th} row $(K_{i1}, K_{i2}, ..., K_{ij}, ...)$ models the receptive field for output neuron i as the array of effective weights from input receptors j to output neuron i. In the particular example when input neurons are photoreceptors and output neurons are retinal ganglion cells, K_{ij} is the effective connection from photoreceptor j to ganglion cell i (implemented via the interneurons in the cell layers of the retina), and collectively, $(K_{i1}, K_{i2}, ..., K_{ij}, ...)$ describe the linear receptive field of this ganglion cell. We consider the problem of finding an optimal K that maximizes the information extracted by O about S, i.e., the mutual information $I(O; S)$[49] between O and S subject to a given cost of the neural encoding, which depends on the responses in a way we will describe shortly.

Therefore, the optimal K should minimize the objective function:

$$E\left(K\right) = \text{neural cost} - \lambda \times I\left(O; S\right) \tag{3}$$

where λ is a parameter whose value specifies a particular balance between the needs to minimize costs and to maximize extracted information. Neural costs can arise from various sources, such as the metabolic energy cost for generating neural activities or spikes[50] and the cost of thicker axons to transmit higher rates of neural firing. We follow a formulation that has been productive in vision[31, 34], and model the neural cost as

$$\text{neural cost} = \sum_i \left\langle O_i^2 \right\rangle,$$

where $\langle ... \rangle$ indicates the average over the stimulus ensemble. This gives

$$E(K) = \sum_i \left\langle O_i^2 \right\rangle - \lambda \times I(O; S) \tag{4}$$

It has been shown [29, 33, 51, 34] that the K that provides the most efficient coding according to $E(K)$ has the following properties. At high signal-to-noise ratio (SNR), K is such that O extracts the difference between correlated channels, and thus avoids transmitting redundant information. Hence, for example, in

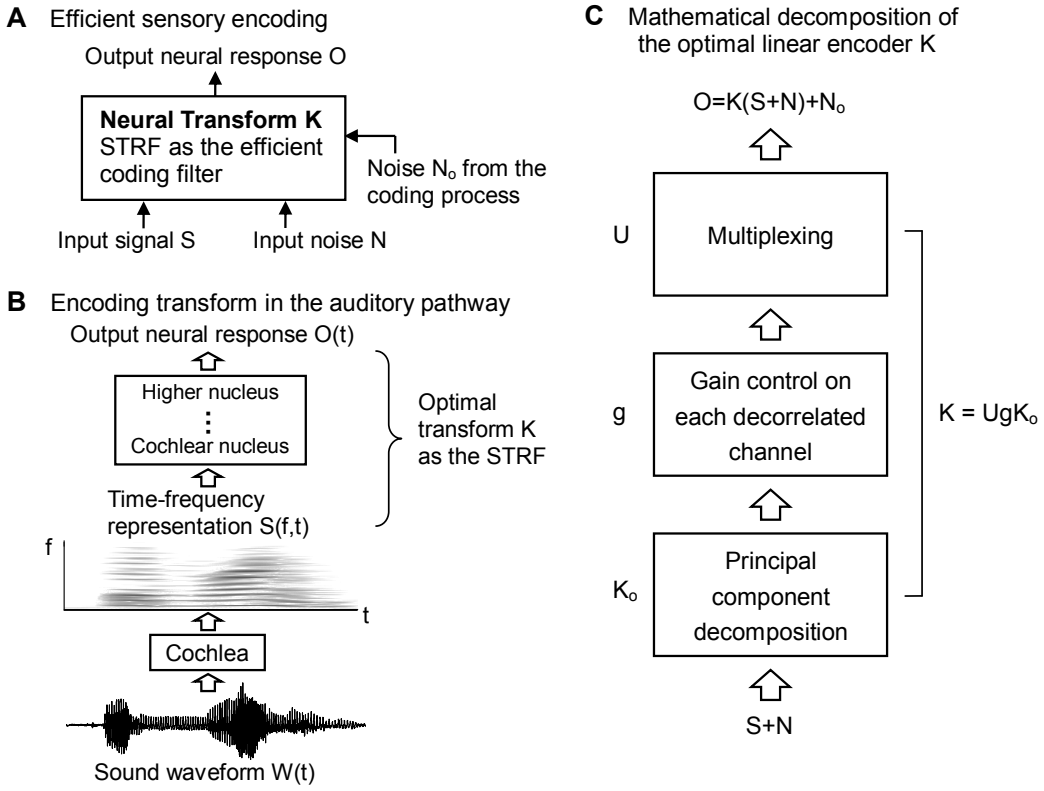

Figure 2: Formulation and components of efficient coding. (A) A schematic plot of the efficient encoding transform. (B) Signal transformation in the auditory system. The cochlea turns the time-varying waveform $W(t)$ into a time-frequency representation $S(f,t)$, as the population activities of the auditory nerves, which is the input to the efficient encoding system. Signal and noise pass through a series of brain nuclei such as cochlear nucleus, superior olive, inferior colliculus, etc. The current work proposes that the effective transform STRF of the spectrogram that is collectively realized by these nuclei is, in its linear form, the optimal filter K implied by the efficient coding principle. The output $O(t)$ is the activity of neurons in a higher nucleus. (C) Three steps of signal flow within the linear encoding step K or STRF in (A) and (B). Note that these three steps are merely abstract algorithmic steps, rather than neural implementation processes for the effective transform K or STRF.

photopic conditions, retinal ganglion cells have center-surround spatial receptive fields which extract the spatial contrast of the input. By contrast, at low SNR, K is a smoothing filter that averages out input noise instead of reducing redundancy. This avoids spending neural cost on transmitting noise. Hence, for example, in scotopic conditions, when SNR can be considered as being low, the receptive fields of retinal ganglion cells expand the sizes of their center regions and weaken their suppressive surrounds [52]. We will apply this framework to the auditory encoding to understand STRFs and their adaptation to stimulus ensembles.

Methods

Auditory encoding system and its comparison to vision

To apply the efficient coding principle to auditory STRFs, we borrow insights from vision by making an analogy between (aspects of) the auditory and visual systems. For simplicity, we start by ignoring input noise. While sound signals are typically air vibrations over time, at the input sampling stage, they are sampled as $S_{f,t}$ from a continuous time-frequency representation $S(f,t)$, namely the response at time t of a hair cell tuned to sound vibration frequency f. This is analogous to visual input sampling, in which the response of a photoreceptor at location i samples the light signal in the form of electromagnetic vibrations. Auditory hair cells are tonotopically arranged in the cochlea, so that neighboring hair cells are tuned to nearby sound frequencies. Therefore, at any instant t , the response pattern $(S_{f_1,t}, S_{f_2,t}, ...S_{f_i,t}, ...)$ as a function of hair cell's location i over the cochlea is an auditory "image" of the pattern of powers across sound frequencies, analogous to a retinal image. (In our formulation, we focus on sampling the intensity or power in $S_{f,t}$, and ignore the phase of the sound wave at frequency f. This is because (1) auditory nerve responses do not encode the phase except for low frequency inputs via phase locking, and (2), as mentioned, our goal is to understand the STRFs which do not concern the phase information.) While a retinal image is two dimensional in space (and one additional dimension in time), the auditory "image" at any instant t is one dimensional in sound frequency f. One may use time t as the second dimension such that $S_{f,t}$ for all f and t collectively can be seen as a single discrete sample of the two-dimensional auditory "image". When input noise N is included, input S becomes $S + N$.

As for vision, we explore whether the auditory STRFs can be partly understood by the goal of efficiently coding auditory information. The sensory input is sampled as $S+N$, the responses of the cochlear hair cells. This input is encoded by the STRFs to give rise to outputs O as the neural activities of a higher nucleus, such as the inferior colliculus (IC) or the auditory cortex (Figure 2B). The STRF is then analogous to a spatial receptive field, such as that of the retinal ganglion cells. Thus the STRF should be determined by the statistics of the auditory inputs, and in particular, the correlation $R^S_{ij} = \left\langle S_{(f,t)_i} S_{(f,t)_j} \right\rangle$ between different inputs $S_{(f,t)_i}$ and $S_{(f,t)_j}$, where $(f,t)_i$ labels a particular spectro-temporal combination of a frequency value f and time t. Note that for $i \neq j$, the frequency f or t, but not both, in the two indices $(f,t)_i$ and $(f,t)_j$ may be equal. (Here, for simplicity we assume, or pre-process the signal, such that all inputs have zero mean,

i.e., $\langle S_{(f,t)_i} \rangle = 0$, just like the input signal fluctuation $S(f,t)$ around the ensemble average in the definition of the STRF in equation (1)). As in vision, natural auditory inputs express substantial correlations between inputs of neighboring frequencies and at neighboring temporal instances. When the input SNR is sufficiently high, an optimal STRF should reduce these correlations to achieve efficient transmission. Such an STRF will have neighboring excitatory and inhibitory regions in the frequency-latency domain, making the neuron be tuned to spectro-temporal contrast and be insensitive to the spectro-temporal redundancy.

Auditory STRF filter as an efficient coding transform

The general formulation and derivation of the efficient coding transform K (or STRF) can be found in its application to vision [34]. Here we outline these results and illustrate their consequences for auditory coding. Let S be the input with p input channels:

$$S = (S_1, S_2, ..., S_p)^T \tag{5}$$

(superscript T denotes vector or matrix transpose). These p input channels may correspond to p auditory nerves if we omit the temporal dimension, p time instances if we focus on a single frequency channel, or they may correspond to p spectro-temporal labels $(f,t)_i$ for $i = 1, 2, ..., p$. Let the input correlation be described by correlation matrix R^S with elements $R^S_{ij} = \langle S_i S_j \rangle$. The optimal transform K that minimizes $E(K)$ in equation (4) can be decomposed in three steps (Figure 2C): (1) a principal component transform to de-correlate the inputs, (2) gain control of each principal component, (3) an ortho-normal or unitary transform on the array of the gain-controlled components to arrive at various output channels. We now elaborate and elucidate these three steps.

The first step is a coordinate rotation, or ortho-normal transform, $S \rightarrow K_o S$, by an ortho-normal matrix K_o that de-correlates the input channels such that each of the channels in the transformed signal $K_o S$ contains a principal component of the original signal. We denote these principal components as $\mathcal{S}_k = \sum_j (K_o)_{kj} S_j$, with sub-index k (instead of i, j) as the indices of the de-correlated channels (later, we also use ω to denote the de-correlated channels in the temporal domain, or (Ω, ω) in spectro-temporal domain). Since the correlation between $\mathcal{S}_k$ and $\mathcal{S}_{k'}$ is $\langle \mathcal{S}_k \mathcal{S}_{k'} \rangle = (K_o R^S K_o^T)_{kk'}$, decorrelation between principal components implies that $K_o R^S K_o^T$ is a diagonal matrix, with $(K_o R^S K_o^T)_{kk'} = \langle \mathcal{S}_k^2 \rangle \delta_{kk'}$, where $\langle \mathcal{S}_k^2 \rangle$ is the k^{th} eigenvalue of matrix R^S and also the average signal power of the k^{th} principal component $\mathcal{S}_k$. As we will see later, when the input correlation $\langle S_{f,t} S_{f',t'} \rangle$ depends mainly on the differences $(f - f', t - t')$ in frequency and time, it turns out that $\mathcal{S}_k$ (with the index k denoting the spectro-temporal modulation frequency (Ω, ω)) is the amplitude of a dynamic or moving ripple that some experiments use to estimate the STRFs of cortical and midbrain neurons [17, 18, 19, 16, 2].

