Sustained Firing of Model Central Auditory Neurons Yields a Discriminative Spectro-temporal Representation for Natural Sounds

Michael A. Carlin, Mounya Elhilali*

Department of Electrical and Computer Engineering, The Center for Language and Speech Processing, Johns Hopkins University, Baltimore, Maryland, United States of America

Abstract

The processing characteristics of neurons in the central auditory system are directly shaped by and reflect the statistics of natural acoustic environments, but the principles that govern the relationship between natural sound ensembles and observed responses in neurophysiological studies remain unclear. In particular, accumulating evidence suggests the presence of a code based on sustained neural firing rates, where central auditory neurons exhibit strong, persistent responses to their preferred stimuli. Such a strategy can indicate the presence of ongoing sounds, is involved in parsing complex auditory scenes, and may play a role in matching neural dynamics to varying time scales in acoustic signals. In this paper, we describe a computational framework for exploring the influence of a code based on sustained firing rates on the shape of the spectro-temporal receptive field (STRF), a linear kernel that maps a spectro-temporal acoustic stimulus to the instantaneous firing rate of a central auditory neuron. We demonstrate the emergence of richly structured STRFs that capture the structure of natural sounds over a wide range of timescales, and show how the emergent ensembles resemble those commonly reported in physiological studies. Furthermore, we compare ensembles that optimize a sustained firing code with one that optimizes a sparse code, another widely considered coding strategy, and suggest how the resulting population responses are not mutually exclusive. Finally, we demonstrate how the emergent ensembles contour the high-energy spectro-temporal modulations of natural sounds, forming a discriminative representation that captures the full range of modulation statistics that characterize natural sound ensembles. These findings have direct implications for our understanding of how sensory systems encode the informative components of natural stimuli and potentially facilitate multi-sensory integration.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mounya@jhu.edu

Introduction

It is widely believed that sensory representations are optimized to process the stimuli to which they are exposed in natural environments [1]. Of particular interest is understanding the computational principles that underlie the generation of observed neural firing patterns. A popular hypothesis explored in recent years assumes that neural populations optimize a sparse code. This means that at any given time, only a small subset of a neural population fires to encode a given stimulus [2]. Such a representation is attractive for reasons of coding efficiency (see, e.g., [3]) and conservation of physiological resources [4]. The sparse coding hypothesis has enjoyed particular success in studies of vision (e.g., [5,6]), and has also been supported more recently by both neurophysiological [7,8] and computational studies [9–11] of the auditory system.

However, it has also been observed that some central auditory neurons, when driven by their preferred stimuli, exhibit *sustained* firing rates. Measuring from auditory thalamus and primary

auditory cortex, Wang *et al.* observed that sustained responses were not simply phase-locked to the fast dynamics of the stimulus, suggesting that this rate-based code represented a meaningful, non-isomorphic transformation of the stimulus [12,13]. Indeed, such a code is particularly important for audition since it directly addresses the issue of how to indicate the continued presence of a sound in a complex acoustic environment. Results from Petkov *et al.* have also illustrated how sustained responses play a role in auditory scene analysis, forming part of the neural basis for the perceptual restoration of foreground sounds against a cluttered background [14]. Moreover, Wang has argued that a rate-based representation is critical for matching fast temporal modulations present in natural sounds to slower rates found in higher cortical areas [15]. Slower dynamics in acoustic signals are believed to be the main carrier of information in speech and music [16]; are commensurate with temporal dynamics of stream formation and auditory grouping [17]; and may play an important role in multi-modal sensory integration [15]. Related computational studies in vision have suggested how this principle may underlie the shapes

Author Summary

We explore a fundamental question with regard to the representation of sound in the auditory system, namely: what are the coding strategies that underlie observed neurophysiological responses in central auditory areas? There has been debate in recent years as to whether neural ensembles explicitly minimize their propensity to fire (the so-called sparse coding hypothesis) or whether neurons exhibit strong, sustained firing rates when processing their preferred stimuli. Using computational modeling, we directly confront issues raised in this debate, and our results suggest that not only does a sustained firing strategy yield a sparse representation of sound, but the principle yields emergent neural ensembles that capture the rich structural variations present in natural stimuli. In particular, spectro-temporal receptive fields (STRFs) have been widely used to characterize the processing mechanisms of central auditory neurons and have revealed much about the nature of sound processing in central auditory areas. In our paper, we demonstrate how neurons that maximize a sustained firing objective yield STRFs akin to those commonly measured in physiological studies, capturing a wide range of aspects of natural sounds over a variety of timescales, suggesting that such a coding strategy underlies observed neural responses.

of simple and complex cell receptive fields in primary visual cortex [18,19]. Importantly, a sustained firing rate, i.e., one that is persistent and therefore slowly changing over time, is related to slow feature analysis, a well-known method for extracting invariances from sensory signals [20] (see Discussion). To the best of our knowledge, however, there are no computational studies that explicitly consider the implications of a sustained firing-based code in central auditory areas.

At first glance, the two coding schemes are seemingly at odds: on the one hand a sparse code seeks to minimize the activity of a neural population whereas a sustained firing-based code requires that neural responses persist over time but still form an efficient representation of the stimulus. However, it appears that central auditory responses can strike a balance between the two strategies, with a large, transient population response at the onset of a sound, and a sparse subset of preferentially driven neurons exhibiting a strong, sustained response throughout the sound's duration [15,21]. This picture suggests a mechanism for detecting and tracking target sounds in noisy acoustic environments and for generating a persistent signal that facilitates a stable perceptual representation. From a computational perspective, a better understanding of these mechanisms can inform models of auditory scene analysis as well as signal processing schemes for hearing prosthetics and automated sound processing systems.

A general computational approach for exploring the effects of particular coding strategies in sensory systems is based on optimizing a statistical objective criterion that quantifies the principle governing the transformation between stimulus and internal representation. Upon convergence, one then compares the emergent representation to known properties of the sensory system being studied [1]. Here, we apply this framework to explore how optimizing a sustained firing criterion influences the shapes of model auditory spectro-temporal receptive fields (STRFs) when processing natural sounds, and we compare the emergent ensembles to those obtained by optimizing a sparse coding objective. STRFs describe the linear mapping between a spectro-temporal stimulus and an instantaneous firing rate [22], and have

proven useful not only for describing basic processing aspects of auditory neurons [23,24], but also for shedding light on the nature of task-driven plasticity [25]. Figure 1 illustrates how a spectro-temporal stimulus is mapped to a set of instantaneous neural firing rates, whose ensemble response according to a desired coding strategy directly shapes the mapping.

In this paper, we show how this framework allows us to not only explore how the timescales of natural sounds are captured by and reflected in an emergent sensory representation, but reveal key similarities between choice of a sustained versus sparse code. Moreover, we demonstrate how a sustained firing-based code suggests a mechanism for an emergent discriminative representation for ensembles of natural stimuli.

Results

We defined a sustained neural response as one where firing rate energy changes relatively slowly and is consequently highly *correlated* over time. In particular, we were interested in the characteristics of ensembles of model STRFs $H = [\mathbf{h}_1\,\mathbf{h}_2\,\mathbf{h}_K]$ that promoted sustained responses over a specified time interval $[t-\Delta T, t]$. Denoting the response of the k'th neuron as $r_k(t) = \mathbf{h}_k^T\mathbf{s}(t)$, where $\mathbf{h}_k$ is the STRF and $\mathbf{s}(t)$ is a spectro-temporal stimulus, we quantified this principle using the following objective function:

$$J_{sus}(H) := \sum_{k=1}^{K}\int_{\Delta T}\alpha_\tau\langle r_k^2(t)r_k^2(t-\tau)\rangle_t\,d\tau, \qquad (1)$$

where $\langle\cdot\rangle_t$ denotes time average. Observe that $J_{sus}(H)$ represents the sum of correlations between signal energies of the k'th neuron over a time interval defined by ΔT across an ensemble of K neurons. If a neuron yielded a sustained response, then each of the $r_k(t)$ would vary smoothly over the specified interval and we expect $J_{sus}(H)$ to be large. Moreover, choice of the correlation interval ΔT allowed us to directly explore the effect of different timescales on the ensembles H that *maximized* Eq. 1. Finally, the weights α_τ were chosen to be linearly decaying for $\tau = 0$ to ΔT, reflecting the intuition that recent activity of a neuron likely has more influence on the current output than the past. Note that these weights could be adapted to specifically model, for example, positive- or negative-monotonic sustained responses observed in physiological studies [13]. Full details of the optimization procedure can be found in Methods.

Alternatively, we explored an objective function that promoted sparsity. A natural way to induce sparsity in a population code is by enforcing a population response whose firing rate distribution is highly peaked near zero (representing frequent *weak* responses), but has long tails (representing infrequent *large* responses), i.e., a distribution with high *kurtosis* [26]. We quantified the sparsity of a population code using sample kurtosis:

$$J_{sp}(H) = \left\langle\frac{\mu_4(t)}{[\sigma^2(t)]^2}\right\rangle_t \qquad (2)$$

where $\mu_4(t) = \frac{1}{K}\sum_k (r_k(t) - \bar{r}(t))^4$ is the fourth central moment at time t, $\sigma^2(t) = \frac{1}{K}\sum_k (r_k(t) - \bar{r}(t))^2$ is the population variance at time t, and $\bar{r}(t)$ is the population mean at time t.

For both $J_{sus}(H)$ and $J_{sp}(H)$, the basic problem was to find an ensemble of STRFs that *maximized* the respective objective function subject to constraints that (1) bounded the amplitude of the filter responses and (2) minimized redundancy among the learned
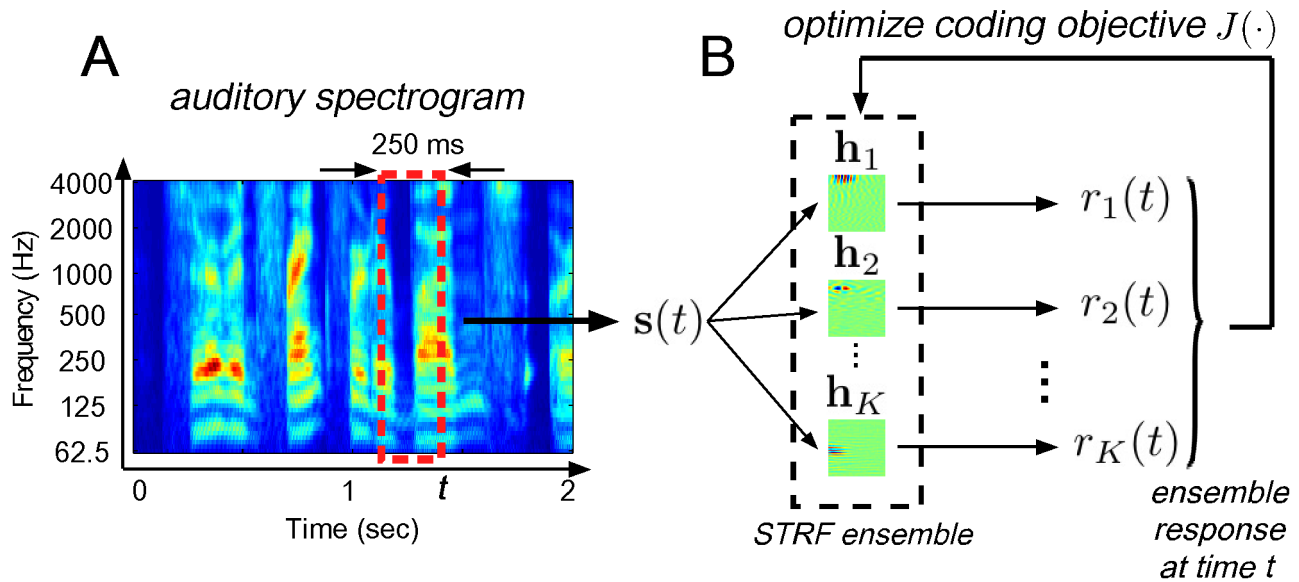
Figure 1. Schematic of the proposed framework. Panel (A) shows an example of an auditory spectrogram for the speech utterance "serve on frankfurter buns…" whereas panel (B) illustrates how spectro-temporal patches are mapped to an ensemble of instantaneous neural firing rates.
doi:10.1371/journal.pcbi.1002982.g001

ensemble. This was achieved by enforcing the responses have unit variance and be mutually uncorrelated, i.e., $\langle r_j(t)r_k(t)\rangle_t = \delta_{jk}$ where δ_{jk} is the Kroenecker delta function (see Methods); we refer to these as *response* constraints. These constraints ensured that the responses had a bounded magnitude and that the STRFs did not all converge to the same solution.

