CBMM Memo No. 060 January 26, 2017

Learning Mid-Level Auditory Codes from Natural Sound Statistics

by

Wiktor Młynarski, Josh H. McDermott

Abstract

Interaction with the world requires an organism to transform sensory signals into representations in which behaviorally meaningful properties of the environment are made explicit. These representations are derived through cascades of neuronal processing stages in which neurons at each stage recode the output of preceding stages. Explanations of sensory coding may thus involve understanding how low-level patterns are combined into more complex structures. Although models exist in the visual domain to explain how mid-level features such as junctions and curves might be derived from oriented filters in early visual cortex, little is known about analogous grouping principles for mid-level auditory representations. We propose a hierarchical generative model of natural sounds that learns combinations of spectrotemporal features from natural stimulus statistics. In the first layer the model forms a sparse convolutional code of spectrograms using a dictionary of learned spectrotemporal kernels. To generalize from specific kernel activation patterns, the second layer encodes patterns of time-varying magnitude of multiple first layer coefficients. Because second-layer features are sensitive to combinations of spectrotemporal features, the representation they support encodes more complex acoustic patterns than the first layer. When trained on corpora of speech and environmental sounds, some second-layer units learned to group spectrotemporal features that occur together in natural sounds. Others instantiate opponency between dissimilar sets of spectrotemporal features. Such groupings might be instantiated by neurons in the auditory cortex, providing a hypothesis for mid-level neuronal computation.

This work was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF - 1231216.

Learning Mid-Level Auditory Codes from Natural Sound Statistics

Wiktor Młynarski*, Josh H. McDermott

Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

mlynar@mit.edu

Abstract

Interaction with the world requires an organism to transform sensory signals into representations in which behaviorally meaningful properties of the environment are made explicit. These representations are derived through cascades of neuronal processing stages in which neurons at each stage recode the output of preceding stages. Explanations of sensory coding may thus involve understanding how low-level patterns are combined into more complex structures. Although models exist in the visual domain to explain how mid-level features such as junctions and curves might be derived from oriented filters in early visual cortex, little is known about analogous grouping principles for mid-level auditory representations. We propose a hierarchical generative model of natural sounds that learns combinations of spectrotemporal features from natural stimulus statistics. In the first layer the model forms a sparse convolutional code of spectrograms using a dictionary of learned spectrotemporal kernels. To generalize from specific kernel activation patterns, the second layer encodes patterns of time-varying magnitude of multiple first layer coefficients. Because second-layer features are sensitive to combinations of spectrotemporal features, the representation they support encodes more complex acoustic patterns than the first layer. When trained on corpora of speech and environmental sounds, some second-layer units learned to group spectrotemporal features that occur together in natural sounds. Others instantiate opponency between dissimilar sets of spectrotemporal features. Such groupings might be instantiated by neurons in the auditory cortex, providing a hypothesis for mid-level neuronal computation.

Introduction

Interaction with the environment requires an organism to infer characteristics of the world from sensory signals. One challenge is that the environmental properties an organism must recognize are usually not explicit in the sensory input. A primary function of sensory systems is to transform raw sensory signals into representations in which behaviorally important features are more easily recovered. To successfully infer the state of the world, the brain must generalize across irrelevant stimulus variation, while maintaining selectivity to the variables that matter for behavior. The nature of sensory codes and the mechanisms by which they achieve appropriate selectivity and invariance are thus a primary target of sensory system research.

The auditory system is believed to instantiate such representations through a sequence of processing stages extending from the cochlea into the auditory cortex. Existing functional evidence suggests that neurons in progressively higher stages of the auditory pathway respond to increasingly complex and abstract properties of sound [1–11]. Yet our understanding of the underlying transformations remains limited, particularly when compared to the visual system.

Feature selectivity throughout the auditory system has traditionally been described using linear receptive fields [12–14]. The most common instantiation is the spectrotemporal receptive field (STRF), which typically characterizes neural activity with a one-dimensional linear projection of the sound spectrogram transformed with a nonlinearity [15]. As a neural data analysis technique, STRFs are widespread in auditory neuroscience and have generated considerable insight in domains ranging from plasticity to speech coding (e.g. [16, 17]).

Despite their utility, it is clear that STRFs are at best an incomplete description of auditory codes, especially in the cortex [18–20]. Experimental evidence suggests that auditory neural responses are strongly nonlinear. As a consequence, auditory receptive fields estimated with natural sounds differ substantially from estimates obtained with artificial stimuli [21]. STRF descriptions also fail to capture the dimensionality expansion of higher representational stages. In contrast to the brainstem, neurons in the auditory cortex seem to be sensitive to multiple stimulus features at the same time [2,22,23]. The presence of strongly non-linear behavior and multiplexing necessitates signal models more sophisticated than one-dimensional, linear features of the spectrogram such as STRFs.

An additional challenge to characterizing mid-level features of sound is that humans lack strong intuitions about abstract auditory structure. By contrast, the study of the visual system has often been influenced by intuitions of how complex representations could emerge by combining lower-level features. For instance, elongated edges and curves, which drive neural responses in V2, can be thought of as conjunctions of the Gabor filters which match receptive fields of V1 neurons (e.g. [24, 25]). We know of few hypotheses for analogous auditory representations of intermediate complexity. In specific signal domains such as speech, progress has been made by cataloging phonemes and other frequently occurring structures, but it is not obvious how to generalize this approach to broader corpora of natural sounds.