The second step is gain control g_k on each component $\mathcal{S}_k$, giving output $g_k \mathcal{S}_k$. Including noise $\mathcal{N}_k$, which is the original input noise N projected to the k^{th} channel by the transform K_o, and the encoding noise $\mathcal{N}_{o,k}$ (in the decorrelated k space), the total output becomes $\mathcal{O}_k = g_k(\mathcal{S}_k + \mathcal{N}_k) + \mathcal{N}_{o,k}$. It can be shown (see

[34]) that the gain g_k that minimizes $E(K)$ in equation (4) is determined by the input signal-to-noise ratio $\langle \mathcal{S}_k^2 \rangle / \langle \mathcal{N}^2 \rangle$ to satisfy

$$g_k^2 \propto \text{Max}\left\{ \left[\frac{1}{2(1 + \langle \mathcal{N}^2 \rangle / \langle \mathcal{S}_k^2 \rangle)} \left(1 + \sqrt{1 + \frac{2\lambda}{(\ln 2)\, \langle \mathcal{N}_o^2 \rangle} \frac{\langle \mathcal{N}^2 \rangle}{\langle \mathcal{S}_k^2 \rangle}} \right) - 1 \right], 0 \right\} \tag{6}$$

where $\langle \mathcal{N}^2 \rangle$ is the variance of $\mathcal{N}_k$, and also of the input noise N (assumed to be independent, identically distributed and Gaussian in each channel) , and $\langle \mathcal{N}_o^2 \rangle$ is the variance of the encoding noise $\mathcal{N}_{o,k}$ in each channel k (and of the encoding noise $N_{o,i}$ in each i since different encoding noise channels are also assumed to be independently and identically distributed).

Note that the total noise at output neuron i is output noise$_i = \sum_j K_{ij} N_j + N_{o,i}$. One effect of the encoding transform K is that noise corrupting different output neurons can be correlated, even when the original input noise is independent. The additional encoding noise $N_{o,i}$ could also be correlated in different output neurons, since it could also reflect a common origin in intermediate stages of the encoding processes. Our assumption of independence between $N_{o,i}$ and $N_{o,j}$ for $i \neq j$ is thus a simplification for mathematical convenience.

Since all the variables are assumed to be Gaussian, each output $\mathcal{O}_k$ extracts the following amount of information

$$I(\mathcal{O}_k; \mathcal{S}_k) = \frac{1}{2} \log \left(1 + \frac{g_k^2 \langle \mathcal{S}_k^2 \rangle}{g_k^2 \langle \mathcal{N}^2 \rangle + \langle \mathcal{N}_o^2 \rangle} \right)$$

about the input S and has an output power $\langle \mathcal{O}_k^2 \rangle = g_k^2 (\langle \mathcal{S}_k^2 \rangle + \langle \mathcal{N}^2 \rangle) + \langle \mathcal{N}_o^2 \rangle$. Since different output channels $\mathcal{O}_k$ from different k are decorrelated from each other, the quantity E in equation (4) is

$$E = \sum_k \langle \mathcal{O}_k^2 \rangle - \lambda \sum_k I(\mathcal{O}_k; \mathcal{S}_k) \tag{7}$$

One can then verify that g_k^2 in equation (6) indeed minimizes this E since $dE/dg_k^2 = 0$ at that value. Note that if $\mathcal{S}_k$ is the amplitude of a moving ripple indexed by k, g_k will be the sensitivity of the neuron to the moving ripple.

We can write these two steps as the product gK_o, where K_o is the principal component transform, and g performs the gain control. g is a diagonal matrix with diagonal elements g_k. The net output is then $O = gK_o(S + N) + N_o$. Consider imposing on this transform an orthonormal or unitary transform U (with $UU^T = 1$), the third step in building the efficient coding filter K, giving $K = UgK_o$. It follows [34] from the properties of unitary matrices that neither the first term nor the second term in E in equation (4) will be affected by U (at least when signal and noise are Gaussian and when the components of N_o are independent and identically distributed).

Each row vector of the matrix K determines the receptive field of a particular output channel or neuron. Without U , $K = gK_o$ would specify receptive fields that would be gain controlled eigenvectors or principal components of the input correlation matrix. For example, they would look like ripples covering the entire

spectro-temporal range. An appropriate choice of non-trivial U will alter the receptive field shape dramatically, giving rise to receptive field properties found in real neurons such as a finite span in input channel space. For example, if we consider only the input frequency channels f for auditory inputs and omit the time dimension, we may prefer that the STRF for an output neuron to be selective to only a finite band of input frequencies such that the neural responses O resemble periphery inputs S while maintaining coding efficiency. It can be shown[34, 45] that this can be achieved by choosing $U = K_o^{-1}$, such that $K = K_o^{-1} g K_o$. We will use this choice, $U = K_o^{-1}$, in building our STRF in frequency domain. However, insensitive to the exact form of U, the critical feature of the STRF comes from the gain g_k specified in the second step of the encoding model (as long as one does not impose additional computational goals that may restrict the final STRFs, see Discussion). We will show later that g_k often corresponds to the modulation transfer functions (MTFs, also called ripple transfer function, RTF,in different literatures) of the STRFs.

We now apply this general framework to the case of auditory encoding. Sound spectrogram $S(f,t)$ is derived from the sound waveform $W(t)$ as follows. The first step is to perform a temporally-windowed Fourier transform of $W(t)$ to obtain the sound spectrum $\hat{W}(\hat{f},t) \propto \int W(\tau)T(t-\tau)e^{-i\cdot 2\pi \hat{f}\tau}d\tau$ as a function of time, where $T(t)$ is a temporal window function (e.g., $T(t) = 1$ for $t \in [0,t_0]$,$T(t) = 0$ otherwise). Since the cochlea performs approximately a log scale frequency analysis, we first let $f = \log\left(\hat{f}\right)$ to obtain $\hat{W}(f,t)$ (although the more accurate form would be $f = 21.4\log_{10}\left(4.37\hat{f}+1\right)$ [53]). Then the input power in f is $\hat{S}(f,t) = \left|\hat{W}(f,t)\right|^2$. One may employ a further logarithmic transform $S(f,t) = \log\hat{S}(f,t)$ to characterize the cochlear response better (through capturing the compressive input/output transform realized by processes in the basilar membrane and hair cells)[54, 55]. However, this further logarithmic transform is not essential for our formulation, and, as pointed out previously[56], it does not significantly affect the qualitative characteristics of the empirical STRFs. If one omits this logarithmic transform, then $S(f,t) = \hat{S}(f,t)$. We then subtract the mean $\langle S(f,t)\rangle$ from $S(f,t)$, and, for simplicity, denote the resulting zero mean signal still by $S(f,t)$, as in the definition of STRF. We next consider discrete samples $S_{f,t}$ of the continuous $S(f,t)$. This leads to the input correlation matrix $R^S_{ij} = \langle S_{(f,t)_i} S_{(f,t)_j}\rangle$.

Finally, we follow the three encoding steps above to obtain the optimal encoding transform as $STRF = K$. In the sub-section "The spectral filter SRF", we discuss the simple case in which the temporal dimension t is omitted. Then, the input vector (equation (5)) is $S = (S_{f_1}, S_{f_2}, ...)^T$, and the input correlation matrix is $R^S_{ij} = \langle S_{f_i} S_{f_j}\rangle$. The efficient encoding procedure specifies the optimal spectral receptive field (SRF) K_{ij} for neuron i, with $O_i = \sum_j K_{ij} S_{f_j}$ + noise. When the temporal dimension is included $S = (S_{(f,t)_1}, S_{(f,t)_2}, ...)^T$, $R^S_{ij} = \langle S_{(f,t)_i} S_{(f,t)_j}\rangle$, and efficient coding specifies the optimal STRF as input weights or selectivity associated with the spectrogram $\{S_{(f,t)_i}\}$.

It is apparent that the optimal SRF and STRF depend on input statistics via the input correlation R^S and the input SNR (through the steps 1 and 2 in the encoding scheme). Therefore, when the stimulus ensemble changes, altering the input correlations and signal intensity, the form of the encoding receptive field should adapt in order to maintain encoding optimality. We propose that it is this that explains the input ensemble

dependence of the STRFs.

A special class of input statistics has translation invariant correlations, i.e., with $R^S_{ij} = \left\langle S_{(f,t)_i} S_{(f,t)_j} \right\rangle$ depending only on the differences $f_i - f_j$ (quantified in octaves) and $t_i - t_j$. This is a reasonable approximation of the input correlations in natural auditory scenes under two conditions. The first is that a local frequency range is considered that is not much larger than the range of the frequencies to which a neuron is sensitive, i.e., in the perspective of a neuron, the dependence of $\left\langle S_{(f,t)_i} S_{(f,t)_j} \right\rangle$ on the frequency is mainly through $f_i - f_j$. This is analogous to approximating spatial correlation of visual inputs as translation invariant to understand the retinal ganglion cell's spatial receptive fields although the spatial sampling density varies substantially with input eccentricity[31, 34]. The second is that the environment is statistically stationary, as then the correlations in time depend only on the temporal difference $t_i - t_j$. It can then be shown that[34] the principal components are moving ripple $\propto e^{i(2\pi\Omega f + 2\pi\omega t)}$, each of which has a 2D modulation frequency (Ω, ω), which can be indexed by $k \equiv (\Omega, \omega)$. The first encoding step is then a 2D Fourier transform $(K_o)_{(\Omega,\omega),j} \sim \exp[-2\pi i(\Omega f_j + \omega t_j)]$ of the input $S(f,t)$ to obtain $\mathcal{S}(\Omega,\omega) \propto \iint S(f,t) e^{-i(2\pi\Omega f + 2\pi\omega t)} df dt$. Meanwhile, the original input can be written as $S(f,t) \propto \iint \mathcal{S}(\Omega,\omega) e^{i(2\pi\Omega f + 2\pi\omega t)} d\Omega d\omega$, i.e., as a weighted sum of the moving ripples[19]. The second encoding step determines the gains for the ripple amplitudes $\mathcal{S}(\Omega,\omega)$ [34] as

$$g^2(\Omega,\omega) \propto Max\left\{\left[\frac{1}{2(1+\langle\mathcal{N}^2\rangle / \langle\mathcal{S}^2(\Omega,\omega)\rangle)}\left(1+\sqrt{1+\frac{2\lambda}{(\ln 2)\langle\mathcal{N}_o^2\rangle}\frac{\langle\mathcal{N}^2\rangle}{\langle\mathcal{S}^2(\Omega,\omega)\rangle}}\right)-1\right], 0\right\} \tag{8}$$

i.e., replacing g_k and $\left\langle \mathcal{S}_k^2 \right\rangle$ in equation (6) by the corresponding $g(\Omega,\omega)$ and $\left\langle \mathcal{S}^2(\Omega,\omega) \right\rangle$. If U is chosen as the inverse Fourier transform

$$U_{i,(\Omega,\omega)} \sim \exp[2\pi i(\Omega f_i + \omega t_i) + i\phi(\Omega,\omega)], \tag{9}$$

with an extra phase function $\phi(\Omega,\omega)$, then the encoding transform is $K_{ij} = \sum_{(\Omega,\omega)} U_{i,(\Omega,\omega)} g(\Omega,\omega)(K_o)_{(\Omega,\omega),j}$. This gives

$$\begin{aligned} & K(f_i, t_i; f_j t_j) \equiv K(f_i - f_j, t_i - t_j) \\ \propto & \iint g(\Omega,\omega) \exp\left[2\pi i(\Omega(f_i - f_j) + \omega(t_i - t_j)) + i\phi(\Omega,\omega)\right] d\Omega d\omega, \end{aligned} \tag{10}$$

which depends only on the differences $f_i - f_j$ and $t_i - t_j$. Applying this transform to input S to give output $O_i(t_i) = \iint df_j dt_j K(f_i - f_j, t_i - t_j) S(f_j, t_j)$, we see, by comparison with equation (1), that the STRF is $STRF(f,t) = K(f_i - f, t)$. This is a temporal filter tuned to sound frequency with a tuning pattern governed by $g(\Omega,\omega)$, and centered around frequency f_i. Changing the center frequency from f_i to f_j is like shifting from one output neuron i to another neuron j. Altering the phase $\phi(\Omega,\omega)$ in equation (9) alters the STRF shape, in particular to ensure its temporal causality. In physiology, modulation tuning function (MTF) is

often mentioned as the Fourier transform of auditory receptive field [19]. Therefore, it is clear from equation (10) that the gain profile $g(\Omega, \omega)$, which is determined by efficient coding, corresponds to the magnitude of the MTF. However, the shape of an STRF is determined by the phase as well as the magnitude of the MTF, and efficient coding does not strongly constrain the phase. Therefore, while we will illustrate the general properties of some example STRFs predicted by the theory by choosing particular U transforms (governed by the additional requirements of spectro-temporal locality and causality), in the Results, we will generally compare physiological data to the magnitudes of the MTFs that the theory predicts.