Emergence of richly structured STRFs

We optimized both the sustained objective $J_{sus}(H)$ and sparsity objective $J_{sp}(H)$ using an ensemble of natural stimuli comprising speech, animal vocalizations, and ambient outdoor sounds. Each ensemble of $K = 400$ filters was initialized at random using zero-mean, unit variance Gaussian noise, and each STRF covered from 0–250 ms in time and 62.5–4000 Hz along the tonotopic axis.

For the sustained objective, we considered a wide range of correlation intervals from very brief ($\Delta T = 10$ ms) to very long ($\Delta T = 2000$ ms). Examples of emergent STRFs for $\Delta T = 125$ ms are shown in Figure 2A. For the spectro-temporal patches shown, red and blue colors indicate that the presence of energy in a particular spectro-temporal region yields excitatory and inhibitory responses, respectively. We observe a variety of STRFs that are highly localized, sensitive to narrowband spectral and temporal events, oriented, and some that are seemingly noise-like and not convergent to any particularly interesting shape. Importantly, such observations about these basic STRF classes align with those made in a number of previous physiological studies (see, e.g., [23,24,27]). Moreover, coverage of the STRFs appears to span the full time-frequency space. These results suggest that the sustained firing objective may underlie part of the coding strategy used by central auditory neurons.
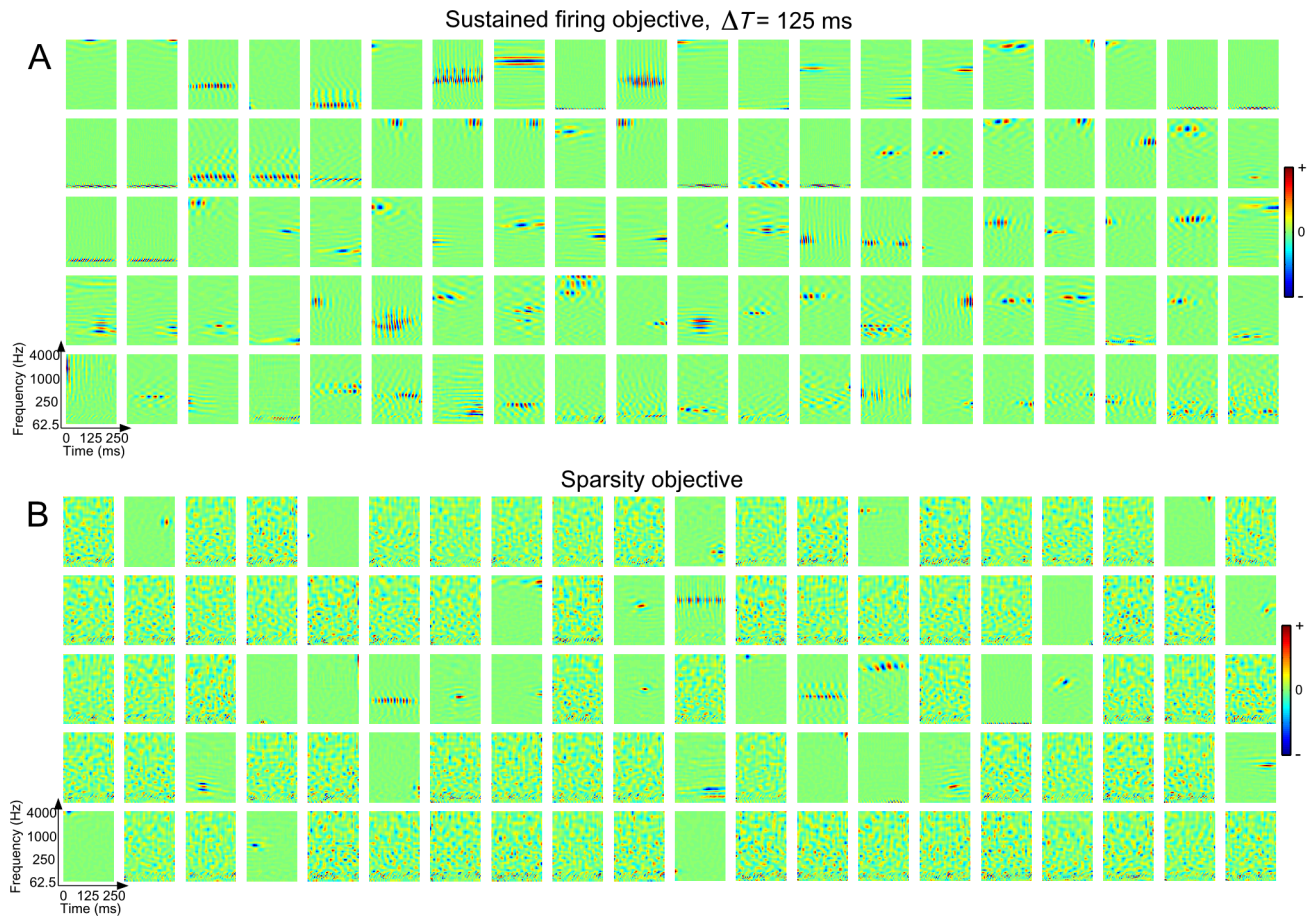
Shown in Figure 2B are examples of emergent STRFs obtained by optimizing the sparsity objective. Indeed, this particular objective yields STRFs that are highly localized and sparsely distributed, with sensitivity to bandlimited spectral and temporal events. While both objective criteria yield noisy STRFs, it is clear that the sparse ensemble is much more noisy, with a less extensive coverage of the basic sound classes as observed with the sustained ensemble.

Ensemble diversity varies smoothly with ΔT

Since the information-bearing components of natural sounds vary concurrently across multiple timescales, it was expected that the structure of STRFs learned under the sustained objective would vary with the correlation interval ΔT. Indeed, inspection of the sustained ensembles for a range of ΔT suggested the presence of a number of latent classes whose membership varied smoothly from short to long correlation intervals. To quantify variations in population diversity over ecologically relevant timescales, we performed unsupervised clustering of the emergent STRFs and studied how class membership changed with objective function and correlation interval.

We pooled STRFs from the sparse ensemble and from the sustained ensembles for $\Delta T = 10, 25, 50, 125, 250, 500, 1000,$ and 2000 ms, yielding a total of 3600 STRFs. We then applied normalized spectral clustering to discover latent classes among the pooled STRFs. In general, spectral clustering algorithms require an affinity matrix that specifies pairwise similarities between the objects being clustered. Viewing this affinity matrix as an undirected graph, spectral clustering finds a partition of the graph into groups whose elements have common similarity with one another. A natural measure of similarity between STRFs can be derived from the two-dimensional cross-correlation between pairs of spectro-temporal patches. Such a measure is similar to that considered by Woolley *et al.* [28] and is desirable since it does not depend on subjective choice of spectro-temporal features to use for clustering. In this work, we defined the measure of similarity between pairs of STRFs as the *absolute value* of the maximum value of the two-dimensional cross-correlation matrix; we used absolute value since we wished to group similar STRFs regardless of whether they were excitatory or inhibitory. Furthermore, as the STRFs tended to be distributed with a variety of phases in the input space, we considered cross-correlations for arbitrary time-frequency shifts (see Methods for details).

Results obtained using normalized spectral clustering of the emergent ensembles into nine classes are shown in Figure 3. In the center panel of the figure, a stacked bar chart illustrates the the percentage of STRFs at a particular ΔT assigned to one of nine

Figure 2. Examples of emergent STRFs. Shown are STRFs learned by optimizing (A) the sustained objective function $J_{sus}(H)$ for $\Delta T = 125$ ms and (B) the sparsity objective function $J_{sp}(H)$. The examples shown here were drawn at random from ensembles of 400 neurons. The sustained STRFs are shown in order of decreasing contribution to the overall objective function whereas the sparse STRFs are shown randomly ordered. Each spectro-temporal patch spans 0–250 ms in time and 62.5–4000 Hz in frequency. For these examples the dynamic range of the STRFs was compressed using a $\sinh(\cdot)$ nonlinearity.
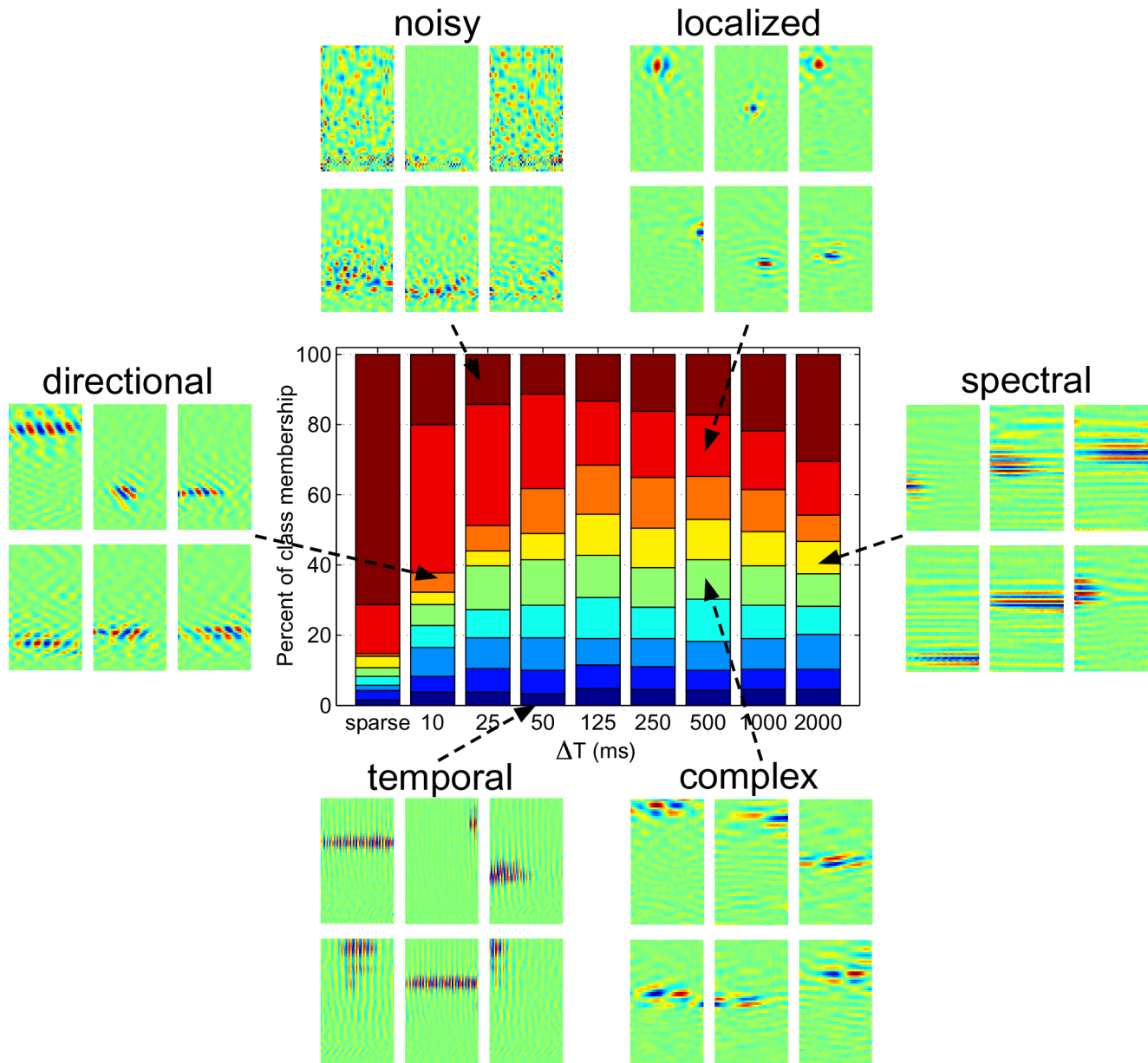doi:10.1371/journal.pcbi.1002982.g002

classes. Different segment colors correspond to each of the nine classes, and segment width is proportional to the number of STRFs assigned to that class. Surrounding the bar chart are examples from six classes that best illustrate how diversity varies with ΔT, namely *noisy*, *localized*, *spectral*, *complex*, *temporal*, and *directional* classes. These labels are qualitative descriptors of each class and not quantitative assessments of the time-frequency characteristics of each category.

Inspection of the cluster groupings reveal rich structural variations over a wide range of correlation intervals. In particular, the STRFs labeled according to the *noisy* class are found to dominate the sparse ensemble, with a large presence in the sustained ensemble for $\Delta T = 10$ ms. Membership in this class drops for ΔT between 10 and 125 ms, and begins to increase at 125 ms. We also observe that short correlation intervals ($\Delta T = 10$, 25, and 50 ms) have a large concentration of *localized* STRFs, with membership dropping with increasing ΔT. While the *temporal* class holds relatively steady across the sustained ensembles, we find that membership in the *directional*, *complex*, and *spectral* classes varied smoothly across ΔT. In general, we find that ensemble diversity is maximized for $\Delta T = 125$ ms (max. entropy of 3.08 bits), but the overall trends suggest rich ensemble structure between 10 and 250 ms, which is notably in the range of the timescales of natural sounds [29,30]. This is further supported by the increasing

presence of *noisy* STRFs for large correlation intervals ($\Delta T = 1000$ and 2000 ms).