In the Results, we will discuss the efficient coding framework for situations both with (e.g., to study temporal aspects of STRFs) and without (e.g., to study their spectral aspects) translation invariance in input statistics.

Results

To illustrate how the framework explains and predicts physiological experiments, we first discuss a few examples when the temporal or the spectral dimension is omitted, and then show a full spectro-temporal STRF.

The spectral filter SRF

We first omit time, treating the input $S(f)$ as varying only in frequency. In this case, the encoding filter reduces from being an STRF to an SRF. We take f_i as one of 250 discrete values $i = 1, 2, ..., 250$, from low to high frequencies; hence input S is a one dimensional vector $S = (S_{f_1}, S_{f_2}, ..., S_{f_{250}})^T$. In simulations, input sample S is generated by smoothing a random noise vector $S' = (S'_{f_1}, S'_{f_2}, ..., S'_{f_{250}})^T$ (Figure 3A), with all the components S'_{f_i} taken to be independent, zero mean, unit variance, Gaussian noise. Specifically

$$S_{f_i} = \sqrt{I_F} \sum_j M_{ij} S'_{f_j} \tag{11}$$

where I_F is a factor to scale the overall input power intensity, and M is the smoothing matrix with elements

$$M_{ij} = A_i \hat{M}_{ij} \tag{12}$$

explained in detail below. Here $A_i = \frac{250-i}{300} + 0.1$ controls the scale of the signal S_{f_i}, which decays with i (like in an environment in which high frequency sounds do not propagate well), and $\hat{M}$ is a normalized smoothing matrix with elements $\hat{M}_{ij} = \tilde{M}_{i-j}/NORM$, in which

$$\tilde{M}_{i-j} = \begin{cases} 0.54 + 0.46 \cos(\frac{2\pi(i-j)}{L}), & \text{if } -L/2 \le i-j \le L/2 \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

$NORM = \left(\sum_{a=-L/2}^{L/2} \tilde{M}_a^2\right)^{1/2}$ is a normalization constant, and L controls the range of frequency difference $|f_i - f_j|$ for significant correlation coefficient between the variation of S_{f_i} and that of S_{f_j}.

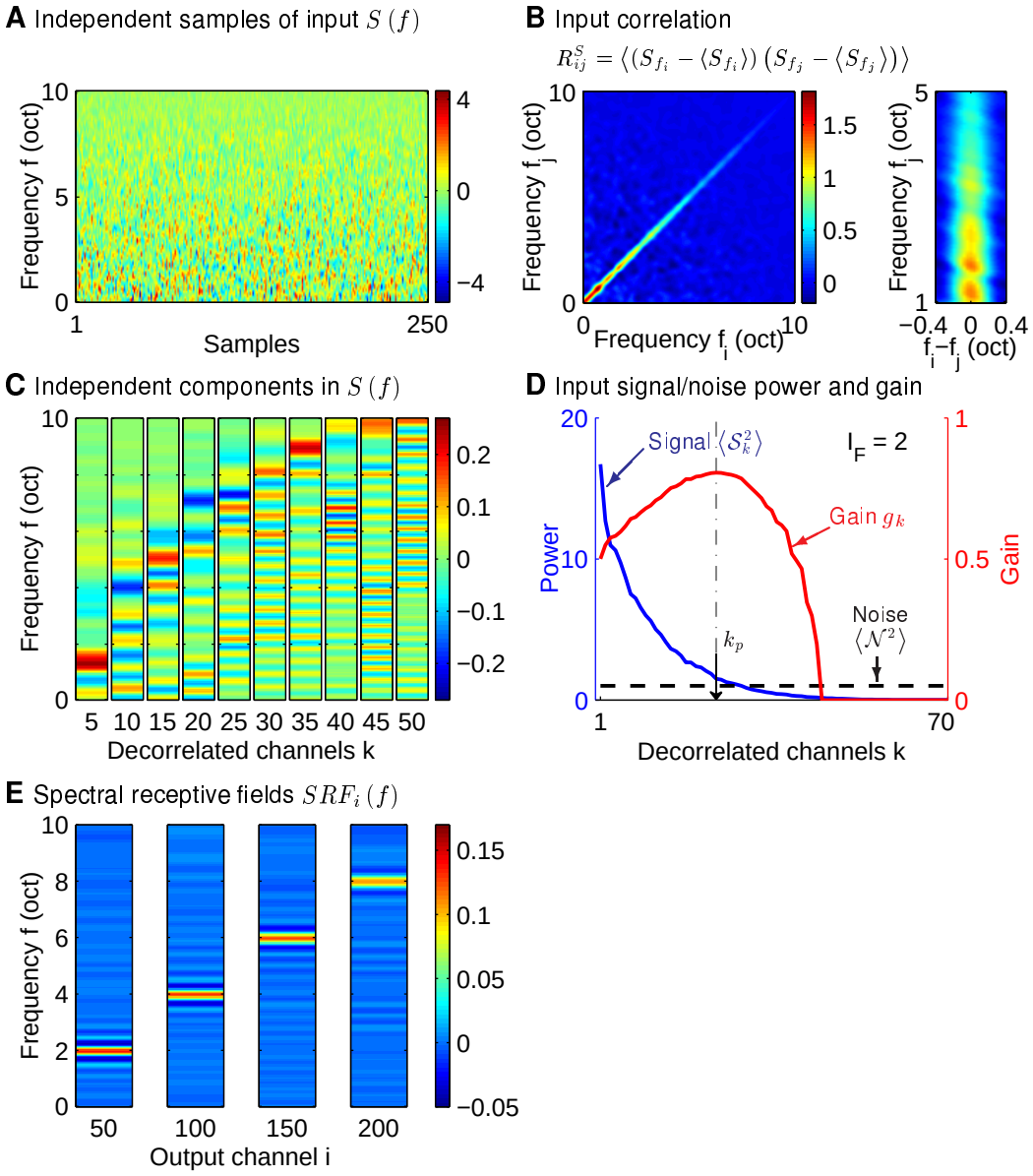

Figure 3: Simulation of the efficient spectral kernel SRF, when the temporal dimension is omitted. (A) 250 samples of input spectra $S(f)$, each of which is smoothed Gaussian white noise in the frequency domain (equations (11 - 13), $I_F = 2$, $L = 14$). (B) Correlation between different frequency channels $S(f)$. Left: Correlation R^S; Right: an zoomed-in view, as R^S_{ij} vs $f_i = f_j$. (C) Ten examples of eigenvectors $V^k(f)$ of the correlation matrix R^S in B; each is an independent component in $S(f)$. Smaller indices k are associated with larger eigenvalues. (D) Gain profile (peaking at k_p), and signal and noise power in decorrelated channels. (E) Four examples ($i = 50$, 100, 150, and 200) of spectral receptive fields $SRF_i(f) = \sum_k g_k V^k(f_i) V^k(f)$; each prefers input frequencies around f_i.

Consequently, each S_{f_i} is also a zero mean Gaussian random variable, and the input correlations comprise a 250x250 matrix $R^S = I_F M M^T$. One could also estimate R^S from input samples S (as when animals adapt their auditory system to environmental sound through experience), in which case element $R^S_{ij} = \left\langle (S_{f_i} - \langle S_{f_i} \rangle) \left(S_{f_j} - \langle S_{f_j} \rangle \right) \right\rangle$. Figure 3B illustrates R^S (obtained numerically from 250 samples of S in Figure 3A, of course one could use more than 250 samples to estimate R^S) for $L = 14$. The correlation $R^S_{ij} = I_F A_i A_j (\hat{M}^2)_{ij}$ scales with strengths of the original signals $S_{f_i}$and S_{f_j} through the scales A_i and A_j, and so decays with frequency f_i and f_j. Thus the statistics of the stimulus ensemble are not translation invariant in the spectral frequency f. Nevertheless, the correlation coefficient

$$C_{ij} \equiv \frac{R^S_{ij}}{\sqrt{R^S_{ii} R^S_{jj}}} = \frac{(\hat{M}^2)_{ij}}{\sqrt{(\hat{M}^2)_{ii} (\hat{M}^2)_{jj}}}$$

does depend mainly on the (frequency) difference $|i - j|$, since $(\hat{M}^2)_{ii}$ is almost independent of i and $(\hat{M}^2)_{ij}$ depends mainly on $|i - j|$ except for the very small or very large i and j. This is evident in the fact that the rate of decay of R^S_{ij} with the difference $|f_i - f_j|$ in Figure 3B is almost constant. Since the stimulus ensemble is not translation invariant, we will use the general formulation to obtain the SRF. From R^S, we obtain its 250 eigenvalues and the corresponding eigenvectors. Each of these is a vector with 250 components. We list them in the order of descending eigenvalues, denoting the k^{th} eigenvector as $V^k \equiv [(K_o)_{k1}, (K_o)_{k2}, ..., (K_o)_{kj}, ...(K_o)_{k250}]^T$, and placing it as the k^{th} row vector of the K_o transform matrix. Figure 3C depicts the eigenvectors for $k = 5, 10, ..., 50$, where smaller k is associated with a larger eigenvalue. Each principal component or eigenvector can be seen as a special input spectrum pattern $S = V^k$, while a general input $S = \sum_k \mathcal{S}_k V^k$ is a linear sum of the principal components with weights $\mathcal{S}_k$. The first encoding step is thus a transformation of the original input S by K_o to obtain the decorrelated signal $\mathcal{S}_k$, for $k = 1, 2, ..., 250$. The average power in $\mathcal{S}_k$ is the k^{th} eigenvalue of matrix R^S

$$\left\langle \mathcal{S}^2_k \right\rangle = (K_o R^S K_o^T)_{kk}$$

The eigenvectors look roughly like oscillating waveforms (spectral oscillations) with different oscillation rates, and are comparable to the sinusoidal bases in the Fourier transform. They also resemble the "ripples" used in physiological experiments. This is because the input correlations are roughly translation invariant, at least within a small range of frequencies in which the signal power $\left\langle S^2_f \right\rangle$ is roughly independent of f (just like in vision when the statistics of inputs sampled at the retina can be seen as roughly translation invariant within a local region). Also note that smaller or larger k is associated with eigenvectors with fewer or more oscillations. This makes k relate monotonically to the spectral modulation frequency (corresponding to the "ripple frequency" Ω in physiological experiments). Larger eigenvalues, i.e., larger signal powers $\left\langle \mathcal{S}^2_k \right\rangle$, are associated with fewer spectral modulations or smaller indices k, because inputs of more similar sound frequencies are more correlated with each other, i.e., R^S_{ij} decreases with increasing $|f_i - f_j|$. The analogy between the eigenvectors and the Fourier bases can be understood as follows: if R^S is strictly translation invariant, then the eigenvectors are sine waves with different spectral modulation frequencies Ω. The eigenvalues are the Fourier transforms of $R^S_{ij} \equiv R^S(f_i - f_j)$, and hence they decrease with the modulation frequency Ω because $R^S(f_i - f_j)$ is non-negative and decreases with increasing $|f_i - f_j|$.