In addition to studying structural variations in the *shapes* of the emergent STRFs, it is also of interest to examine the structure of the STRF *outputs* in response to natural sounds. In particular, we sought to address the extent to which enforcing sustained responses does indeed yield responses that persist over time. We defined the k'th neuron to be significantly "active" when its firing rate $r_k(t)$ exceeded ± 1 standard deviation over time. While this is not meant to be a precise measure of a neuron's activation (since, for instance, the firing rate is not used to modulate a Poisson spike generation process), such a measure nevertheless quantifies and characterizes a strong versus weak ensemble response to natural stimuli.

Shown in Figure 4A are the distribution of activation times for individual neurons for ensembles of $\Delta T = 10$ and 125 ms in response to a held-out set of natural stimuli. The neurons are shown sorted according to decreasing median activation time, and the interquartile ranges of activation time are indicated by the shaded regions. We observed that the most diversity in median activation times across ensembles occurred in approximately the top 10% of the *most persistent* neurons. To summarize these observations, we considered the distribution of median activation times of the top 10% of neurons with most persistent responses

Figure 3. Spectral clustering results. Shown are nine clusters obtained by pooling STRFs from the sparse as well as sustained ensembles for $\Delta T = 10$, 25, 50, 125, 250, 500, 1000, and 2500 ms. Shown in the center is a stacked bar chart where segment color corresponds to class label and segment width is proportional to the number of STRFs assigned to a particular class in a given ensemble. The surrounding panels show examples of STRFs drawn from six illustrative classes, namely, *noisy*, *localized*, *spectral*, *complex*, *temporal*, and *directional*.
doi:10.1371/journal.pcbi.1002982.g003

(i.e., the top 40 neurons); these distributions are illustrated as boxplots in Figure 4B.

As noted previously with the clustering results, shorter ΔT values favor mostly localized and noisy STRFs and consequently it was expected that activations would be brief. Interestingly, however, we observe that with increasing ΔT, median activations peak between 50 and 500 ms and fall off for large ΔT despite the STRFs being optimized to promote sustained responses over long intervals. This overall trend aligns with the previous clustering results that demonstrate how population diversity is maximized over intervals corresponding to timescales that predominate natural stimuli. The STRFs corresponding to the top 10% most persistent responses for $\Delta T = 125$ ms are shown in Supplementary

Figure 1, and we find that they generally have a spectral tuning, but are fairly narrowband and localized.

Additionally, we considered the responses of the top 40 most persistent responses obtained using the sparsity objective function; the distribution of median activations is in the first column of Figure 4B. We find that the sparse ensemble yields responses most similar to those for short ΔT.

Comparison of emergent sustained ensembles to physiology

How do the emergent STRFs learned under the sustained firing objective compare to those observed in physiological studies? Broadly speaking, we find that the emergent STRFs share many of
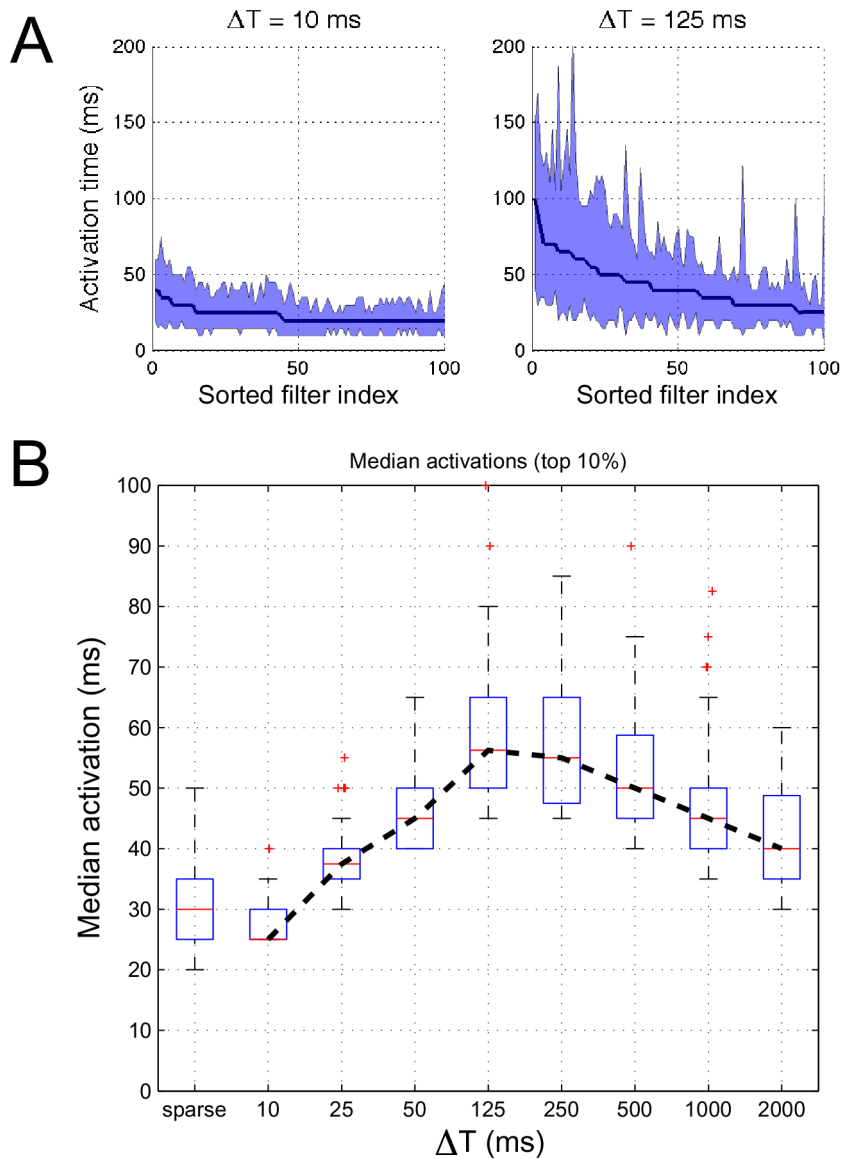
Figure 4. Analysis of the temporal activations of emergent ensembles. Panel (A) shows the median activation time of individual neurons (solid lines, sorted in decreasing order) for $\Delta T = 10$ and 125 ms, respectively, for STRFs that optimize the sustained objective function. The shaded region illustrates the corresponding interquartile range. Panel (B) shows the distributions (as boxplots) of median activation times of the top 10% "most persistent" neurons for sparse and sustained ensembles for increasing ΔT.
doi:10.1371/journal.pcbi.1002982.g004

the trends with biological receptive fields typically observed in animal models. We explored this issue by comparing our model ensembles with a set of 1586 STRFs recorded from awake, non-behaving ferret primary auditory cortex using TORC [31] and speech stimuli [27,32] (see Methods for more details). Where applicable, we also compared our results with reported results from anesthetized ferrets by Depireux *et al.* [23] and cats by Miller *et al.* [24] in the literature.

Illustrative examples of the types of STRFs found in the neural data are shown in Figure 5. In particular, we find neural STRFs that are qualitatively similar those found in the *localized*, *complex*, *noisy*, and *directional* clusters shown earlier in Figure 3. Because the temporal and spectral sampling rates used in our model are higher than those used in the physiological data, we did not find good matches with the *temporal* and *spectral* classes.

To visualize the overlap between the spectro-temporal modulation coverage of the neural and model STRFs, we used the ensemble modulation transfer function (eMTF). The eMTF is derived by averaging the magnitude of the 2D Fourier Transform of each neuron in a given ensemble, and jointly characterizes modulations in time (rate, in Hz) and in frequency (scale, in cyc/oct). We first applied normalized spectral clustering to the neural STRFs to obtain nine clusters. Next, we computed the eMTF for each cluster, extracted isoline contours at the 65% level, and overlaid these curves on the eMTF of the model STRFs for $\Delta T = 125$ ms. These results are shown in Figure 6 and illustrate the overlap between the model and neural data, particularly at the "edges" of the neural STRF modulations. While the overlap is not complete, it is clear that the modulation spectra of each ensemble are not disjoint. Moreover, the model eMTF suggests a general
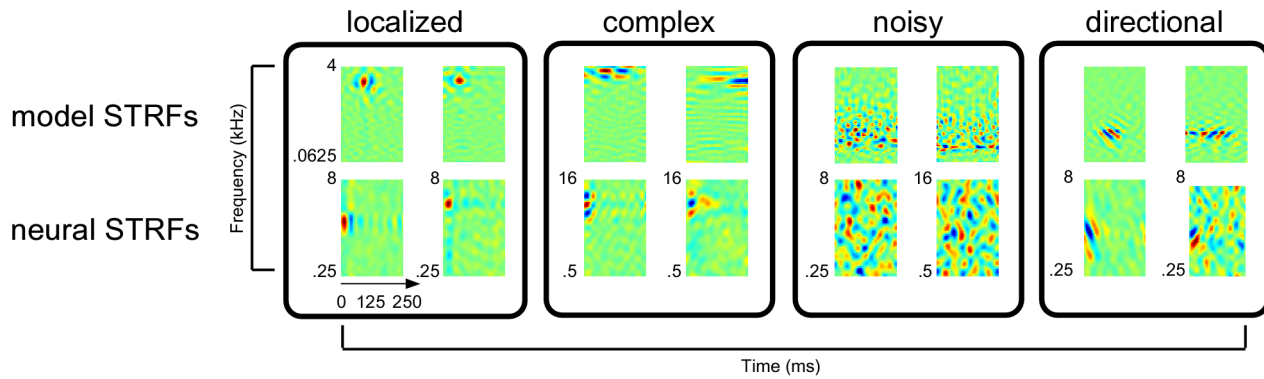
Figure 5. Comparison of emergent STRFs learned according to the sustained objective function with examples estimated from ferret auditory cortex.
doi:10.1371/journal.pcbi.1002982.g005

ensemble sensitivity to relatively fast modulations; this point is explored further in a later section ("Emergent STRFs capture spectro-temporal modulation statistics of stimulus").

To better characterize the relationship between the neural and model data, we employed a statistical comparison of the distribution of the two datasets. If the models truly generated STRFs similar to those in physiological studies, then one might expect a nearest-neighbor (NN) similarity distribution akin to one derived from the neural ensemble we considered. We computed the symmetric KL-divergence between each of the model and within-physiology NN similarity distributions (shown in Supplemental Figure 2). We found that the sustained-response (presented here) and sustained-shape (presented later in this paper) distributions had KL divergences of 0.80 and 0.85, respectively, whereas the sparse distribution had a KL distance of 1.05. KL typically measures the expected number of bits required to code samples from one distribution using codes from the other. While these

numbers are difficult to assess in absolute terms, they give a sense of how the different model optimizations and constraints compare to each other. These numbers reveal that the sustained ensembles are similarly comparable to the physiology, whereas the sparse ensemble has a somewhat worse match. Of course, caution must be taken with these numbers because the set of neural STRFs we analyzed represent only a subset of mappings that likely exist in central auditory areas.

Next, we measured a variety of parameters from the neural and model STRFs (for $\Delta T = 125$ ms) that more fully characterized the extent of spectro-temporal coverage and modulation sensitivity of the ensembles (see Methods), the results of which are summarized in Figure 7.