The second encoding step is to assign the gain g_k to each of these channels $\mathcal{S}_k$ according to equation (6), giving $\mathcal{S}_k \rightarrow g_k \mathcal{S}_k$ (see Figure 3D; $I_F = 2$, $\langle \mathcal{N}^2 \rangle = 1$ and $\lambda / \langle \mathcal{N}_o^2 \rangle = 10$). Note that while the signal power $\langle \mathcal{S}_k^2 \rangle$ decreases with increasing k, the gain magnitude g_k first increases with k and then decreases and drops to zero at higher k.

The gain for small k is low since the SNR $\langle \mathcal{S}_k^2 \rangle / \langle \mathcal{N}^2 \rangle$ is high enough to make amplifying $\mathcal{S}_k$ less necessary. From equation (6)[34],

$$g_k^2 \propto \langle \mathcal{S}_k^2 \rangle^{-1} \qquad \text{when } \langle \mathcal{S}_k^2 \rangle / \langle \mathcal{N}^2 \rangle \rightarrow \infty \tag{14}$$

This implies that $g_k^2 \langle \mathcal{S}_k^2 \rangle$ = constant for sufficiently large SNRs. When each principal component $\mathcal{S}_k$ is a modulation frequency mode, this gain profile g_k is often called whitening. At smaller signal powers, the gain increases so as to utilize the channel's dynamic range fully. However, when SNR is too small, for example, when noise power is higher than signal power $\langle \mathcal{S}_k^2 \rangle / \langle \mathcal{N}^2 \rangle < 1$, gain decreases with decreasing $\langle \mathcal{S}_k^2 \rangle$[34]. This is because such input components are dominated by noise, and amplifying noise increases neural cost. Thus, in general, when $\langle \mathcal{S}_k^2 \rangle$ decreases with increasing k, the gain profile has a band-pass shape, first increasing, and then decreasing with increasing k (see the red curve in Figure 3D). The peak of the gain occurs at $k = k_p$, where $\langle \mathcal{S}_k^2 \rangle / \langle \mathcal{N}^2 \rangle \simeq 1$.

Third, taking $U = K_o^{-1}$ in order to localize the receptive Fields as best as possible, the overall encoding transform is $K = K_o^{-1} g K_o$. Here, the gain matrix is diagonal having elements $g_{kk} = g_k$. When $K_o^T = K_o^{-1}$ (as when the eigenvectors are real and othornormalized)

$$K_{ij} = (K_o^T g K_o)_{ij} = \sum_k g_k (K_o)_{ki} (K_o)_{kj} = \sum_k g_k V_i^k V_j^k.$$

As the overall encoding transform gives outputs $O = KS + noise$, where $noise = KN + N_o$, the i^{th} output neuron O_i has its SRF as a vector of weights for inputs S_{f_j} of various frequencies $j = 1, 2, ..., 250$

$$SRF_i = (K_{i1}, K_{i2}, ..., K_{i250}) = \sum_k g_k V_i^k V^k$$

It can thus be seen as a weighted sum of the eigenvectors V^k of the input correlation matrix, with weights $g_k V_i^k$ for output neuron i. Figure 3E shows SRFs for four different output neurons (or channels i). These SRFs have different preferred frequencies f, so that the preferred frequencies of all the output neurons span the whole input frequency range. The shapes of the SRF depend on the input statistics via the dependence of V^k and g_k on the input correlation matrix R^S. In particular, for sufficiently high input SNR, while a neuron is excited by its preferred frequency, it is suppressed by nearby frequencies. This form of contrast enhancement achieves a measure of decorrelation between neighboring output neurons that would otherwise reflect the strong correlations between neighboring frequencies. For SRFs tuned to higher frequencies, the center excitatory regions are larger and the surround suppression is weaker. This is because SNRs are weaker for higher frequency inputs (the dependency of SRF on SNR will be discussed in the next sub-section). If the input statistics are strictly translation invariant, the SRFs for different output channels will have the same shape, and will just be centered on different frequencies.

Adaptation of SRF to input signal-to-noise ratio

When sound intensity decreases, the basilar membrane in the cochlea undergoes a smaller vibration. This decreases the magnitudes of input signals S, and so, if the level of the noise stays unchanged, the signal-to-noise ratio $\langle \mathcal{S}_k{}^2 \rangle / \langle \mathcal{N}^2 \rangle$ will decrease. This will change the optimal encoding gain g_k via equation (6), and thus change the final SRFs. In our example, we simulate the change in input intensity by changing I_F in equation (11).

Figure 4A shows three example input intensity profiles $\langle \mathcal{S}_k^2 \rangle$, and the corresponding gain profiles g_k. While an overall change of input intensity merely scales the profile $\langle \mathcal{S}_k^2 \rangle$ up and down, the gain profile g_k does not trivially scale up and down. When input intensity decreases, the k at which $\langle \mathcal{S}_k^2 \rangle / \langle \mathcal{N}^2 \rangle = 1$ becomes smaller, thereby decreasing the k_p at which g_k peaks. Consequently, the gain profile turns from being band-pass to being low-pass (Figure 4A).

The non-zero gain at higher k implies sensitivity to weaker principal components with more spectral oscillations (or higher "ripple frequencies"). Thus, as input intensity decreases, the overall SRF filter changes in two ways (Figure 4B): (1) it fluctuates less (i.e., has fewer excitatory and inhibitory regions, and with decreased strength inhibitory regions); (2) the width of the excitatory and inhibitory regions increases, as the result of losing contributions from spectral modulations V^k with higher modulation frequencies.

The insights from Figure 4B can help to understand the difference between the four SRFs in Figure 3E. Given the I_F as in Figure 3, one may divide the whole sound frequency range into two ranges of equal bandwidth, one for the lower and the other for the higher f's, and treat the two ranges as if they were two different stimulus ensembles. If one ignores the overall sound frequency difference between these two ensembles, then these two ensembles differ from each other only in their SNRs, with a higher SNR for the ensemble for the lower sound frequencies f. In this perspective, one can understand why a SRF tuned to the lower frequencies in Figure 3E has a narrower excitatory region and a stronger surround suppression than a SRF tuned to higher frequencies, using the insights gained from Figure 4. (In comparing Figure 4B with Figure 3E, one should note that each SRF in Figure 4B is depicted by zooming to the frequency region around the preferred frequency f of the SRF.) One may even view the four SRFs in Figure 3E as if they were each exposed to one of the four different stimulus ensembles that differ in SNRs (and in sound frequency f, and we ignore this difference). Within each of these stimulus ensembles, the input statistics may be seen as approximately translation invariant, since $\langle S_f^2 \rangle$ is almost independent of f and the correlation $\langle S_f S_{f'} \rangle$ is approximately only a function of the frequency difference $f - f'$ within a small range of frequency f.

Adaptation of SRF to input signal correlation

As well as adapting to the input SNR, the SRF can adapt to the signal correlations in the input. These can also vary across auditory environments. We generate two stimulus ensembles (Ensemble$_{short}$ and Ensemble$_{long}$) based on equation (11), with short and long range (in frequency space) correlations between

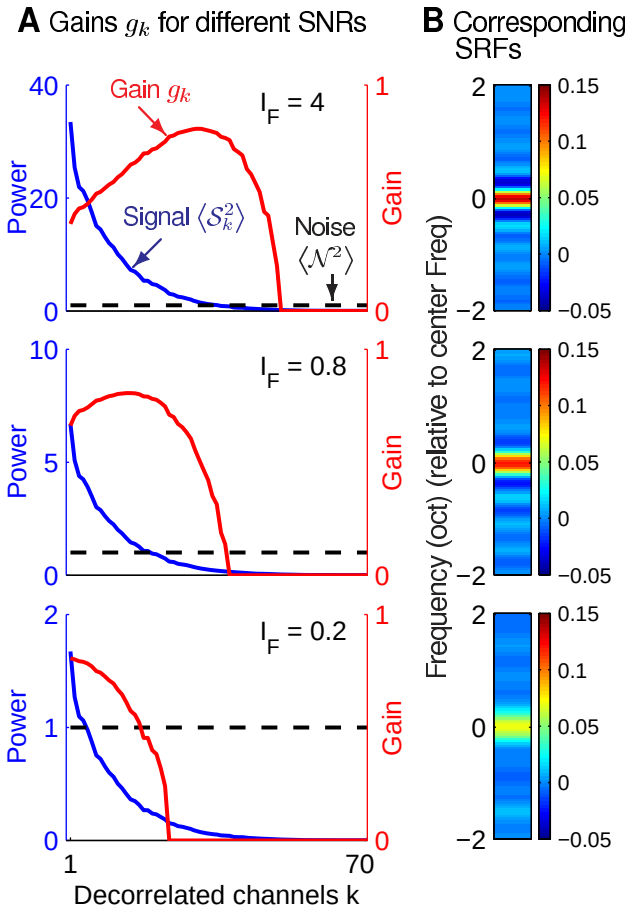

Figure 4: The effect of signal-to-noise ratio (SNR) on gain g_k and the spectral receptive field (SRF). Same stimulus ensemble as in Figure 3A except the overall SNR has been scaled by I_F. (A) Gain control (red), signal (blue), and noise power (black) under high, medium and low SNR. (B) The corresponding SRFs of one output neuron (channel #120) in the three SNR cases.

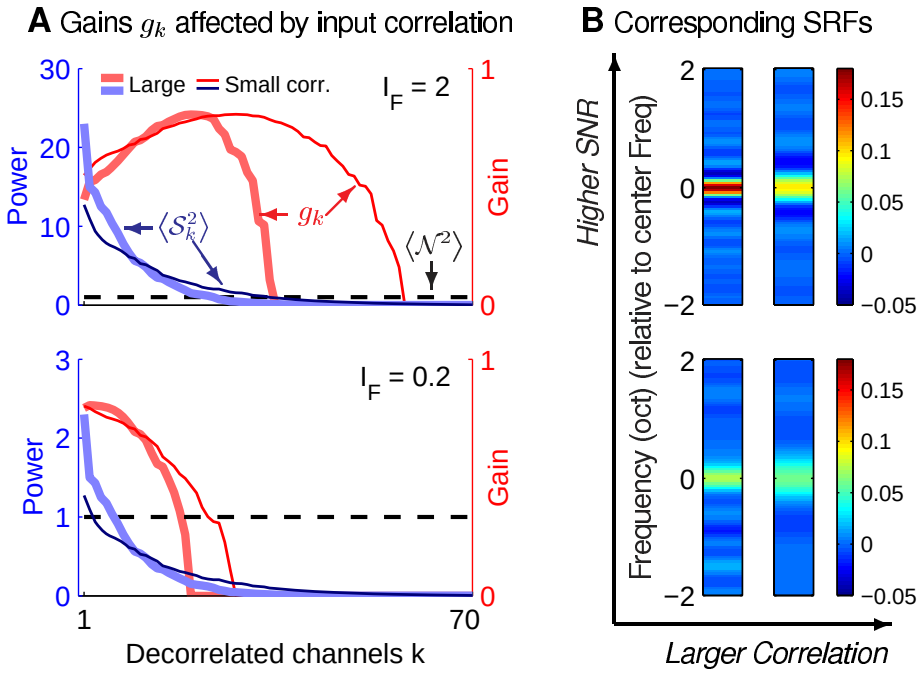

Figure 5: Adaptation of gain g_k and spectral filter kernel SRF to input correlations under high/low SNR. Same input ensemble as that in Figure 3A, except that the smoothing parameter, $L = 10$ and $L = 20$, are set for short and long range correlations, respectively. Analogous figure format as in Figure 4, with added illustrations of the adaptation to input correlations. The thick and thin curves correspond to quantities for inputs with large and small correlations respectively, blue/red curves plot signal power $\langle \mathcal{S}_k^2 \rangle$ and gain g_k respectively.

inputs $S_{f_i}$and S_{f_j} of different sound frequencies. We do this by setting the smoothing length L in equation (13) to be $L_{short} = 10$ and $L_{long} = 20$. Since short and long range correlations give respectively smaller and larger correlations or degrees of input redundancy, in this paper, we use the terms short/long-range and small/large correlations interchangeably. The two stimulus ensembles are made to have the same overall signal power $\sum_k \langle \mathcal{S}_k^2 \rangle$, and consequently their $\langle \mathcal{S}_k^2 \rangle$ vs. k curves cross each other at a particular frequency k_x (Figure 5A). In Ensemble$_{long}$, signal power $\langle \mathcal{S}_k^2 \rangle$ is more concentrated in lower k's, and the "bandwidth" of gain, i.e., the range of k's with substantial g_k, is consequently narrower.

If $\langle \mathcal{S}_k^2 \rangle > \langle \mathcal{N}^2 \rangle$ at $k = k_x$, the k at which signal power $\langle \mathcal{S}_k^2 \rangle / \langle \mathcal{N}^2 \rangle = 1$ is larger in Ensemble$_{short}$ (Figure 5A, upper panel, $I_F = 2$, $\langle \mathcal{N}^2 \rangle$=1, $\lambda/\langle \mathcal{N}_o^2 \rangle = 10$). Thus, the frequency k_p at which gain g_k peaks is also larger in Ensemble$_{short}$. If the SNR is lower, so that $\langle \mathcal{S}_k^2 \rangle < \langle \mathcal{N}^2 \rangle$ at $k = k_x$, then k_p is instead smaller in Ensemble$_{short}$ than in Ensemble$_{long}$. However, this is less apparent since gain profiles in both ensembles become "low-pass" in k implying that there is no obvious "peak position" (Figure 5A, lower panel, $I_F = 0.2$). Nevertheless, the cutoff frequency k where $g_k = 0$ is always smaller for Ensemble$_{long}$ (Figure 5A), and the optimal SRFs for it consequently enjoy a greater spectral extent (i.e., the SRFs are non-zero for a larger range of f (Figure 5B). Intuition for this effect is that for it to be effective as either a contrast enhancing filter at a high SNR, or a smoothing filter at a low SNR, the SRF's spectral extent should match the range of the input correlations.

The temporal filter TRF

We can similarly ignore the frequency dimension of the input to understand the temporal receptive field (TRF). This is determined from the way $O_t = \sum_{t'} K_{tt'} S_{t'}$+noise, the input temporal sequence $S = (S_{t_1}, S_{t_2}, ..., S_{t_i}, ...)$ is transformed to the output temporal sequence O. In a statistically stable auditory environment, the input correlation should be time shift invariant, i.e., $R^S_{tt'} = \langle S_t S_{t'} \rangle$ should depend only on $t - t'$. Denote $R^S_{tt'} = R^S(t - t')$. Then, the de-correlating transform K_o should just be a Fourier transform $(K_o)_{\omega t} \propto e^{-i \cdot 2\pi \omega t}$ with the principal component $\mathcal{S}_\omega \propto \sum_t e^{-i \cdot 2\pi \omega t} S_t$ being the Fourier Amplitude of the relevant mode. Here we use index ω instead of k to denote the principal component to signify the association with the temporal Fourier amplitude. The average power $\langle \mathcal{S}_\omega^2 \rangle \propto \int dt R^S(t) e^{-i \cdot 2\pi \omega t}$ is simply the Fourier transform of the input temporal correlation. If we set $A_i = 1$ in equation (12) to generate inputs with shift invariant correlation, then $\langle \mathcal{S}_\omega^2 \rangle = I_F M^2(\omega)$ where $M(\omega)$ is the Fourier amplitude of $M(i - j) = M_{ij}$. The gain control $\mathcal{S}_\omega \to g_\omega \mathcal{S}_\omega$ in the second encoding step is determined by equation (6) (substituting ω for k). The final TRF will be the transform $K = UgK_o$ given an appropriate choice of U.

However, the actual procedure to obtain the TRF is trickier in that the U transform in the third encoding step to give the overall $K = UgK_o$ has to be chosen to satisfy the causality constraint. That is, the output O_t at time t should only depend on past input $S_{t'}$ for $t' \leq t$, i.e., $K_{tt'} = 0$ for $t' > t$. Moreover, it is better for the TRF to have a short temporal span and latency, an outcome that can be achieved by assuming that the optimal temporal filter K has a minimum phase-shift[57]. Short latency can feasibly be implemented by

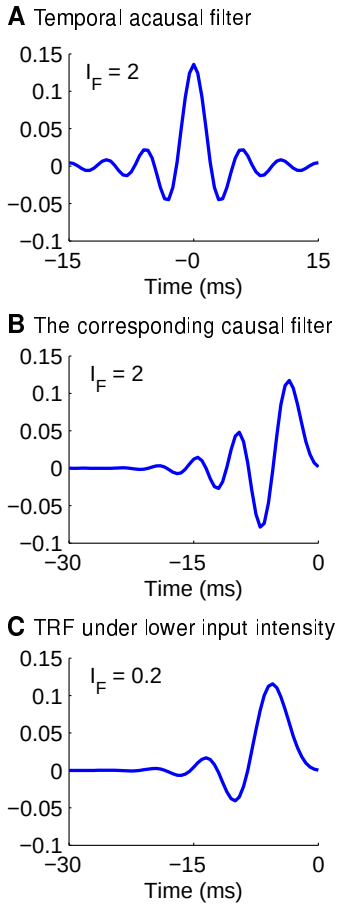

Figure 6: Simulation of temporal receptive field TRF, when the spectral dimension is omitted. The same stimulus ensemble is used as in Figure 3A, except the factor $A_i = 1$ in equation (12) to ensure translation invariance of correlation. (A;B) Demonstration of transforming an acausal temporal filter (A) to its causal minimum-phase counterpart (B) at a relatively high input SNR. (C) TRF for a relatively low input SNR.

neural synaptic and membrane mechanisms that typically have time constants no longer than a few hundred milliseconds[58]. Hence, these offer credible constraints on the TRF. Note that if we choose $U = K_o^{-1}$, i.e., $U_{t\omega} \propto e^{i\cdot 2\pi\omega t}$, then $K_{tt'} \propto \sum_\omega g_\omega e^{i\cdot 2\pi\omega(t-t')}$would be an even function of $t-t'$ and thus not a causal temporal filter given gains g_ω that are all real. The filter K can be made causal and minimal phase by choosing another U simply as $U_{t\omega} \propto e^{i\cdot 2\pi\omega t+i\phi(\omega)}$ with a particular phase function $\phi(\omega)$, so that $K_{tt'} \propto \sum_\omega g_\omega e^{i\cdot 2\pi\omega(t-t')+i\phi(\omega)}$. Instead of directly obtaining this phase function $\phi(\omega)$, we can also equivalently obtain this minimum phase shift causal filter by transforming the acausal K using standard procedures in signal processing theory as follows (see [57] for the proof). Given a non-causal filter $K\left(\hat{t}\right)$ with finite non-zero values in discrete time $\hat{t} = -M, -M+1, ..., 0, ..., N-M-1, N-M$, first let $t = \hat{t} + M$ to make a causal filter $K(t)$ whose nonzero values are at $t = 0, 1, ..., N$. Second define

$$\tilde{K}(z) = K(0) + K(1) z^{-1} + K(2) z^{-2} + \cdots + K(N) z^{-N}.$$

Among the N complex roots of the equation $\tilde{K}(z) = 0$, let z_i denote the roots with $|z_i| > 1$ and z_j the other roots with $|z_j| \leq 1$. Third, let

$$\begin{aligned}\tilde{K}_{\min} &= z^{-N} \textstyle\prod_i (z - 1/z_i) \prod_j (z - z_j) \\ &= K_m(0) + K_m(1) z^{-1} + K_m(2) z^{-2} + \cdots + K_m(N) z^{-N}\end{aligned}$$

The coefficients $K_m(t)$, $t = 0, 1, ..., N$ are the values of the desired causal minimum phase filter. One example of this process is demonstrated in Figure 6A (before the minimum phase adjustment) and Figure 6B (after the minimum phase adjustment)($I_F = 2, L = 14$).

The temporal kernel also depends on the SNR and the input correlations. The change in g_ω when sound intensity becomes lower is similar to that in the spectral case: from band-pass to low-pass. A temporal kernel under lower SNR is demonstrated in Figure 6C. The changes in g_ω and TRF with input correlations are analogous to those in the spectral case as well (figure not shown).

The two dimensional STRF

Finally, we show examples of the two dimensional $STRF(f, t)$. Here, we extended the assumption of shift invariance in the input correlations to the spectral dimension for the convenience of calculation. This assumption is reasonable when individual STRFs cover sufficiently small ranges of frequencies that the correlation in the spectral space is almost translation invariant within that range, as we see in our SRF examples. Then, spectral and temporal dimensions can be de-correlated at the same time by performing a 2-D Fourier transform on inputs $S(f, t)$, with the moving ripples as decorrelated channels, each denoted by a 2D index (Ω, ω) marking the spectral and temporal modulation frequencies.

Let the signal power in the de-correlated channels (Ω, ω) for input $S(f, t)$ be $\left\langle \mathcal{S}^2(\Omega, \omega) \right\rangle = I_F F(\Omega, \omega)$. Here, $F(\Omega, \omega)$ typically decays with modulation frequency $|\Omega|$ and $|\omega|$ since most natural inputs have input correlation $\langle S(f, t) S(f', t') \rangle$ that decays with $|f - f'|$ and $|t - t'|$. I_F is a scale factor that controls the SNR.