Based on the distribution of directionality indices, shown in panel (A), we observe that the model STRFs are largely symmetric, with the majority of neurons having no preference for upward or downward moving input stimuli (mean ≈ 0). As indicated by the
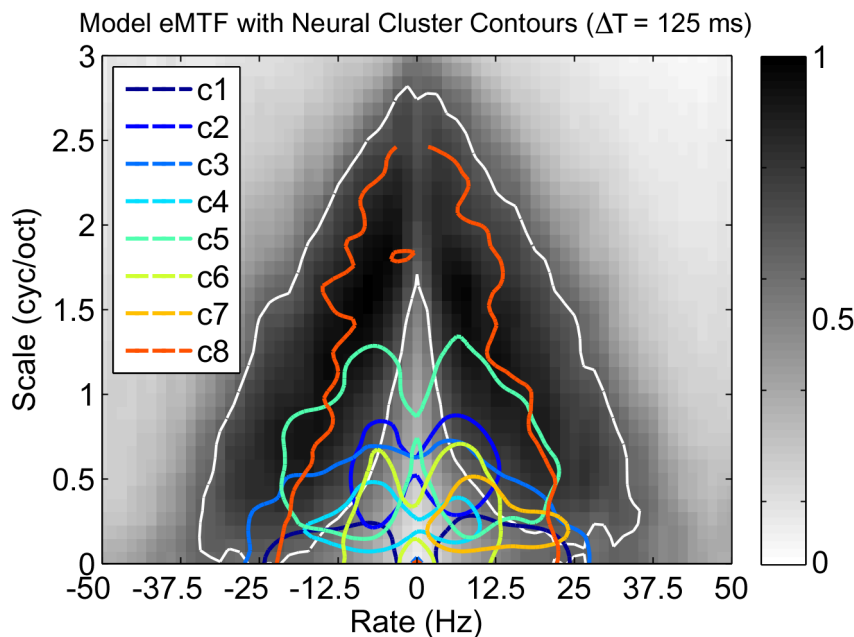
Figure 6. Cluster analysis of neural STRFs. Illustration of the overlap between the eMTFs of neural STRF clusters and that of the response-constrained sustained objective model STRFs; class 9 comprised mostly noisy STRFs with an exceedingly broad eMTF and its contour is omitted here for clarity. The white contour corresponds to the model eMTF at the 65% level.
doi:10.1371/journal.pcbi.1002982.g006
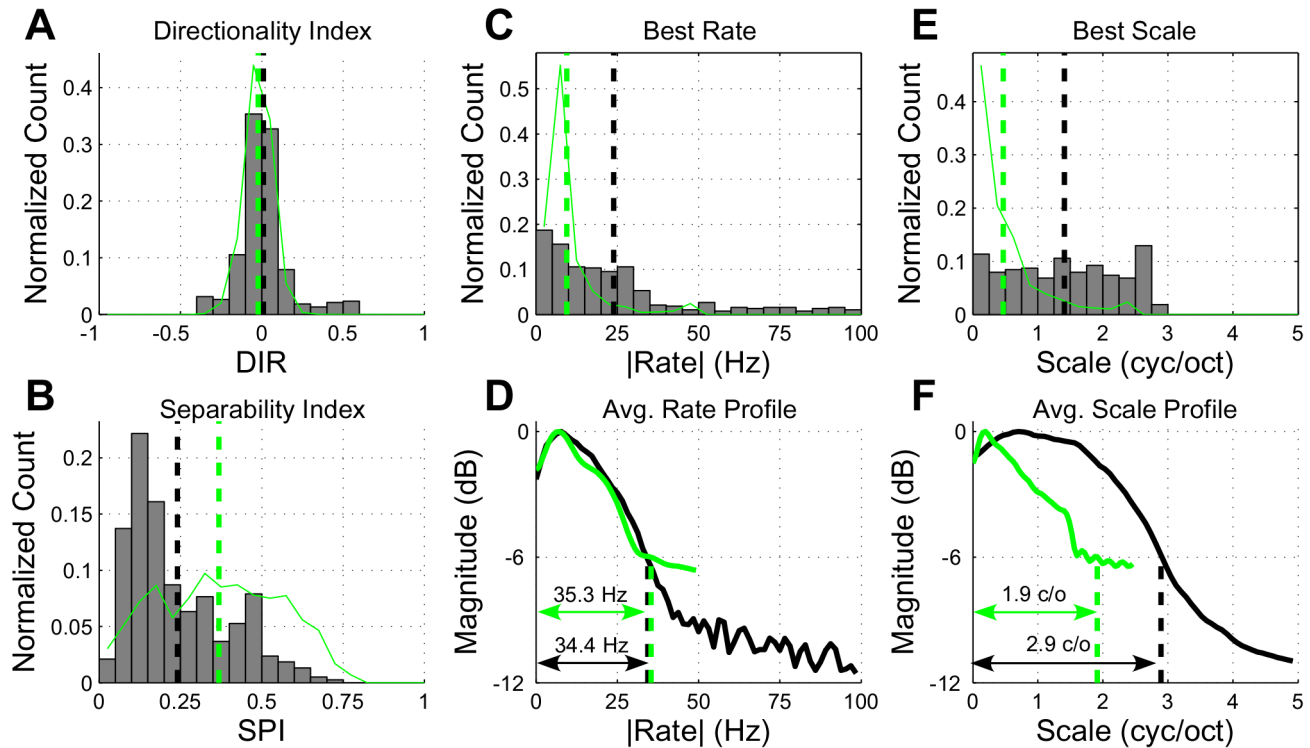
Figure 7. Ensemble analysis of STRFs learned under the sustained objective function for $\Delta T = 125$ ms. In panels (A), (B), (C) and (E), the histograms show the distribution of model parameters whereas the thin green lines show the distribution of the physiological data. The black and green dashed vertical lines show population means for the model and neural data, respectively. In panels (D) and (F), the black and green lines correspond to the model and neural STRFs, respectively, with the dashed lines indicating 6-dB upper cutoff frequencies. Refer to the text for more details.

doi:10.1371/journal.pcbi.1002982.g007

tails of this distribution, however, a subset of neurons have a strong directional preference. This agrees with the neural STRFs, and similar observations have been made in MGB and primary auditory cortex of cats by Miller *et al.*, as well as in measurements by Depireux *et al.* from primary auditory cortex of ferrets. Furthermore, panel (B) illustrates that a large number of model STRFs are fairly separable, with a peak in the separability index (SPI) distribution around 0.10 and an average value of 0.26. This trend aligns with values reported in the literature by Depireux *et al.* in measurements from ferret auditory cortex (mean of approx. 0.25). However, it is worth noting that this low level of separability is not uniformly reported across physiological studies of receptive field of mammalian auditory cortex. For instance, the physiological data analyzed in the current study (examples of which are shown in Figure 5) do yield a higher average SPI (mean = 0.37).

The temporal modulation statistics of the model STRFs, as quantified by best rate (BR), also align generally with results reported from mammalian thalamus and cortex. In panel (C) we observe a broad, bandpass distribution of best rates, with an average of 23.9 Hz. Reported physiological results from Miller *et al.* show similarly broad ranges of temporal tuning with preferences around 16 Hz and 30 Hz range for cortex and thalamus, respectively. The neural STRFs we analyzed show a somewhat slower tuning, with an average BR of 9.5 Hz. Furthermore, in panel (D), we computed the normalized average rate profile from the model STRFs. We observe a peak at 7.8 Hz, with an upper 6-dB cutoff of 34.4 Hz. Here we find a close overlap with the rate profile computed from the neural STRFs as well as with average profile results as reported by Miller *et al.* (peak at 12.8 Hz; upper 6-dB cutoff at 37.4 Hz).

The spectral modulation statistics of the model STRFs, as quantified by best scale, are generally faster than those reported from studies of thalamic and cortical nuclei. The distribution of best scales shown in panel (E) is bandpass with a wide range of slow to fast spectral coverage, with an average tuning of 1.40 cyc/oct. The neural STRFs, in contrast, are tuned to much slower scales (mean = 0.47 cyc/oct). Similarly, results from Miller *et al.* in MGB indicate a generally slower tuning (0.58 cyc/oct), whereas measurements from cortical neurons, while having a similarly wide range of tunings as with the model, indicate a slower average value of 0.46 cyc/oct and an upper cutoff of approx. 2 cyc/oct.

Finally, the ensemble average scale profile, shown in panel (F), is bandpass and exhibits a peak at 0.7 cyc/oct with an upper 6-dB cutoff of 2.9 cyc/oct. The neural STRFs, however, are much slower with peak at 0.2 cyc/oct and an upper cutoff of 1.9 cyc/oct. This is similar to observations from MGB by Miller *et al.*, where they reported that the ensemble average scale profile is generally low-pass, with average scale profile peaks and upper 6-dB cutoffs at 0 cyc/oct and 1.3 cyc/oct, respectively, with similar observations in cortex.

In summary, while we cannot map the emergent STRFs to any exact synapse, they nevertheless reflect the general processing characteristics of various stations along in the central auditory pathway. There is good alignment with the neural STRFs and reported results in mammalian MGB and primary auditory cortex with respect to directional sensitivity and spectro-temporal separability. The temporal modulation statistics of the emergent sustained STRFs appear to be most similar to those measured from thalamus and cortex. Furthermore, the model STRFs are generally faster with regard to spectral modulations than those measured from thalamus and cortex.

Emergence of a sparse population code

To explore the relationship between STRFs optimized to promote sustained responses and those that explicitly maximize population sparsity, we compared the average responses of the sustained ensemble for $\Delta T = 125$ ms with the sparse ensemble. Specifically, we used the converged STRFs to analyze a held-out set of natural stimuli, computed a histogram of the population responses at each time, and computed the average histogram across the entire test input (see Methods). Since the sparse ensemble was optimized to yield a highly kurtotic firing rate distribution, it was of interest to examine the shape of the distribution when promoting sustained responses.

Results comparing the average histograms of sustained versus sparse responses is shown in Figure 8, with log-probabilities shown on the vertical axis to emphasize differences between the tails of the distributions. The main observation is that both the sustained and sparse ensembles have distributions that have long tails and are are highly peaked around a firing rate of zero. For reference, we show the average histograms obtained by filtering the stimulus through the first 400 principal components of the stimulus (see Supplemental Figure 3) as well as through a set of 400 random STRFs; a zero-mean, unit variance Gaussian distribution is also shown. Therefore, despite promoting temporally persistent responses, the sustained responses yield a population response that is not altogether different from an ensemble that explicitly maximizes kurtosis. Interestingly, this observation was also made by Berkes and Wiscott in the context of complex cell processing in primary visual cortex (see Sec. 6 of [33]).

## Emergent STRFs capture spectro-temporal modulation statistics of stimulus

Finally, we sought to explore the consequences of relaxing the constraint that the responses be mutually uncorrelated. Rather than directly constrain the *responses*, we considered constraints to the *shapes* of the model STRFs. This was achieved by solving

$$\arg\max_{H} J_{sus}(H) \quad \text{subject to} \quad H^T H = I,$$

i.e., we require the STRFs to form an orthonormal basis. So long as the stimuli are bounded, this set of constraints meets our requirements that (1) the output of the STRFs be bounded and (2) we minimize redundancy in the learned ensemble. As before, the optimization is described in the Methods. We consider an ensemble size of $K = 400$ STRFs initialized at random. Examples of shape-constrained STRFs that optimize the sustained objective function for $\Delta T = 125$ ms are shown in Figure 9. Again, we observe STRFs that are bandpass, localized, oriented, and sensitive to a variety of spectral and temporal input. However, there was an apparent difference between the speed of the spectro-temporal modulations and those from STRFs learned subject to the response constraints.

It is well known that natural sound ensembles are composed largely of slow spectro-temporal modulations [29,30,34]. However, the emergent STRFs learned subject to *response* constraints appear to be tuned to relatively *fast* spectral and temporal modulations, whereas the STRFs learned subject to *shape* constraints appear to have a broader tuning. To further examine how both sets of constraints jointly capture and are related to the spectro-temporal modulations observed in stimulus, we compared the average 2D modulation profile of the stimulus to the eMTFs derived from both sets of constraints.