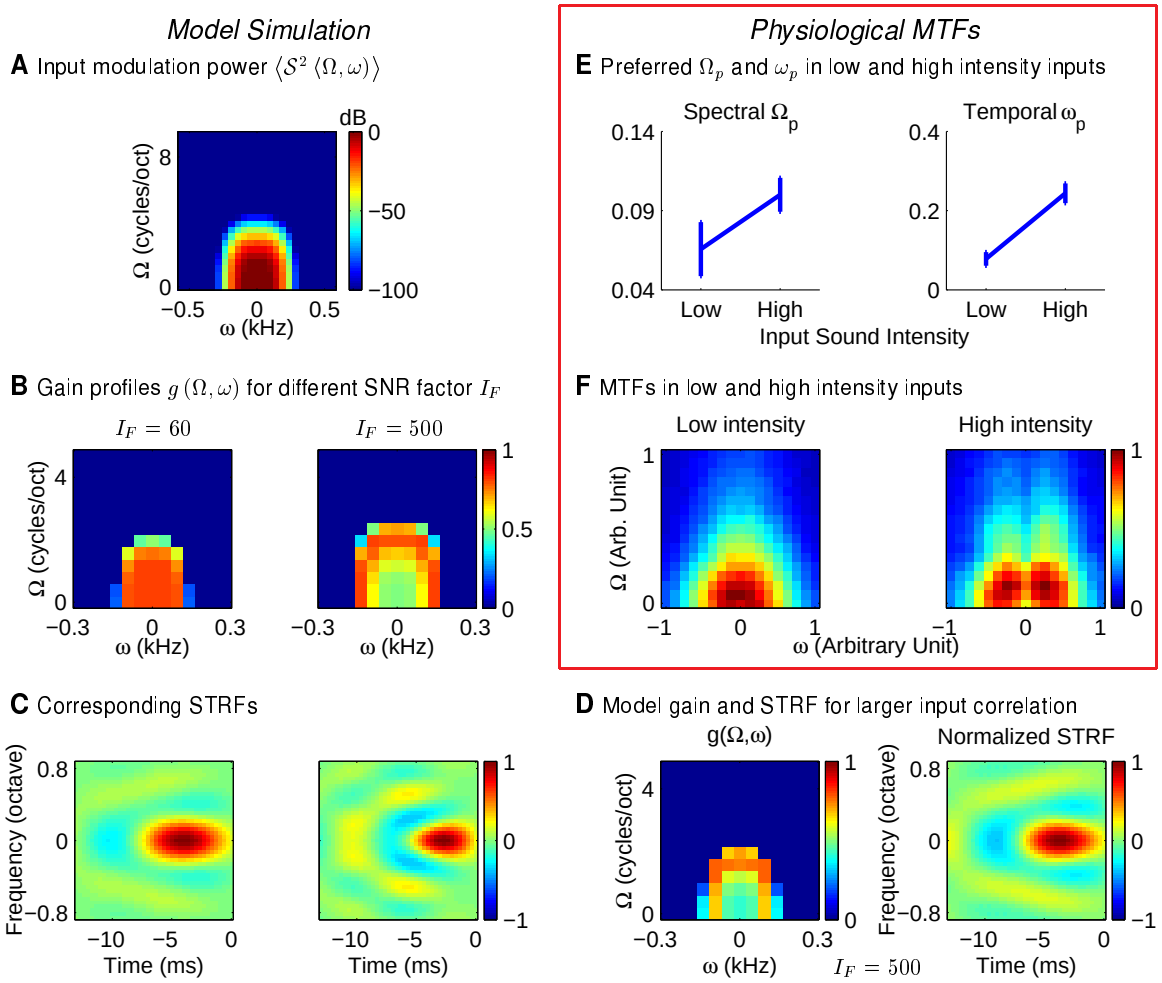

Figure 7: The 2D STRFs/MTFs implied by efficient coding and found physiologically. (A) input power $\langle \mathcal{S}^2(\Omega,\omega)\rangle$ (equation (15), $\Omega_0 = 4$, $\omega_0 = 4$) in decorrelated channels. (B, C) MTF profile $g(\Omega,\omega)$ and the corresponding STRFs with two SNRs (scaled by I_F's). (D) $g(\Omega,\omega)$ and STRF as in B;C (when $I_F = 500$) except with larger input correlations ($\Omega_0 = 3.2$, $\omega_0 = 3.2$ in equation (15)). (E;F) Modulation transfer functions (MTFs) and their properties at low and high input sound intensities averaged over 40 IC neurons from Lesica and Grothe[7]. Here, (Ω_p, ω_p) is the spectral-temporal modulation frequency where the MTF peaks. Modulation frequencies in E and F are normalized by the same value across cells and intensities. Error bars in E indicate standard errors. The magnitude patterns of the MTFs for all neurons are normalized to peak value 1. Their average across neurons at each input intensity is then normalized to the same peak value and displayed in F.

We use the following example in our simulations

$$\left\langle \mathcal{S}^2(\Omega,\omega)\right\rangle = \frac{I_F}{NORM}\exp[-(|\Omega|/\Omega_0)^3 - \alpha(|\omega|/\omega_0)^3] \tag{15}$$

where $\alpha = 1.8$, Ω_0 and ω_0 are parameters that control input correlation, and $NORM = \sum_\Omega \sum_\omega \exp[-(|\Omega|/\Omega_0)^3 - \alpha(|\omega|/\omega_0)^3]$ is a normalization factor. Figure 7A shows an example with $\Omega_0 = 4, \omega_0 = 4$. According to equation (8), the gain $g(\Omega,\omega)$ can be obtained as shown in Figure 7B ($\left\langle \mathcal{N}^2\right\rangle = 1$, $\lambda/\langle \mathcal{N}_o^2\rangle = 10$, and $I_F = 60,500$). In particular, in the frequency range (Ω,ω) in which noise is negligible relative to the signal, the gain

$$g(\Omega,\omega) \propto (\left\langle \mathcal{S}^2(\Omega,\omega)\right\rangle)^{-1/2} \tag{16}$$

specifies the whitening filter of equation (14). This gain profile changes from being a band-pass to a low-pass two dimensional filter as the SNR is lowered.

As we noted before, efficient coding predicts the gain $g(\Omega,\omega)$, or the modulation transfer function (MTF), but does not precisely determine the STRF shape. The latter depends on the less constrained U transform. Therefore, we qualitatively compare our $g(\Omega,\omega)$ for two different I_F's with the MTFs obtained from physiological experiments under two different input sound levels. Figure 7E and Figure 7F are obtained from data on STRFs of 40 cells in the inferior colliculus of animals exposed to natural rain sound at low and high sound levels[7]. We first did a two-dimensional Fourier transform on the STRF of each cell to obtain its MTF. Then the spectral modulation frequency Ω_p and the temporal modulation frequency ω_p where the MTF has its maximum value were identified and normalized by a fixed value across cells. The average Ω_p and ω_p across all cells are shown in Figure 7E. These two "peak frequencies" both increased when sound intensity increased. The physiological MTF averaged across all cells (Figure 7F) also becomes higher pass, both spectrally and temporally, under higher sound intensities, as predicted by efficient coding (Figure 7B).

For completeness, we illustrate in Figure 7C the model STRFs from the gain profiles $g(\Omega,\omega)$, using an inverse Fourier transform with a proper phase function $\phi(\Omega,\omega)$ as the candidate U matrix. Specifically, the model STRF is

$$STRF(f,t) = \int d\Omega \int d\omega g(\Omega,\omega) e^{i\cdot 2\pi\Omega f + i\cdot 2\pi\omega t + i\phi(\Omega,\omega)}$$

where the phase $\phi(\Omega,\omega)$ is chosen to make the STRF causal, and with minimum phase shifts in the temporal dimension. In practice, the STRF is obtained as follows, by extending our method for obtaining the causal 1-D TRF. For each Ω, we first obtain the temporal acausal filter

$$K(\Omega,t)_{acausal} = \int g(\Omega,\omega) e^{i\cdot 2\pi\omega t} d\omega$$

and then transformed this into a causal minimum phase filter $K(\Omega,t)$ as for the one dimensional TRF filter. The final two-dimensional STRF is then

$$STRF(f,t) = \int K(\Omega,t) e^{i\cdot 2\pi\Omega f} d\Omega$$

In general the model STRF has its highest amplitude at the preferred frequency on the spectral axis and for short latencies (i.e., the early part of the temporal axis). At low I_F, the STRF has a large excitatory region and a weak inhibitory surround (Figure 7C). At larger I_F, the STRF involves more excitatory and inhibitory regions with an increased inhibitory strength. Overall this has a more band-pass gain profile. Meanwhile, the bandwidth for the gain $g(\Omega, \omega)$ increases with I_F, thus shrinking the width of the main excitatory region. Therefore, adaptation to higher sound levels makes the frequency-time tuning curve sharper, or equivalently more narrowly tuned and so, at a single cell level, supporting a more precise read out of the time and frequency of auditory input. Qualitatively, physiologically observed STRFs adapt to the input intensity in the same way [7] (also see[14]).

The model also predicts changes to MTFs and STRFs for different input correlations. Figure 7D shows the gain function $g(\Omega, \omega)$ and STRF for an example in which the input has longer-range correlations in both spectral and temporal dimensions (we set $\Omega_0 = 3.2, \omega_0 = 3.2$ while holding $I_F = 500$ as in the high SNR case in Figure 7B and 7C). The peak modulation frequency in $g(\Omega, \omega)$ is decreased, and the excitatory region is wider compared with counterparts in Figure 7B and 7C at high SNR. This is consistent with our 1-D results in the spectral dimension (Figure 5).

Discussion

Summary of findings and predictions

In summary, this study set out to understand the computational role of auditory spectro-temporal receptive fields (STRFs). In particular, we generalized previous work[26] by proposing that STRFs are efficient codes for inputs which retain maximal information for a given neural cost associated with the output. We analyzed this proposal in detail for the case that input signals and noise are approximated as Gaussian. Mathematically, the STRF transform can be shown[34] to be composed of three abstract steps: input de-correlation, gain control, and multiplexing. For typical input statistics that are shift-invariant in sound frequency and time, the transform can be compared with two sorts of experimental data. First, gain control corresponds to the magnitude of the modulation transfer function of the STRFs. Second, by choosing the form of multiplexing to arrange the STRFs to have minimal phase, one can predict their full form. That the STRFs or the MTFs adapt to input statistics is a direct prediction of this efficient coding framework, since both the information conveyed and the neural coding cost depend on these statistics. Our efficient coding proposal is thus experimentally testable.

We made two particular predictions about the adaptation of the STRFs, one associated with input intensity, the other with input correlation. For the case of intensity, we predicted that the MTF of the STRFs should become more low pass when input intensity is lowered. Intuitively, as long as inputs at nearby frequencies and times are correlated, a low pass filter smoothes the input to reduce noise, whereas a band pass filter extracts differences between input frequencies and times to remove redundancy. Compared with a band

pass STRF, a low pass STRF has one or all of the following characteristics: (1) it has fewer excitatory and inhibitory regions; (2) each excitatory/inhibitory region has a larger size; (3) the secondary or opponent region, e.g., the inhibitory region for a STRF with an primary excitatory region, is weaker. All three characteristics help to smooth noise, a necessary strategy for weak inputs. In contrast, a band-pass filter has the opposite characteristics, so as not to increase the neural cost due to the transmission of redundant input information. These predictions are analogous to those seen in adaptations of visual coding to input SNR [29, 33, 34, 51, 52]. They also generalize previous accounts of the adaptation of the temporal auditory filter [26] to input intensity.

For the case of adaptation to input correlation, our framework predicts that the sizes of the excitatory and inhibitory regions of the STRFs should adapt to the range of input correlations. That is, input ensembles with longer range correlations in frequency and/or time should lead to STRFs with larger excitatory and inhibitory regions in the corresponding feature dimensions. Longer range input correlations are typically equivalent to greater input modulation power in the lower modulation frequency range in the stimulus ensemble. Equally, larger excitatory/inhibitory regions in the STRF are typically equivalent to its MTF being tuned to lower modulation frequencies. Thus, our prediction can be stated equivalently as saying that a stimulus ensemble with greater input power in the lower modulation frequency range, spectrally and/or temporally, should lead to neural MTFs tuned to the lower modulation frequency ranges. We demonstrated this form of adaptation for SRFs in Figure 5, and for STRFs in Figure 7 In particular, with a sufficiently high SNR, the MTF profile $g(\Omega, \omega)$ should whiten the ensemble specific input modulation power $\langle \mathcal{S}^2(\Omega, \omega) \rangle$.

Experimental evidence and tests of the predictions

Various experimental observations pertain to these predictions about adaptation to input intensity. Lesica and Grothe [7] presented natural rain sounds to gerbils and found that, for a majority of cells in inferior colliculus (IC), the STRFs have more excitatory/inhibitory regions for higher input sound levels, and only have excitatory regions, or at least very weak inhibitory regions for lower sound levels. Nagel and Doupe[14] conducted a similar study in field L of songbirds, an area analogous to mammalian auditory cortex. In both spectral and temporal dimensions, they found that the excitatory/inhibitory regions of the STRFs become smaller and sharper under higher sound intensity, while the number of such regions do not increase. These results paralleled those of an earlier study in which they only examined the temporal dimension of the receptive fields [58]. Both studies are consistent with our proposal that the MTF changes from lower to higher pass when input intensity (and hence, SNR) increases. They thus offer complementary confirmation of our predictions.

As mentioned in the Introduction, Lesica and Grothe[26] also examined the adaptation of the temporal receptive field(TRF) to vocalizations and ambient noises. They found that the TRF changed from being bandpass to lowpass when noise was mixed into the ensemble of vocalizations, and accounted for this finding in terms of efficient temporal coding. Their result can be understood as a special case of adaptation to SNR

in our framework, focusing on the temporal dimension of the STRF, and treating the addition of noise as a reduction in input SNR. According to the principle of efficient coding, the spectral receptive field should also have changed from bandpass to lowpass when this noise was added.