An interesting view of how the emergent STRFs capture the spectro-temporal modulations of the stimulus is illustrated in Figure 10 for $\Delta T = 125$ ms. Shown is the average 2D modulation profile of the stimulus overlaid with a single isoline contour (at the 65% level) of the eMTFs learned subject to response (thick red lines) and shape constraints (thick black lines). We also show the constellation of BR versus BS for each ensemble (indicated by '×' and 'Δ' for response and shape constraints, respectively). As implied by the contours, the response constraints yield STRFs that follow the spectro-temporal
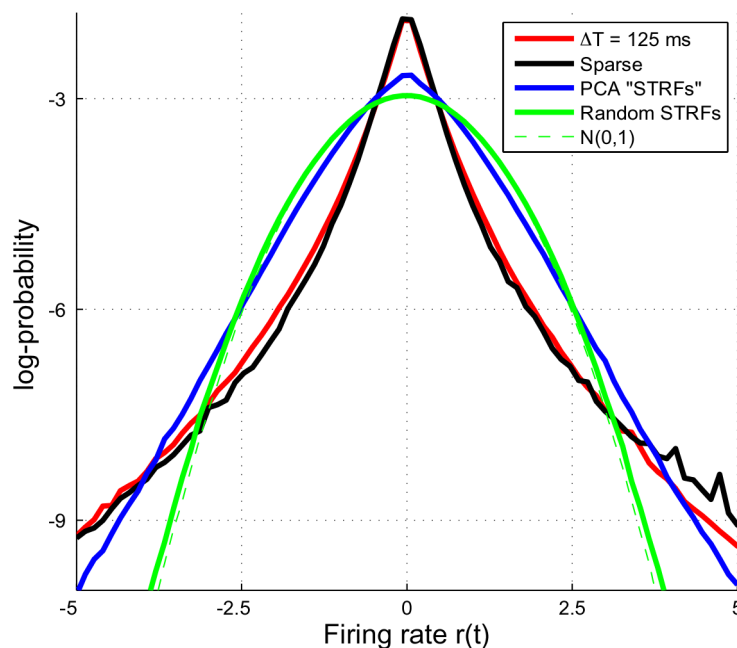
Figure 8. Average population response histograms for STRFs learned under the sustained and sparse objectives subject to response constraints.
doi:10.1371/journal.pcbi.1002982.g008
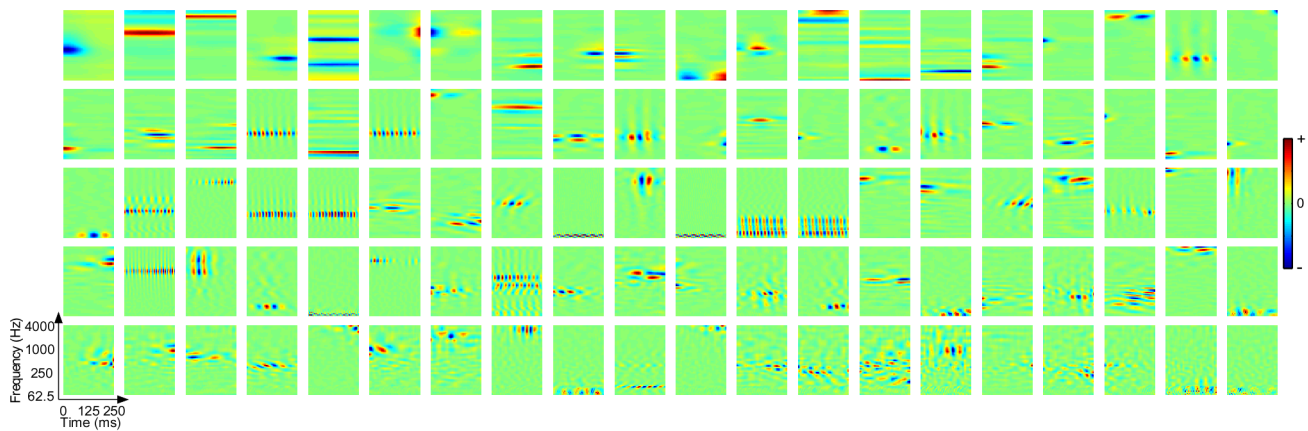
Figure 9. Examples of STRFs learned under the sustained objective function ($\Delta T = 125$ ms) subject to orthonormality constraints on the shapes of the filters. The examples shown here were drawn at random from an ensemble of 400 neurons, and the STRFs are shown in order of decreasing contribution to the overall objective function. Each spectro-temporal patch spans 0–250 ms in time and 62.5–4000 Hz in frequency. For these examples the dynamic range of the STRFs was compressed using a $\sinh(\cdot)$ nonlinearity.
doi:10.1371/journal.pcbi.1002982.g009

"edge" of the stimulus, while the shape constraints explicitly capture most of the "slowness" of the stimulus. As mentioned previously, the response constraints effectively force the temporal response of the sustained ensemble to be sparse, which consequently results in highly selective STRFs that tend to be tuned to fast modulations. Nevertheless, they implicitly capture the spectro-temporal extent of the stimulus. Moreover, since the shape constraints effectively force the STRFs to form a basis that spans the input space, this results in neurons that explicitly capture the slow modulations of the stimulus. Similar observations were made across the range of ΔT, and for each case it was clear that the spectro-temporal modulations of the stimulus are fully captured by the combination of both sets of constraints.

Discussion

In this paper, we considered a framework for studying how choice of a sustained firing versus sparse coding objective affects the shapes of model spectro-temporal receptive fields in central auditory areas. The sparse coding objective considered here, namely that of maximizing population kurtosis, yields STRFs that are mostly noisy. Those that do converge are generally highly localized. In contrast, enforcing the sustained firing objective subject to the same response constraints yields richly structured ensembles of STRFs whose population diversity varies smoothly with the correlation interval ΔT. Of course, the observed
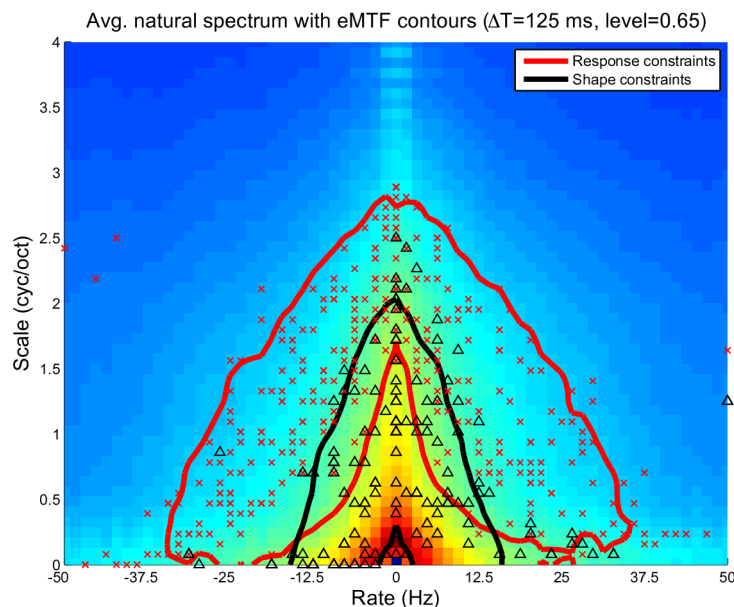
Figure 10. Spectro-temporal modulations in the stimulus are fully captured by STRFs that promote sustained responses subject to response and shape constraints. Here, the average MTF of the stimulus is overlaid with contours (at the 65% level) of the ensemble MTFs for both constraints for $\Delta T = 125$ ms. For each ensemble we also show the constellations for best rate vs. best scale (marked by '×' and 'Δ' for response and shape constraints, respectively). For the response constraints, we show the contour line and BR/BS constellations for STRFs that contribute to 99% of the objective function.
doi:10.1371/journal.pcbi.1002982.g010

structural variations are necessarily biased due to construction of the stimulus. Nevertheless, this diversity, as revealed by the results of the unsupervised clustering, paired with the responses of the most persistent STRFs, supports the notion that sustained neural firings are preferred in the range of timescales predominant in natural sounds. While we do not necessarily attribute the emergent sustained STRFs to any particular synapse in the auditory pathway, we instead note that the observed filters exhibit general similarities to physiological observations made in auditory thalamus and cortex.

We also observed that enforcing the sustained firing objective with response constraints yields an ensemble firing rate distribution that is similar, on average, to one where population sparsity was explicitly enforced. This supports the proposal that the two coding objectives are not necessarily at odds, and that in some sense a sustained firing objective yields "sparsity for free." Of course, the sustained firing and sparse coding objectives could be quantified in many different ways (see, e.g., Hashimoto [35] and Carlson *et al.* [11]), but the present study is a promising step in understanding their relationship in the central auditory system from a computational perspective.

Finally, to explore the consequences of relaxing the constraint that the responses be mutually uncorrelated, we explored an alternative set of orthonormality constraints on the sustained firing objective. While still minimizing a notion of redundancy, we observed that the emergent ensembles are generally slower, potentially better capturing the slow spectro-temporal modulations known to be present in natural sounds. This experiment further demonstrated the utility of the considered framework for directly addressing questions about coding schemes and various sets of constraints in representing sound in central auditory areas.

Emergence of a discriminative spectro-temporal representation for natural sounds

The combination of shape and response constraints on the sustained objective function yield STRF ensembles that appear to jointly capture the full range of spectro-temporal modulations in the stimulus. However, the distinct differences in MTF coverage illustrate the tradeoff between redundancy and efficiency in sensory representations. In particular, the shape constraints yield STRFs that are somewhat akin to the first few principal components of the stimulus (see Supplemental Figure 3). This is not surprising given that the objective function defines a notion of variance of linear projections, the component vectors of which are constrained to form an orthonormal basis. However, since the responses are not strictly enforced to be uncorrelated, orthonormality imposed on the filter shapes does not necessarily reduce redundancy in the resulting neural responses.

In contrast, the response constraints yield STRFs that are highly selective to the input and are thus comparatively "fast" in the modulation domain. This representation can be thought of as more efficient since at any given time only a few neurons have a large response. However, while the shapes of individual STRFs fail to explicitly capture the slow spectro-temporal modulations predominant in natural sounds, it instead appears that the ensemble MTF of the response-constrained STRFs collectively forms a contour around the high-energy modulations of the stimulus that implicitly capture its spectro-temporal extent.

Is this contouring of the average modulation spectrum of natural sounds something performed by the auditory system? The neural STRFs we considered certainly had an eMTF that reflects a tuning to slower modulations near the MTF origin. However, there is some evidence that the auditory system uses an "edge"-sensitive, discriminative modulation profile for analyzing sound.

Woolley *et al.* [36], in an avian study, showed that the eMTF of neurons from Field L (the avian A1 analog) has a bandpass temporal modulation profile (at low scales) that facilitates a discriminative tuning of temporal modulations among classes of natural sounds. Nagel and Doupe [37] have also shown examples of avian Field L STRFs that orient themselves near the spectro-temporal "edge" of the stimulus space. Moreover, Rodriguez *et al.* [38], in a study of mammalian IC neurons, showed that neural bandwidths can scale to better capture fast, but less frequently occurring, modulations. In light of these observations, the modulation profiles observed from the sustained STRFs for both response and shape constraints are consistent with the notion that the auditory system makes an explicit effort to capture all modulations present in natural sounds: fast, feature-selective, and consequently *discriminative* modulations, as well as frequently occurring slow modulations.

A neural code for sensory processing

The notion that sustained neural firings form part of the neural representation of sensory systems is not limited exclusively to the auditory modality. In fact, the sustained firing objective considered in this paper is related to a broad class of sensory coding strategies referred to collectively under the *temporal slowness hypothesis*. This concept proposes that the responses of sensory neurons reflect the time-course of the information-bearing components of the stimulus—which are often much slower with respect to the fast variations observed in the stimulus—and may therefore reflect invariant aspects of the sensory objects in the environment. Examples of early neural network models exploring slowness as a learning principle were considered by Földiák [39], Mitchison [40], and Becker [41]. More recently, a number of computational studies, particularly in vision, have established slowness as a general sensory coding strategy and have revealed relationships with a number of general machine learning techniques. Here we outline the connections between the sustained firing criterion considered in this study and previous work.

Our definition of the sustained firing objective, J_{sus}, was adapted from a notion of temporal stability proposed by Hurri and Hyvärinen termed *temporal response strength correlation* (TRSC) [18]. This study considered modeling of simple cells in primary visual cortex, and their objective function was defined as

$$J_{TRSC} = \sum_{k=1}^{K} \langle r_k^2(t) r_k^2(t-\tau) \rangle_t \qquad (3)$$

for a single fixed τ. By maximizing J_{TRSC} subject to the decorrelation constraints $\langle r_j(t) r_k(t) \rangle_t = \delta_{jk}$, they showed the emergence of spatial receptive fields similar to those observed in simple cells in primary visual cortex. It is clear that the objective functions J_{sus} and J_{TRSC} are equivalent for a single time step, but the main difference between the two is that we sought to enforce temporal stability over a time *interval* $[t-\Delta T, t]$, rather than between two *distinct* times t and $(t-\tau)$. Interestingly, optimization of the TRSC objective was shown by Hyvärinen to yield a solution to the blind source separation problem [42], suggesting perhaps that in the auditory domain, such a criterion may underlie separation of overlapping acoustic sources.

The sustained firing objective is also related to a well-known model of temporal slowness known as *slow feature analysis* (SFA) [20]. The computational goal of SFA is to find a mapping of an input that extracts the slow, and presumably more invariant, information in the stimulus. Briefly, for an input $\mathbf{x}(t)$, linear SFA finds mappings $y_k(t) = h_k^T \mathbf{x}(t)$ that minimize

$$J_{SFA} := \langle (y_k(t) - y_k(t-1))^2 \rangle_t \tag{4}$$

subject to $\langle y_k(t) \rangle_t = 0$, $\langle y_k^2(t) \rangle_t = 1$, and $\langle y_j(t)y_k(t) \rangle_t = 0 \forall j < k$. Note that the input $\mathbf{x}(t)$ is not necessarily the raw stimulus but could represent a non-linear expansion of the input, akin to applying a kernel function in a support vector machine [43]. Therefore, SFA finds a mapping of the input that varies little over time and whose outputs are bounded and mutually uncorrelated. In the visual domain, Berkes and Wiskott found that SFA could explain a variety of complex cell phenomena in primary visual cortex such as the emergence of Gabor-like receptive fields, phase invariance, various forms of inhibition, and directional sensitivity [33]. Similar to our study, they also found the emergence of a sparse population code based on SFA. More importantly, however, they established a link between SFA at the level of complex cells and J_{TRSC}, which in turn links to the sustained firing objective J_{sus} explored in our study. Specifically, they showed that when a complex cell output is expressed as a quadratic form $y(t) = \sum_k r_k^2(t)$ [35,44], the SFA objective could be written as

$$J_{SFA} = \sum_{k=1}^{K} \langle r_k^2(t) r_k^2(t-1) \rangle_t + \sum_{j \neq k} \langle r_j^2(t) r_k^2(t-1) \rangle_t \tag{5}$$

which is equivalent to maximizing J_{TRSC} (and thus J_{sus} for a single time-step) plus cross-correlation terms. As noted by Berkes and Wiskott, this relationship suggests that sustained firing rates at the level of simple cells are modulated as part of a hierarchical cortical processing scheme in primary visual cortex. Given the increasing understanding of such hierarchical circuits in the auditory system [45], the possibility that sustained firing rates are varied as part of a higher-order processing strategy in primary auditory areas is an exciting prospect worth further exploration.

Other important relationships exist between SFA and a number of general machine learning principles. Blaschke *et al.* [46] established a relationship between SFA and independent component analysis, a widely used method for blind source separation (see, e.g., [47]). Klampfl and Maass [48] showed that under certain slowness assumptions about the underlying class labels in observed data, SFA finds a discriminative projection of the input similar to Fisher's linear discriminant. Furthermore, SFA has links to methods for nonlinear dimensionality reduction: Creutzig and Sprekeler [49] described the link between SFA and the information bottleneck whereas Sprekeler [50] showed a connection between SFA and Laplacian eigenmaps.

In summary, the temporal slowness hypothesis forms a sound basis for learning a representation from data with rich temporal structure. Slowness as a learning principle has also been shown to explain the emergence of simple and complex cell properties in primary visual cortex. As described above, the sustained firing principle considered in this paper has fundamental links to SFA, which in turn is related to a number of general machine learning strategies. To the best of our knowledge, ours is the first thorough study that establishes a link between the temporal slowness hypothesis and an emergent spectro-temporal representation of sound in central auditory areas.

Implications for automated sound processing systems

The ensemble modulation coverage results are particularly interesting since it is widely thought that "slow" spectro-temporal modulations carry much of the message-bearing information for human speech perception. Furthermore, it is known in the speech processing community that features that capture slow temporal [51] and joint spectro-temporal modulations [52,53] are important for noise-robust automatic speech recognition. The observed contouring effect resulting from the sustained firing criterion may thus reflect a mechanism to detect the spectro-temporal "edges" of the message-bearing components of the stimulus, and possibly contribute to a noise-robust representation of sound. We have recently considered this principle and have demonstrated that 2D bandpass filters derived from eMTF contours learned from a speech-only stimulus yield state-of-the-art noise-robust acoustic features for automatic speech recognition [54]. Moreover, it is possible that the contour level may be chosen adaptively as a function of ambient signal-to-noise ratio to better capture variations in the high-energy modulations of the stimulus. Also, since the emergent STRFs capture general spectro-temporal patterns that characterize the stimulus, it is possible that ensembles of STRFs could be learned in various speech-plus-noise scenarios to perhaps better characterize noise-corrupted acoustic environments. Such hypotheses can be readily verified experimentally and may have practical impact to automated sound processing systems in noisy acoustic environments.

Concluding remarks

Finally, the framework considered in this paper can be extended in a number of ways. For instance, to address the linearity limitation of the STRF, it is worthwhile to consider a model based on a linear-nonlinear cascade [55]. As mentioned earlier, the auditory pathway is necessarily hierarchical, and warrants consideration of hierarchical computational models. Indeed, recent physiological evidence also indicates that the representation becomes increasingly complex and nonlinear as one moves from away thalamo-recipient layers in primary auditory cortex (for a review, see [45]). Finally, a recent computational study in vision by Cadieu and Olshausen [56] proposes a hierarchical generative model that explicitly unifies notions of sparse coding and temporal stability. In particular, a two-layer network learns a sparse input representation whose activations vary smoothly over time, whereas a second layer modulates the plasticity of the first layer, resulting in a smooth time-varying basis for image sequences. One can imagine that such a framework could be extended to spectro-temporal acoustic stimuli.

Methods

Stimulus description and preparation

An ensemble of natural sounds comprising segments of speech, animal vocalizations, and ambient outdoor noises was assembled for use as stimuli. Two sets were generated, one for training and one for evaluating the response characteristics of the STRFs. Phonetically balanced sentences read by male and female speakers were used [57]. Examples of animal vocalizations included barking dogs, bleating goats, and chattering monkeys [58]. The ambient sounds included, for example, babbling creeks and blowing wind, and other outdoor noises. The speech utterances were approximately three seconds each and comprised 50% of the stimulus. The animal vocalizations and ambient sounds formed the remaining 50% of the stimulus (25% each), were broken into three-second segments, and were windowed using a raised cosine window to avoid transient effects. Finally, segments from each class were downsampled to 8 kHz, standardized to be zero-mean and unit variance, and randomly concatenated to yield a waveform approximately three minutes in overall length, i.e., ~ 90 seconds of speech, ~ 45 seconds of animal vocalizations, and ~ 45 seconds of ambient outdoor noises.

We used a computational model of peripheral processing to account for the transformation of a monaural acoustic stimulus to a joint time-frequency representation in the auditory midbrain; this representation is referred to as an *auditory spectrogram* [59,60]. The auditory spectrogram represents the time-varying spectral energy distribution on the (logarithmic) tonotopic axis, and accounts for the physiology of inner hair cell transduction and filtering on the auditory nerve, enhanced frequency selectivity in the cochlear nucleus via a lateral inhibitory network, and the loss of phase locking to stimuli observed in midbrain nuclei. The specific model details have been presented previously and as such we forego a detailed description here, except to note that we sampled the log-frequency axis over six octaves with ten equally spaced channels per octave, with a short-term integration interval of 5 ms, i.e., we obtained a 60 channel spectral vector every 5 ms. An example auditory spectrogram is shown for a segment of speech in Figure 1A.

Spectro-temporal receptive fields

To quantify the relationship between a spectro-temporal stimulus and its corresponding response in central auditory areas, we used the spectro-temporal receptive field. Such a functional characterization of a neuron is useful for identifying the components of the stimulus to which it is most sensitive. An STRF models the linear transformation of a time-varying spectro-temporal input to an instantaneous firing rate, i.e.,

$$r(t) = \int\int h(\tau,f)s(t-\tau,f)d\tau df + r_0, \qquad (6)$$

where $h(t,f)$ is an LTI filter that defines the STRF, $s(t,f)$ is a spectro-temporal stimulus, and r_0 is the average firing rate. Without loss of generality, we assume $r_0 = 0$. Observe that the mapping represents convolution in time and integration across all frequencies, and we can interpret the STRF as a matched filter that acts on the input auditory spectrogram.

For discrete-time signals and filters, and assuming that $h(t,f)$ has a finite impulse response, we can express Eq. 6 compactly in vector notation as

$$r(t) = \mathbf{h}^T \mathbf{s}(t), \qquad (7)$$

where $\mathbf{s}(t), \mathbf{h} \in \mathbb{R}^d$ are column vectors denoting the stimulus and filter, respectively [61]. Furthermore, to express the response $\mathbf{r}(t) = [r_1(t) r_2(t) \cdots r_K(t)]^T \in \mathbb{R}^K$ of an *ensemble* of K neurons, we concatenate the STRFs into a matrix $H := [\mathbf{h}_1 \mathbf{h}_2 \cdots \mathbf{h}_K] \in \mathbb{R}^{d \times K}$ and write

$$\mathbf{r}(t) = H^T \mathbf{s}(t). \qquad (8)$$

From the stimulus auditory spectrogram, we extracted 250 ms spectro-temporal segments once every 5 ms. Each segment was stacked columnwise into a vector $\mathbf{s}(t) \in \mathbb{R}^d$ where $d = 3000$ (i.e., 50 vectors/segment $\times$ 60 channels). A total of $\sim$30 k spectro-temporal vectors were extracted from the stimulus. We subtracted the local mean from each segment and scaled each vector to be unit norm [18], and note that this pre-processing was also applied to the test stimulus used for evaluating the STRF response characteristics. Finally, each spectro-temporal input patch was processed by the ensemble of STRFs to yield a population response $\mathbf{r}(t)$. Figure 1B illustrates the procedure for obtaining stimulus vectors $\mathbf{s}(t)$ and response vector $\mathbf{r}(t)$.

Optimization

To constrain the responses of the STRFs to have unit variance and be mutually uncorrelated, we first note that the individual constraints can be written as

$$\langle r_j(t)r_k(t) \rangle_t = \mathbf{h}_j^T \langle \mathbf{s}(t)\mathbf{s}^T(t) \rangle_t \mathbf{h}_k = \mathbf{h}_j^T C_{\mathbf{s}} \mathbf{h}_k = \delta_{jk},$$

which can then be compactly expressed as an ensemble constraint

$$H^T C_{\mathbf{s}} H = I, \qquad (9)$$

where $C_{\mathbf{s}} := \langle \mathbf{s}(t)\mathbf{s}^T(t) \rangle_t \in \mathbb{R}^{d \times d}$ denotes the sample covariance matrix and $I \in \mathbb{R}^{K \times K}$ is the identity matrix. Since $C_{\mathbf{s}}$ is real-symmetric, it is unitarily diagonalizable as $C_{\mathbf{s}} = E \Lambda E^T$, where $E \in \mathbb{R}^{d \times d}$ is a matrix of (columnwise) eigenvectors with corresponding eigenvalues along the diagonal of $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_d) \in \mathbb{R}^{d \times d}$. Substituting this decomposition into Eq. 9, we obtained

$$H^T C_{\mathbf{s}} H = H^T (E \Lambda E^T) H$$
$$= H^T E \Lambda^{1/2} \Lambda^{1/2} E^T H$$
$$= U^T U,$$

where $U := \Lambda^{1/2} E^T H \in \mathbb{R}^{d \times K}$. By recasting the constraints, we can rewrite the original matrix of STRFs as $H = E \Lambda^{-1/2} U$ and consequently

$$\mathbf{r}(t) = H^T \mathbf{s}(t) = U^T \Lambda^{-1/2} E^T \mathbf{s}(t) = U^T \mathbf{z}(t),$$

where $\mathbf{z}(t) := \Lambda^{-1/2} E^T \mathbf{s}(t)$ corresponds to a *whitening* of the input acoustic data, i.e., $\mathbf{z}(t)$ has a spherical covariance matrix. For computational efficiency, we reduced the dimensionality of the input using a subset of the principal components of the stimulus, i.e.,

$$\mathbf{z}(t) \approx \Lambda_m^{-1/2} E_m^T \mathbf{s}(t),$$

where Λ_m and E_m, $m < d$, are the matrices of eigenvalues and eigenvectors, respectively, that captured 95% of the variance of the input. In this work, we found $m = 468$. Therefore, the core problem we wished to solve is:

$$\arg \max_U J(U) \quad \text{subject to} \quad U^T U = I, \qquad (10)$$

where $J(\cdot)$ corresponded to either the sustained firing or sparse coding objective function.

To optimize this nonlinear program, we used the gradient projection method due to Rosen, the basic idea of which is as follows [62,63]. Let $U_{(n)}$ denote the n'th update to the matrix of (rotated and scaled) STRFs U, let $\alpha > 0$ be a learning rate, and let $m \in \mathbb{N}$ be an integer used to adjust the learning rate. Assume $U_{(n)}$ is a matrix with orthonormal columns that is a feasible solution to the problem in Eq. 10. We updated U via gradient ascent as follows:

$$U_{(n+1)} = \mathcal{P}\left(U_{(n)} + \frac{\alpha}{2^m} \cdot \frac{\partial J(U_{(n)})}{\partial U} \right), \qquad (11)$$

where $\mathcal{P} : \mathbb{R}^{d \times K} \to \mathbb{R}^{d \times K}$ is a projection of the gradient update so that $U_{(n+1)}$ satisfies the orthonormality constraint required in Eq. 10. If the update was such that $J(U_{(n+1)}) < J(U_{(n)})$, we set

$m \leftarrow m+1$ and recomputed the projected gradient update, repeating until $J(U)$ was non-decreasing. Finally, learning ceased when the relative change between $J(U_{(n)})$ and $J(U_{(n+1)})$ fell below a threshold η or a maximum number of iterations were reached; in our experiments, we stopped learning for $\eta < 0.1\%$ or a maximum number of 30 iterations. Upon convergence, the desired STRFs were obtained using $H = E\Lambda^{-1/2}U$. Note that for the case of the sustained firing objective, J_{sus} was formed from the sum of K independent terms, allowing us to directly sort the emergent STRFs according to their contribution to the overall objective function; such a sorting was not possible for the sparsity objective.