There are as yet few physiological experiments that pertain to our prediction about adaptation to input correlations. One study by Woolley et al [11] examined the STRFs of midbrain neurons in zebra finch in response to bird songs or modulation-limited noise. Compared to that of the noise, the input modulation power of the songs is more concentrated in lower modulation frequencies. The MTFs of the STRFs matched the corresponding modulation frequency spans, consistent with our theoretical prediction.

The studies by Woolley et al [11] and Lesica and Grothe [26] could be extended to different ensembles of natural stimuli, e.g., songs, speech, animal vocalization, and environmental background, each with its own particular input correlations [59]. Findings from such extended studies would provide a stern test of the efficient coding framework. Generally, the input modulation power $\langle \mathcal{S}^2(\Omega,\omega)\rangle$ in natural sounds decays with increasing modulation frequency (Ω,ω), at a rate that is specific to the ensemble [59]. Ensembles with faster decays have longer range input correlations (or larger correlations), as modelled in our Figure 5A and Figure 7BCD. We predict that this decay rate in $\langle \mathcal{S}^2(\Omega,\omega)\rangle$ should dictate the shape of the neural MTFs $g(\Omega,\omega)$, such that ensembles with faster decay should lead to neural MTFs focusing on lower modulation frequency ranges. In particular, for high input SNR, the MTF profile should be that of a whitening filter $g(\Omega,\omega) \propto (\langle \mathcal{S}^2(\Omega,\omega)\rangle)^{-1/2}$, with the upper frequency limit (Ω,ω) for this whitening (beyond which MTF quickly decays to zero) being around the frequency at which $\langle \mathcal{S}^2(\Omega,\omega)\rangle$ is comparable to the power level of the noise. The recent study by Rodriguez et al [59] showed that inferior colliculus (IC) neurons, when examined collectively as a population, do seem to whiten typical natural stimuli, in that the population MTF $g(\Omega,\omega)$ increases with frequency (Ω,ω) (up to a high frequency limit). This is to be expected for an efficient code, since natural input power $\langle \mathcal{S}^2(\Omega,\omega)\rangle$ decreases with frequency. However, the neural STRFs in this study were obtained (using the moving ripple stimuli) without specific adaptation to any particular natural stimulus ensemble. We predict that if the STRFs had been measured under adaptation to the natural sounds for high SNR, then the neural MTF profile, at a neural population level if not at individual neuron level, should be ensemble specific, i.e., whitening the input power $\langle \mathcal{S}^2(\Omega,\omega)\rangle$ of the adapting stimuli.

The neural implementation of the efficient STRF and its adaptations

We seek the overall effective STRF rather than its realization. Thus, it is important to note that the three separate steps of our mathematical analysis of the efficient STRFs are purely abstract. They do not correspond to an actual physiological implementation. In principle, when a receptive field is entirely linear, it can as well be implemented in a single step, as in multiple linear steps in a cascade. Meanwhile, the observation that STRFs adapt to changes in the statistics of auditory inputs, and indeed that visual receptive fields expand when the visual environment changes from bright outdoors to dark indoors[52], attest to the availability of the mechanisms for implementing (and thus adapting) efficient sensory coding.

We speculate that the adaptation of a STRF in a midbrain auditory neuron is likely to involve gain control in many intervening and distributed neural processes upstream along the auditory pathway [60]. Even a simple adaptation of efficient coding, in the large monopolar cells (LMCs) in an insect compound eye to changes in the distribution of input contrasts in the visual environment, involves multiple stages of processes, some in the photoreceptors and others in lamina from the receptors to the LMCs[61]. Synaptic and intrinsic mechanism were also found in the adaptation of retinal bipolar and ganglion cells to temporal contrast [62, 63]. Considering the multiple synapses from the hair cells to IC or auditory cortex, and the many recurrent and feedback networks with both excitatory and inhibitory connections [64, 65] in this pathway (for example, medial olivocochlear (MOC) efferent effects [66]), we speculate that gain control processes are likely to include synaptic facilitation and depression and distributed channel based adaptations. They should collectively achieve the effective adaptation in the gain such as the g_k in equation (6) and/or the underlying eigenmodes. Because there are multiple, redundant, and distributed synapses from the auditory periphery to the neuron whose STRF we model, a STRF could be implemented in multiple ways. Such implementational redundancy is likely to be needed to accommodate the many forms of adaptation that might be needed, given a limited degree of flexibility in any individual mechanism.

The timescale of STRF adaptation to sound levels or input SNRs should be less than several or tens of seconds, or even shorter, since, in the physiological experiments, the stimulus duration for one sound intensity level is 40s in[7] and 5s in[14], while adaptation to mixing noise into the vocalization inputs occurs within hundreds of milliseconds in [26]. Adaptation has been observed to occur over multiple time scales, ranging from tens of milliseonds to minutes in the fly visual system[67]. In the auditory systems, midbrain neurons adapt to sound levels within hundreds of milliseconds[68, 69], while cortical adaptation happens over multiple timescales and is likely to arise from network activities [70, 71] . We still know too little about the actual mechanisms for STRF adaptation[26] or sensory adaptation in general, although it has been suggested that channel based mechanisms at the cellular level are plausible candidates[67]. Understanding the computational roles of the STRFs should motivate future investigations of these mechanisms.

Limitations of the framework

As an initial attempt to understand the computational role of the STRFs, our framework has various limitations. First, the STRF model as a whole is quantitatively inaccurate since it specifies a linear mapping between sensory inputs and neural responses (in each adapted state). The accuracy could be improved in future work through the addition of a static nonlinearity after the STRF [6, 7]. However, this would not be expected to lead to a qualitative change in STRFs or their adaptation. Extensions to dynamic nonlinearities would be much more complex. Second, for analytical convenience, we assumed that the input statistics are Gaussian, meaning that there are no input signal correlations higher than second order. The same approximation was made for the case of efficient visual coding, in the absence of good information about higher order input correlations [30, 32, 34]. Subsequent work using independent component analysis (ICA) on natural visual images avoided the Gaussian assumption, leading to models of visual encoding in

primary visual cortex V1 [72, 73]. This approach has been adopted to understand the STRFs in the auditory cortex[74] and avian primary auditory area field L[75], although it cannot predict adaptation to SNR and its whitenning prediction does not go beyond that obtained under the Gaussian assumption. It is still controversial whether higher order statistics are the cause for the dramatic difference between the V1 encoding and that in the retina and the lateral geniculate nucleus[34]. Furthermore, higher order correlations in natural visual inputs contribute much less redundancy (measured in signal entropy) than second order correlations [36, 37, 38]. This may explain why the Gaussian assumption was not overly deleterious to the predictions of the efficient coding principle in vision. Although higher order correlations in auditory inputs are also poorly understood, they do cause auditory adaptation, e.g., in stimulus-specific adaptation to complex temporal patterns of tones[76]. To what extent higher order input statistics can influence auditory encoding remains to be answered in future studies.

Our focus on coding efficiency ignores aspects of auditory processing devoted to additional tasks such as sound source localization or stream segmentation. The observed STRFs may reflect elements of both efficient coding and requirements associated with these tasks. In fact, some variations are possible within the context of an efficient code. For instance, we have so far restricted ourselves by making all neurons share the same MTF profile predicted by efficient coding (by restricting the U transform to that in equation (9)). Relaxing this restriction would allow other STRFs. In particular, different neurons in the coding population could be tuned to different modulation frequency regions within the (Ω, ω) extent covered by the overall MTF envelope $g(\Omega, \omega)$, and could have different shapes. Accordingly, different STRFs could have different spectral bandwidths (or resolution) and shapes, in addition to preferring different center frequencies f. Indeed, in the auditory cortex, different neurons exhibit different spectral resolutions, and even prefer different motion directions of the spectral ripples[77, 78, 19]. (Analogously, primary visual cortical neurons are tuned to multiple spatial sizes and prefer different orientations, a coding scheme that can be shown to be consistent with efficient coding[36].) Such a collection of STRFs could satisfy the joint goals of coding efficiency and detecting ecologically meaningful auditory objects (such as vocalizations). Diversity in the shape and bandwidth of the STRFs is already present, although perhaps less so sub-cortically, e.g., in inferior colliculus[78]. When different neurons have different STRF bandwidths, our prediction that the input modulation power will be whitened by the neural MTFs should be modified, such that the 'neural MTFs' should mean the collective MTF of the whole neural population within a particular auditory stage (such as IC, see[59]).

There could be alternative formulations (other than equation (4)) of the efficient coding principle, in particular, in the formulation of the neural cost. Our formulation neural cost $= \sum_i \langle O_i^2 \rangle$ causes the degeneracy of the efficient coding solution, i.e., the existence of many choices of the equally efficient coding transforms, when the signals are gaussian. Other formulations of the neural cost could break this degeneracy. For example, formulation neural cost $= \sum_i H(O_i)$ in terms of the summation of individual neural channel capacity (or entropy $H(O_i)$), or neural cost $= \sum_i \langle |O_i| \rangle$ in terms of the total activity level, would generate neural codes to encourage very different MTFs for different neurons. In both audition and vision, the MTFs (in audition) and the contrast sensitivity functions (the vision analog of the MTFs) for different neurons tend to be similar

in the sensory periphery (cochlear nucleus and retina), but they are increasingly disparate further towards the central brain. These changes could be caused by the different cost functions in the nervous system, or, as discussed in the previous paragraph, due to the breaking of the degeneracy by additional computational tasks further downstream along the sensory pathway.

Redundancy reduction and information preservation are two essential ingredients of the efficient coding principle. While this principle has been quite successful in understanding the retinal coding, it cannot explain the enormous increase in the redundancy of the visual coding in the primary visual cortex (in which the number of neurons are about 100 times as many as those in the retina)[34], nor the drastic loss of visual information outside the focus of attention in the higher visual areas without introducing task-dependent factors. It remains to be investigated how much and in what form the efficient coding will take further along the auditory pathway. One can expect that more processes will be devoted to solving specific auditory tasks, in addition to the task of sensory encoding, in the higher stages of auditory processing.

Concluding remarks

This study was partly inspired by the success of the efficient coding principle in understanding receptive fields in the early stages of visual processing, and the way these receptive fields adapt across sensory environments. Analogies between visual and auditory processes have been explored by previous researchers [79], and we expect that they can be carried further in higher level sensory processes including segmentation, selective attention [80], and even object recognition.

In conclusion, efficient coding provides a plausible computational interpretation of various recent experimental observations on STRFs, and notably the way they adapt to input environments. By making testable predictions, it motivates experimental directions which should hopefully lead to further insights and understanding.

Acknowledgement:

We are very grateful to Nick Lesica for providing us with the STRF data of 40 inferior colliculus neurons [7], from which we obtained the physiological MTF plots in Figure 7. We would also like to thank very much Dr. Bo Hong and three anonymous reviewers for their very helpful comments, and to thank very much Peter Dayan for editing the English of the manuscript. This study is supported in part by the Gatsby Charitable Foundation, Tsinghua University 985 fund, and National Science Foundation of China grant 60675029.

Author Contributions:

Began the project as a course project: L. Zhao. Supervised and completed the project: L. Zhaoping.

References

[1] Aertsen AM, Johannesma PI (1981) The spectro-temporal receptive field. A functional characteristic of auditory neurons. Biol Cybern 42: 133-43.

[2] Escabi MA, Schreiner CE (2002) Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. J Neurosci 22: 4114-4131.

[3] Klein DJ, Depireux DA, Simon JZ, Shamma SA (2000) Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design. J Comput Neurosci 9: 85-111.

[4] Theunissen FE, Sen K, Doupe AJ (2000) Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. J Neurosci 20: 2315-31.