Of course, the above procedure required a suitable projection $\mathcal{P}(\cdot)$, and one was derived as follows [64]. In general, for a matrix $A \in \mathbb{R}^{m \times n}$, we wish to find a matrix $V \in \mathbb{R}^{m \times n}$ with orthonormal columns that minimizes

$$\|A-V\|_F^2 \quad \text{subject to} \quad V^T V = I.$$

Introducing a symmetric matrix of Lagrange multipliers $L \in \mathbb{R}^{n \times n}$, and recalling that $\|A\|_F^2 = Tr(AA^T)$, we sought to find a stationary point of the Lagrangian

$$l(V,L) = \text{Tr}\left[(A-V)(A-V)^T\right] + \text{Tr}\left[L(V^T V - I)\right].$$

Computing the (elementwise) partial derivative of $l(V,L)$ w.r.t. V and setting it to 0 we obtained [65]

$$A = V(I+L).$$

Observing that

$$A^T A = (I+L)V^T V(I+L) = (I+L)^2,$$

we have that

$$(I+L) = (A^T A)^{1/2}.$$

Assuming A had full column rank, then an optimal orthogonal matrix that minimized $\|A-V\|_F^2$ that can be used for the projection in Eq. 11 was found as

$$\mathcal{P}(A) = V = A(A^T A)^{-1/2}. \tag{12}$$

Finally, to optimize a given objective function subject to the STRFs being orthonormal, i.e., $\mathbf{h}_j \mathbf{h}_k = \delta_{jk}$, we solve

$$\arg \max_H J(H) \quad \text{subject to} \quad H^T H = I.$$

Here we can again use Rosen's projected gradient method in Eq. 11 along with the projection defined in Eq. 12, but the only difference from before is that it *does not* require pre-whitening of the stimulus.

Characterizing individual STRFs

We first characterized the emergent STRFs based on parameters that described their individual spectro-temporal and modulation tuning.

Separability index. We used a measure of separability to quantify how well an STRF $h(t,f)$ could be decomposed into a product of purely temporal and spectral functions, i.e., as

$h(t,f) = h_T(t) \cdot h_S(f)$ [23]. Generally speaking, by treating an STRF as a matrix $T \in \mathbb{R}^{m \times n}$, separability can be assessed by considering the singular value decomposition of T:

$$T = U\Sigma V^T = \sum_{i=1}^r \sigma_i u_i v_i^T,$$

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are unitary, $\Sigma \in \mathbb{R}^{m \times n}$ is a matrix such that the *singular values* $\sigma_i \geq 0$ lie along the "diagonal", and $r = \text{rank}(T)$. The separability index SPI was defined as

$$SPI = 1 - \frac{\sigma_1^2}{\sum_{i=1}^r \sigma_i^2}.$$

If T is nearly rank-1, we expect σ_1 to dominate and consequently SPI is small, indicating that $T \approx \sigma_1 u_1 v_1^T$, i.e., that the STRF is approximately separable as a product of only two functions. It was often the case that STRFs with a simpler structure, e.g., localized or purely spectral, had small values of SPI. More complex STRFs, particularly those that were noisy, had larger values SPI since they were poorly approximated by a low-rank decomposition.

Modulation transfer function. To characterize spectro-temporal modulation tuning in the Fourier domain, we computed the *modulation transfer function* (MTF) of an STRF, illustrated in Figure 11B [24]. The MTF was obtained by computing the magnitude of the 2D Fourier transform of a thresholded STRF; here we set all values of the STRF that did not exceed ± 1 standard deviation to zero. The MTF summarizes the joint sensitivity of an STRF to temporal modulations (*rate*, in Hz) and spectral modulations (*scale*, in cyc/oct).

Best spectral and temporal modulation rates. We selected the peak of the MTF to estimate best rate (BR) and best scale (BS). We expected that BR and BS would summarize an STRF's preference for fast or slow temporal and spectral modulations.

Average rate and scale profiles. By folding the MTF along the $\omega = 0$Hz axis, we summarized the temporal and spectral modulation sensitivity of the STRF by summing along each axis, yielding rate and scale profiles; these are illustrated in Figure 11C. These profiles can also be averaged across an ensemble of neurons to yield a population rate or scale profile.

Directionality index. To characterize whether a neuron preferred upward vs. downward stimuli, we computed a directionality index by considering the relative difference in spectro-temporal modulation energy in the first and second quadrants in the Fourier domain. This was quantified as [23]

$$DIR = \frac{E_1 - E_2}{E_1 + E_2},$$

where E_1 and E_2 denote the energy in the first and second quadrant, respectively. By convention, $DIR > 0$ indicates a preference for *downward* moving spectro-temporal patches whereas $DIR < 0$ indicates a preference for *upward* moving spectro-temporal patches.

Compactness. To quantify a notion of compactness for an STRF, we used the *isoperimetric quotient*, which considers the ratio of the area of an ellipsoid to its perimeter, i.e.,

$$COMP = \frac{4\pi \cdot \text{area}}{\text{perimeter}^2}.$$
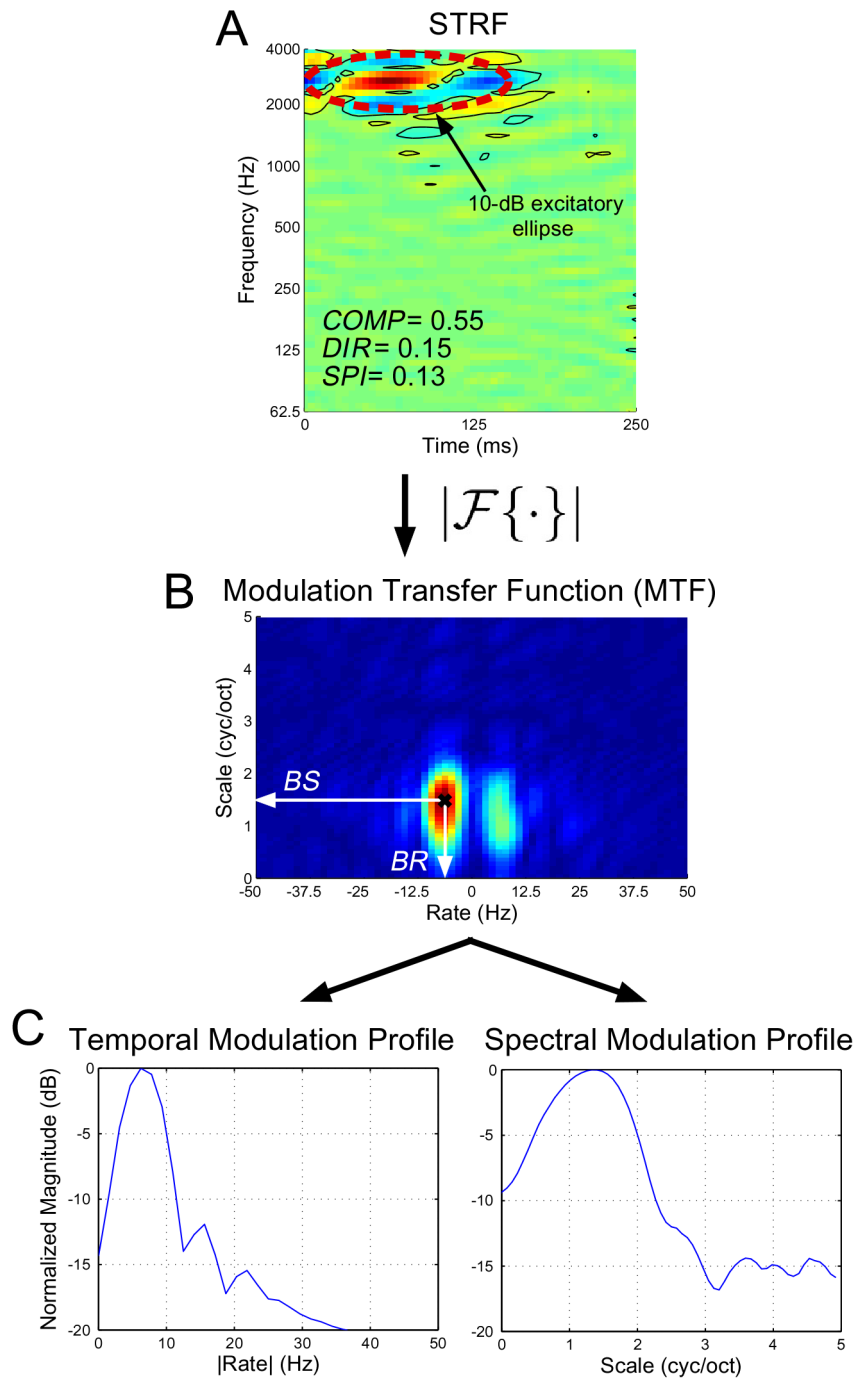
Figure 11. Extracting basic spectro-temporal parameters for an individual STRF. Panel (A) shows a typical STRF, with solid contour lines indicating those regions that exceed $\pm$ one standard deviation. The dashed red line shows the projected 10-dB ellipse from which we estimated spectral bandwidth. As indicated, the STRF is rather elongated with no strong directional preference, and the pattern is highly separable. Panel (B) shows the MTF computed from the magnitude of the 2D Fourier Transform of the STRF in (A); from here we estimate BR and BS. Panel (C) shows the normalized temporal and spectral modulation profiles obtained from the MTF.
doi:10.1371/journal.pcbi.1002982.g011

The area and perimeter were computed from the 10-dB excitatory ellipse which was derived by (1) performing a least-squares fit of a single Gaussian envelope to a thresholded STRF, (2) finding the isoline corresponding to a drop of 10-dB from the maximum of the envelope, and (3) projecting this ellipse onto the spectro-temporal plane. The compactness measure describes the degree to which the coverage of an STRF is spherical ($COMP = 1$) versus

elongated ($COMP < 1$), and was used for characterizing localized vs. non-localized STRFs for the purpose of grouping STRF clusters (described below).

Characterizing STRF ensembles

Next, we considered measures that characterized a variety of ensemble-based spectro-temporal and modulation properties.

Ensemble modulation transfer function. By averaging the MTF obtained from each STRF, we obtained an ensemble MTF (eMTF) that characterized the average spectro-temporal modulation sensitivity of a given ensemble [24]. This representation was used to relate the average modulation tuning of an ensemble to the modulations present in the stimulus.

Median activation of most persistent neurons. In addition to analyzing the shapes of the emergent STRFs, we explored the ensemble firing rate characteristics of the emergent neurons. Using a held-out set of natural stimuli, we measured the activation of a neuron as the length of time a response was maintained above ± 1 standard deviation (over time) for that particular neuron. We sorted each STRF according to its median activation time, and considered the median responses of the top 10% "most persistent" neurons for a given ensemble (as these subsets appeared to vary most across ΔT). The distributions of these activations were then used to study the extent to which enforcing a sustained response was reflected in a neuron's output.

Average population response histogram. In order to compare distributions of population responses across ensembles, we computed averaged response histograms as follows. Upon convergence of a given ensemble, we filtered a held-out set of natural sound stimuli through the emergent STRFs to obtain a population response. At each time t, we computed a histogram of the population response, and computed the average histogram across the duration of the stimulus. These averaged histograms could then be used to compare the average population response characteristics across ensembles.

When comparing the receptive field ensembles from the sparse and sustained sets, we only included the responses of highly structured, non-noisy STRFs as determined by the clustering results outlined next. This step was necessary to keep the comparison between objective functions fair since the sparse ensemble was dominated by noisy STRFs. This inclusion criterion resulted in 115 and 347 neurons for the sparse and sustained ensembles, respectively.

For comparison, we also calculated the response histograms for stimuli filtered through the first 400 principal components of the stimulus (Supplemental Figure 3) as well as through a set of 400 random STRFs. Recall that the magnitudes of the emergent STRFs were constrained so that that their responses had unit variance over time. Accordingly, we normalized the responses of the principal components and random STRFs to also have unit variance to make a fair comparison.

Average stimulus 2D modulation profile

To summarize the spectro-temporal modulations present in the natural sound stimulus, we averaged the magnitude of the 2D Fourier transform of 250 ms patches (non-overlapping) of the auditory spectrogram.

Grouping canonical classes of STRFs

The optimization procedure resulted in a set of richly structured patterns that suggested the presence of a number of latent classes whose membership varied with both choice of objective function and correlation interval ΔT. To quantify these variations, we applied the normalized spectral clustering algorithm of Ng *et al.* [66].

We defined the similarity s_{mn} between a given pair of STRFs $h_m(t,f)$ and $h_n(t,f)$ by computing the normalized 2D cross-correlation matrix for arbitrary shifts in time and frequency and selecting the maximum of the *absolute* value of this matrix, i.e.,

$$s_{mn} = \max_{i,j} |c_{ij}(m,n)|,$$

where

$$c_{ij}(m,n) = \frac{\sum_t \sum_f h_m(i,j) h_n(t+i,f+j)}{||h_m(t,f)||_F \cdot ||h_n(t,f)||_F}.$$

Importantly, the absolute value of the cross correlation was used here since we wished to group STRFs regardless of whether they were excitatory or inhibitory. Next, we pooled all STRFs we sought to cluster and constructed a pairwise similarity matrix $S = [s_{mn}] \in \mathbb{R}^{N \times N}$. Viewing S as a fully connected graph with edge weights specified by s_{mn}, spectral clustering finds a partitioning of the graph into k groups such that edges between groups have low similarity whereas edges within a group have high similarity.

Defining the degree matrix $D = \text{diag}(d_1, d_2, \cdots, d_N)$ where $d_m = \sum_n s_{mn}$ and unnormalized graph Laplacian $L = D - S$, the normalized spectral clustering algorithm is as follows:

1. Compute the normalized Laplacian $L_{sym} = D^{-1/2} L D^{-1/2}$.
2. Compute the first k eigenvectors $\{v_1 v_2 \cdots v_k\}$ corresponding to the largest k eigenvalues of L_{sym}.
3. Let $V = [v_1 v_2 \cdots v_k] \in \mathbb{R}^{N \times k}$ and form a matrix W from V by normalizing each row to have unit Euclidean norm.
4. Denoting $w_n \in \mathbb{R}^k$ as the n'th row of W, cluster the set of points $\{w_n\}$ using the k-means algorithm to obtain clusters $\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_k$.

We clustered the STRFs initially into 12 groups. While this number was necessarily an arbitrary choice, it was found to sufficiently capture variations in population diversity with ΔT. However, we found that (i) three of the resulting clusters could be reasonably labeled as *noisy*, whereas (ii) two of the resulting clusters could be reliably labeled as *localized*; merely reducing the number of initial classes did not merge the clusters, but instead blurred distinctions among the other major categories we sought to study. We interpreted *noisy* patterns as those with no obvious spectro-temporal structure and not indicative of any subset of the stimulus.

Merging of the initial 12 classes was achieved by computing the average *SPI* of STRFs from the initial class labels and ranking the classes in descending order. Indeed, the three *noisy* classes had the highest average *SPI* and consequently resulted in a group with average *SPI* greater than 0.5. Similarly, the localized STRFs were typically highly spherical and sorting the initial clusters by *COMP* resulted in the two *localized* classes to be ranked highest. Consequently, we grouped these two clusters that had an average *COMP* of greater than 0.69. This resulted in a final cluster count of nine classes.

Analysis of Neural STRFs

We obtained ensembles of neural STRFs estimated using TORC [31] and speech stimuli [27,32]. There were 2145 TORC and 793 speech STRFs, and each STRF was pre-processed to cover 110 ms in time (sampling rate = 100 Hz) and span 5 octaves in frequency (sampling rate = 5 cyc/oct). For the spectral clustering analysis, we subsampled the TORC set by randomly selecting 793 STRFs and combined them with the speech STRFs, yielding a total of 1586 STRFs in the neural data set. In this way, the neural data analysis was not biased towards one stimulus type or the other.

Supporting Information

Figure S1   STRFs corresponding to the top 10% ''most persistent'' responses for $\Delta T = 125$ ms.
(TIF)

Figure S2   Distributions of nearest-neighbor similarities for the model ensembles (response- and shape-constrained sustained objective vs. the sparse objective) and the neural ensemble.
(TIF)

Figure S3   Top 100 principal components of the natural stimulus ensemble.
(TIF)

Author Contributions

Conceived and designed the experiments: MAC ME. Performed the experiments: MAC. Analyzed the data: MAC ME. Wrote the paper: MAC ME.

References

1. Simoncelli EP, Olshausen BA (2001) Natural image statistics and neural representation. Annu Rev Neurosci 24: 1193–1216.
2. Olshausen BA, Field DJ (2004) Sparse coding of sensory inputs. Curr Op Neurobio 14: 481–487.
3. Rosenblith WA, editor (1961) Sensory Communication. Cambridge (Massachusetts): MIT Press.
4. Laughlin SB (2001) Energy as a constraint on the coding and processing of sensory information. Curr Op Neurobio 11: 475–480.
5. Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381: 607–609.
6. Vinje WE, Gallant JL (2000) Sparse coding and decorrelation in primary visual cortex during natural vision. Science 287: 1273–1276.
7. DeWeese MR, Wehr M, Zador AM (2003) Binary spiking in auditory cortex. J Neurosci 23: 7940–7949.
8. Hromádka T, DeWeese MR, Zador AM (2008) Sparse representation of sounds in the unanesthetized auditory cortex. PLoS Bio 6: e16.
9. Klein DJ, König P, Körding KP (2003) Sparse spectrotemporal coding of sounds. EURASIP J Appl Sig Proc 2003: 659–667.
10. Smith EC, Lewicki MS (2006) Effcient auditory coding. Nature 439: 978–982.
11. Carlson NL, Ming VL, DeWeese MR (2012) Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. PLoS Comp Bio 8: e1002594.
12. Wang X, Lu T, Snider RK, Liang L (2005) Sustained firing in auditory cortex evoked by preferred stimuli. Nature 435: 341–346.
13. Wang X, Lu T, Bendor D, Bartlett E (2008) Neural coding of temporal information in auditory thalamus and cortex. Neuroscience 157: 484–493.
14. Petkov CI, O'Connor KN, Sutter ML (2007) Encoding of illusory continuity in primary auditory cortex. Neuron 54: 153–165.
15. Wang X (2007) Neural coding strategies in auditory cortex. Hearing Research 229: 81–93.
16. Elhilali M, Fritz JB, Klein DJ, Simon JZ, Shamma SA (2004) Dynamics of precise spike timing in primary auditory cortex. J Neurosci 24: 1159–1172.
17. Elhilali M, Shamma SA (2008) A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. J Acoust Soc Am 124: 3751–3771.
18. Hurri J, Hyvärinen A (2003) Simple-cell-like receptive fields maximize temporal coherence in natural video. Neural Comp 15: 663–691.
19. Körding KP, Kayser C, Einhäuser W, König P (2004) How are complex cell properties adapted to the statistics of natural stimuli? J Neurophys 91: 206–212.
20. Wiskott L, Sejnowski TJ (2002) Slow feature analysis: unsupervised learning of invariances. Neural Comp 14: 715–770.
21. Middlebrooks JC (2005) Auditory cortex cheers the overture and listens through the finale. Nature Neurosci 8: 851–852.
22. Aertsen AMHJ, Johannesma PIM (1981) The spectro-temporal receptive field. Biol Cybernetics 42: 133–143.
23. Depireux DA, Simon JZ, Klein DJ, Shamma S (2001) Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J Neurophys 85: 1220–1234.
24. Miller LM, Escabí MA, Read HL, Schreiner CE (2002) Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J Neurophys 87: 516–527.
25. Fritz J, Shamma S, Elhilali M, Klein D (2003) Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. Nature Neurosci 6: 1216–1223.
26. Willmore B, Tolhurst DJ (2001) Characterizing the sparseness of neural codes. Network: Computation in Neural Systems 12: 255–270.
27. David SV, Mesgarani N, Fritz JB, Shamma SA (2009) Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli. J Neurosci 29: 3374–3386.
28. Wooley SMN, Gill PR, Fremouw T, Theunissen FE (2009) Functional groups in the avian auditory system. J Neurosci 20: 2780–2793.
29. Rosen S (1992) Temporal information in speech: acoustic, auditory, and linguistic aspects. Phil Trans R Soc Lond B 336: 367–373.
30. Singh NC, Theunissen FE (2003) Modulation spectra of natural sounds and ethological theories of auditory processing. J Acoust Soc Am 114: 3394–3411.
31. Klein DJ, Depireux DA, Simon JZ, Shamma SA (2000) Robust spectrotemporal reverse correlation for the auditory system: Optimizing stimulus design. J Comp Neurosci 9: 85–111.
32. Mesgarani N, David SV, Fritz JB, Shamma SA (2009) Inuence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. J Neurosci 102: 3329–3339.
33. Berkes P, Wiskott L (2005) Slow feature analysis yields a rich repertoire of complex cell properties. J Vision 5: 579–602.
34. Attias H, Schreiner CE (1997) Temporal low-order statistics of natural sounds. Adv. Neural Inf. Proc. Sys. 9: 27–33.
35. HashimotoW(2003) Quadratic forms in natural images. Network: Computation in Neural Systems 14: 765–788.
36. Woolley SMN, Fremouw TE, Hsu A, Theunissen FE (2005) Tuning for spectrotemporal modulations as a mechanism for auditory discrimination of natural sounds. Nature Neurosci 8: 1371–1379.
37. Nagel KI, Doupe AJ (2008) Organizing principles of spectro-temporal encoding in the avian primary auditory area Field L. Neuron 58: 938–955.
38. Rodriguez FA, Chen C, Read HL, Escabí MA (2010) Neural modulation tuning characteristics scale to efficiently encode natural sound statistics. J Neurosci 30: 15969–15980.
39. Földiák P (1991) Learning invariances from transformational sequences. Neural Comp 3: 194–2000.
40. Mitchison G (1991) Removing time variation with the anti-Hebbian differential synapse. Neural Comp 3: 312–320.
41. Becker S (1993) Learning to categorize objects using temporal coherence. Adv Neural Inf Proc Sys 5: 361–368.
42. Hyvärinen A (2001) Blind source separation by nonstationary of variance: a cumulant-based approach. IEEE Trans Neural Networks 12: 1471–1474.
43. Bishop CM (2006) Pattern Recognition and Machine Learning. New York: Springer. 740 p.
44. Berkes P, Wiskott L (2006) On the analysis and interpretation of inhomogeneous quadratic forms as receptive fields. Neural Comp 18: 1868–1895.
45. Sharpee TO, Atencio CA, Schreiner CE (2011) Hierarchical representations in the auditory cortex. Curr Op Neurobio 21: 761–767.
46. Blaschke T, Berkes P, Wiskott L (2006) What is the relation between slow feature analysis and independent component analysis? Neural Comp 18: 2495–2508.
47. Hyvärinen A, Karhunen J, Oja E (2001) Independent Component Analysis. New York: John Wiley and Sons. 504 p.
48. Klampfl S, Maass W (2010) A theoretical basis for emergent pattern discrimination in neural systems through slow feature extraction. Neural Comp 22: 2979–3035.
49. Creutzig F, Sprekeler H (2008) Predictive coding and the slowness principle: an informationtheoretic approach. Neural Comp 20: 1026–1041.
50. Sprekeler H (2011) On the relation of slow feature analysis and Laplacian eigenmaps. Neural Comp 23: 3287–3302.
51. Hermansky H, Morgan N (1994) RASTA processing of speech. IEEE Trans Speech and Audio Process 2: 382–395.
52. Nemala SK, Patil K, Elhilali M (2011) Multistream bandpass modulation features for robust speech recognition. In: Interspeech 2011, 12th Annual Conference of the International Speech Communication Association; 27–31 August 2011; Florence, Italy.
53. Nemala SK (2012) Robust Speech Recognition by Humans and Machines: The Role of Spectro-Temporal Modulations. Ph.D. thesis, Johns Hopkins University.
54. Carlin MA, Patil K, Nemala SK, Elhilali M (2012) Robust phoneme recognition using biomimetic speech contours. In: Interspeech 2012, 13th Annual Conference of the International Speech Communication Association; 9–13 September 2012; Portland, Oregon.
55. Calabrese A, Schumacher JW, Schneider DM, Paninski L, Wooley SMN (2011) A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds. PLoS ONE 6: e16104.

56. Cadieu CF, Olshausen BA (2009) Learning transformational invariants from natural movies. Adv Neural Inf Proc Sys: 21: 209–216.

57. Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS, et al. (1993) TIMIT Acoustic-Phonetic Continuous Speech Corpus, Linguistic Data Consortium, Philadelphia.

58. (2006) The BBC Sound Effects Library Original Series, http://www.soundideas.com.

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

60. Chi T, Ru P, Shamma SA (2005) Multiresolution spectrotemporal analysis of complex sounds. J Acoust Soc Am 118: 887–906.

61. Theunissen FE, David SV, Singh NC, Hsu A, Vinje WE, et al. (2001) Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Neuron: Computation in Neural Systems 12: 289–316.

62. Rosen JB (1961) The gradient projection method for nonlinear programming: part II. Nonlinear constraints. J Soc Indust Appl Math 9: 514–532.

63. Luenberger DG (1969) Optimization by Vector Space Methods. New York: Wiley.

64. Horn BKP, Hilden HM, Negahdaripour S (1988) Closed form solution of absolute orientation using orthonormal matrices. J Optical Soc A 5: 1127–1135.

65. Horn RA, Johnson CR (1985) Matrix Analysis. Cambridge Univ. Press.

66. Ng AY, Jordan MI, Weiss Y (2002) On spectral clustering: analysis and an algorithm. Adv Neural Inf Proc Sys 14: 849–856.