[5] Eggermont JJ, Johannesma PIM, Aertsen AMHJ (1983) Reverse-correlation methods in auditory research. Q Rev Biophys 16: 341–414.

[6] Eggermont JJ, Aertsen AMHJ, Johannesma PIM (1983a) Quantitative characterisation procedure for auditory neurons based on the spectro-temporal receptive field. Hearing Res 10: 167–190.

[7] Lesica NA, Grothe B (2008) Dynamic spectrotemporal feature selectivity in the auditory midbrain. J Neurosci 28: 5412-5421.

[8] Eggermont JJ, Aertsen AMHJ, Johannesma PIM (1983b) Prediction of the responses of auditory neurons in the midbrain of the grass frog based on the spectro-temporal receptive field. Hearing Res 10: 191-202.

[9] Christianson GB, Sahani M, Linden JF (2008) The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields. J Neurosci 28: 446-455.

[10] Gourevitch B, Norena A, Shaw G, Eggermont JJ (2009) Spectrotemporal receptive fields in anesthetized cat primary auditory cortex are context dependent. Cereb Cortex 19: 1448-1461.

[11] Woolley SMN, Gill PR, Theunissen FE (2006) Stimulus-dependent auditory tuning results in synchronous population coding of vocalizations in the songbird midbrain. J Neurosci 26: 2499-2512.

[12] Yu JJ, Young ED (2000) Linear and nonlinear pathways of spectral information transmission in the cochlear nucleus. P Natl Acad Sci U S A 97: 11780-11786.

[13] Young ED, Oertel D (2003) The cochlear nucleus. In: Shepherd G, editor, Synaptic Organization of the Brain, Oxford Press, chapter 4. 5 edition, pp. 125-164.

[14] Nagel KI, Doupe AJ (2008) Organizing principles of spectro-temporal encoding in the avian primary auditory area field L. Neuron 58: 938-955.

[15] Kim PJ, Young ED (1994) Comparative analysis of spectro-temporal receptive fields, reverse correlation functions, and frequency tuning curves of auditory-nerve fibers. J Acoust Soc Am 95: 410-422.

[16] Versnel H, Zwiers MP, van Opstal AJ (2009) Spectrotemporal response properties of inferior colliculus neurons in alert monkey. J Neurosci 29: 9725-9739.

[17] Shamma SA, Versnel H (1995) Ripple analysis in ferret primary auditory cortex. II. Prediction of unit responses to arbitrary spectral profiles. Audit Neurosci 1: 255–270.

[18] Kowalski N, Depireux DA, Shamma SA (1996) Analysis of dynamic spectra in ferret primary auditory cortex. II. Prediction of unit responses to arbitrary dynamic spectra. J Neurophysiol 76: 3524-3534.

[19] Depireux DA, Simon JZ, Klein DJ, Shamma SA (2001) Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J Neurophysiol 85: 1220-1234.

[20] Schnupp JWH, Mrsic-Flogel TD, King AJ (2001) Linear processing of spatial cues in primary auditory cortex. Nature 414: 200–204.

[21] Nelken I, Bar-Yosef O (2008) Neurons and objects: the case of auditory cortex. Front Neurosci 2: 107-113.

[22] Barbour DL, Wang X (2003) Contrast tuning in auditory cortex. Science 299: 1073-1075.

[23] Ahrens MB, Linden JF, Sahani M (2008) Nonlinearities and contextual influences in auditory cortical responses modeled with multilinear spectrotemporal methods. J Neurosci 28: 1929-1942.

[24] Lewicki MS (2002) Efficient coding of natural sounds. Nat Neurosci 5: 356-363.

[25] Smith EC, Lewicki MS (2006) Efficient auditory coding. Nature 439: 978-982.

[26] Lesica NA, Grothe B (2008) Efficient temporal processing of naturalistic sounds. PLoS One 3: e1655.

[27] Barlow HB (1961) Possible principles underlying the transformation of sensory messages. In: Rosenblith WA, editor, Sensory Communication, Cambridge, MA: MIT Press. pp. 217-234.

[28] Laughlin S (1981) A simple coding procedure enhances a neuron's information capacity. Z Naturforsch C 36: 910–912.

[29] Srinivasan MV, Laughlin SB, Dubs A (1982) Predictive coding: a fresh view of inhibition in the retina. P Roy Soc Lond B Bio 216: 427–459.

[30] Linsker R (1990) Perceptual neural organization: some approaches based on network models and information theory. Annu Rev Neurosci 13: 257–281.

[31] Atick JJ, Redlich AN (1990) Towards a theory of early visual processing. Neural Comput 2: 308–320.

[32] Atick JJ (1992) Could information theory provide an ecological theory of sensory processing? Network-Comp Neural 3: 213–251.

[33] van Hateren JH (1992) A theory of maximizing sensory information. Biol Cybern 68: 23-9.

[34] Zhaoping L (2006) Theoretical understanding of the early visual processes by data compression and data selection. Network-Comp Neural 17: 301-334.

[35] Nelken I, Rotman Y, Yosef OB (1999) Responses of auditory-cortex neurons to structural features of natural sounds. Nature 397: 154–157.

[36] Li Z, Atick JJ (1994) Toward a theory of the striate cortex. Neural Comput 6: 127–146.

[37] Petrov Y, Zhaoping L (2003) Local correlations, information redundancy, and sufficient pixel depth in natural images. J Opt Soc Am A 20: 56–66.

[38] Hosseini R, Sinz F, Bethge M (2010) Lower bounds on the redundancy of natural images. Vision Res 50: 2213-2222.

[39] Field DJ (1987) Relations between the statistics of natural images and the response properties of cortical cells. J Opt Soc Am A 4: 2379–2394.

[40] Kersten D (1987) Predictability and redundancy of natural images. J Opt Soc Am A 4: 2395–2400.

[41] Ruderman DL, Bialek W (1994) Statistics of natural images: Scaling in the woods. Phys Rev Lett 73: 814–817.

[42] Reinagel P, Zador AM (1999) Natural scene statistics at the centre of gaze. Network-Comp Neural 10: 341–350.

[43] Daugman JG (1989) Entropy reduction and decorrelation in visual coding by oriented neural receptive fields. IEEE T Bio-Med Eng 36: 107–114.

[44] Atick JJ, Li Z, Redlich AN (1992) Understanding retinal color coding from first principles. Neural Comput 4: 559–572.

[45] Atick JJ, Li Z, Redlich AN (1993) What does post-adaptation color appearance reveal about cortical color representation? Vision Res 33: 123–129.

[46] Li Z, Atick JJ (1994) Efficient stereo coding in the multiscale representation. Network-Comp Neural 5: 157–174.

[47] Li Z (1995) Understanding ocular dominance development from binocular input statistics. In: Bower J, editor, Proceeding of Computational Neuroscience Conference. Monterey, California: Kluwer Academic Publishers, pp. 397-402.

[48] Chechik G, Anderson MJ, Bar-Yosef O, Young ED, Tishby N, et al. (2006) Reduction of information redundancy in the ascending auditory pathway. Neuron 51: 359-68.

[49] Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27: 379-423.

[50] Levy WB, Baxter RA (1996) Energy efficient neural codes. Neural Comput 8: 531–543.

[51] Atick JJ, Redlich AN (1992) What does the retina know about natural scenes? Neural Comput 4: 196–210.

[52] Barlow HB, Fitzhugh R, Kuffler SW (1957) Change of organization in the receptive fields of the cat's retina during dark adaptation. J Physiol-London 137: 338-354.

[53] Glasberg BR, Moore BCJ (1990) Derivation of auditory filter shapes from notched-noise data. Hearing Res 47: 103–138.

[54] Escabi MA, Miller LM, Read HL, Schreiner CE (2003) Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. J Neurosci 23: 11489-11504.

[55] Gill P, Zhang J, Woolley SMN, Fremouw T, Theunissen FE (2006) Sound representation methods for spectro-temporal receptive field estimation. J Comput Neurosci 21: 5-20.

[56] Young ED, Calhoun BM (2005) Nonlinear modeling of auditory-nerve rate responses to wideband stimuli. J Neurophysiol 94: 4441-4454.

[57] Oppenheim AV, Willsky AS, Nawab SH (1997) Signals and systems. Prentice Hall, 2 edition.

[58] Nagel KI, Doupe AJ (2006) Temporal processing and adaptation in the songbird auditory forebrain. Neuron 51: 845-859.

[59] Rodriguez FA, Chen C, Read HL, Escabi MA (2010) Neural modulation tuning characteristics scale to efficiently encode natural sound statistics. J Neurosci 30: 15969-15980.

[60] Robinson BL, McAlpine D (2009) Gain control mechanisms in the auditory pathway. Curr Opin Neurobiol 19: 402–407.

[61] Laughlin SB, Hardie RC (1978) Common strategies for light adaptation in the peripheral visual systems of fly and dragonfly. J Comp Physiol A 128: 319-340.

[62] Rieke F (2001) Temporal contrast adaptation in salamander bipolar cells. J Neurosci 21: 9445-9454.

[63] Kim KJ, Rieke F (2001) Temporal contrast adaptation in the input and output signals of salamander retinal ganglion cells. J Neurosci 21: 287-299.

[64] Le Beau FE, Rees A, Malmierca MS (1996) Contribution of gaba-and glycine-mediated inhibition to the monaural temporal response properties of neurons in the inferior colliculus. J Neurophysiol 75: 902-919.

[65] Caspary DM, Palombi PS, Hughes LF (2002) Gabaergic inputs shape responses to amplitude modulated stimuli in the inferior colliculus. Hearing Res 168: 163–173.

[66] Guinan Jr JJ (2006) Olivocochlear efferents: anatomy, physiology, function, and the measurement of efferent effects in humans. Ear Hearing 27: 589-607.

[67] Wark B, Lundstrom BN, Fairhall A (2007) Sensory adaptation. Curr Opin Neurobiol 17: 423-429.

[68] Dean I, Robinson BL, Harper NS, McAlpine D (2008) Rapid neural adaptation to sound level statistics. J Neurosci 28: 6430-6438.

[69] Dean I, Harper NS, McAlpine D (2005) Neural population coding of sound level adapts to stimulus statistics. Nat Neurosci 8: 1684–1689.

[70] Ulanovsky N, Las L, Farkas D, Nelken I (2004) Multiple time scales of adaptation in auditory cortex neurons. J Neurosci 24: 10440-10453.

[71] Ulanovsky N, Las L, Nelken I (2003) Processing of low-probability sounds by cortical neurons. Nat Neurosci 6: 391–398.

[72] Olshausen BA, Field DJ (1997) Sparse coding with an overcomplete basis set: A strategy employed by v1? Vision Res 37: 3311–3325.

[73] Bell AJ, Sejnowski TJ (1997) The "independent components" of natural scenes are edge filters. Vision Res 37: 3327-3338.

[74] Klein DJ, König P, Körding KP (2003) Sparse spectrotemporal coding of sounds. EURASIP J Appl Sig P 7: 659–667.

[75] Greene G, Barrett DGT, Sen K, Houghton C (2009) Sparse coding of birdsong and receptive field structure in songbirds. Network-Comp Neural 20: 162–177.

[76] Nelken I (2004) Processing of complex stimuli and natural scenes in the auditory cortex. Curr Opin Neurobiol 14: 474–480.

[77] Wang K, Shamma SA (1995) Spectral shape analysis in the central auditory system. IEEE T Speech Audi P 3: 382–395.

[78] Schreiner CE, Read HL, Sutter ML (2000) Modular organization of frequency integration in primary auditory cortex. Annu Rev Neurosci 23: 501–529.

[79] Shamma SA (2001) On the role of space and time in auditory processing. Trends Cogn Sci 5: 340–348.

[80] Fritz JB, Elhilali M, David SV, Shamma SA (2007) Auditory attention-focusing the searchlight on sound. Curr Opin Neurobiol 17: 437–455.