An alternative approach to understanding sensory representations that is less reliant on domain-specific intuition is that of efficient coding [26,27]. The efficient coding hypothesis holds that neural codes should exploit the statistical structure of natural signals, allowing such signals to be represented with a minimum of resources. Numerous studies have demonstrated that tuning properties of neurons in early stages of the visual and auditory systems are predicted by statistical models of natural images or sounds [28–39]. Although the early successes of this approach engendered optimism, applications have largely been limited to learning a single stage of representation, and extensions to multiple levels of sensory processing have proven difficult. The underlying challenge is that there are many possible forms of high-order statistical dependencies in signals, and the particular dependencies that occur in natural stimuli are typically not obvious. The formulation of models capable of capturing these dependencies requires careful analysis and design [40–43], and perhaps good fortune, and is additionally constrained by what is tractable to implement. In the auditory system in particular, it remains to be seen whether modeling statistical signal regularities can reveal the complex acoustic structures and invariances that are believed to be represented in higher stages of the auditory system.

The primary goal of the present work was to discover such high-order structure in natural sounds and generate hypotheses about not-yet-observed intermediate-level neural representations. To this end we developed a probabilistic generative model of natural sounds designed to learn a novel stimulus representation - a population code of naturally occurring combinations of basic spectrotemporal patterns, analogous to spectrotemporal receptive fields (STRFs).

The resulting representations learned from corpora of natural sounds suggest grouping principles in the auditory system. In particular, a class of model units appears to encode opponency between different sets of features. These units were activated and inhibited by different types of natural stimulus features which do not typically occur together in natural audio. Although not yet described in the auditory system, such tuning patterns appear to be analogous to phenomena such as end-stopping or cross-orientation suppression in the visual system. The representations learned by our model also resemble some recently reported properties of auditory cortical neurons, providing further evidence that natural-scene statistics can predict neural representations in higher sensory areas.

Methods and Models

Overview of the hierarchical model

To learn mid-level auditory representations, we constructed a hierarchical, statistical model of natural sounds. The model structure is depicted in Fig 1. The model consisted of a stimulus layer and two latent layers that were adapted to efficiently represent a corpus of audio signals. Because our goal was to learn mid-level auditory codes, we did not model the raw sound waveform. Instead, we assumed an initial stage of frequency analysis, modeled after that of the mammalian cochlea. This frequency analysis results in a spectrogram-like input representation of sound, which we term a 'cochleagram,' that provides a coarse model of the auditory nerve input to the brain (Fig 1 A, bottom row). This input representation is an $F \times T$ matrix, where F is the number of frequency channels and T the number of time-points. Our aim was to capture statistical dependencies in natural sounds represented in this way.

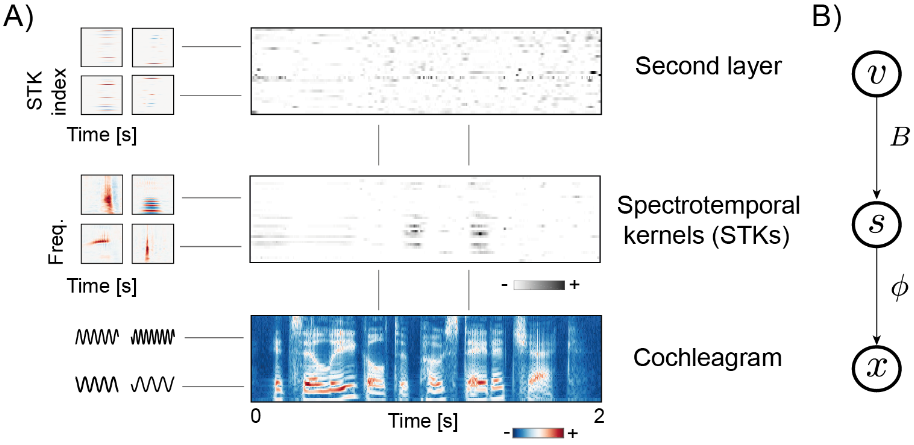

Figure 1: **Overview of the hierarchical model.** A) A spectrogram (bottom-row) is encoded by a set of spectrotemporal kernels (middle row). The features learned by the second layer encode temporal patterns of multiple STK activations. B) A graphical model depicting statistical dependencies among variables.

The first layer of the model (Fig 1, middle row) was intended to learn basic acoustic features, analogous to STRFs of early auditory neurons. Through the remainder of the paper, we refer to the features learned by the first layer as spectrotemporal kernels (STKs), in order to differentiate them from neurally derived STRFs. The second layer, depicted in the top row of Fig 1 A, was intended to learn patterns of STK co-activations that frequently occur in natural sounds.

The model specifies a probability distribution over the space of natural sounds, and its parameters can be understood as random variables whose dependency structure is depicted in Fig 1 B. The spectrogram x is represented with a set of spectrotemporal features ϕ convolved with the latent activation time-courses s. The second latent layer encodes the magnitudes of s with the basis functions B convolved with their activation time-courses v. In the following sections, we present the details of each layer.

First layer of model - convolutional, non-negative sparse coding of spectrograms

The first layer of the model was designed to learn basic spectrotemporal features. One previous attempt to learn sparse spectrotemporal representations of natural sounds [35] produced structures reminiscent of receptive fields of neurons in the auditory midbrain, thalamus, and cortex. Here, we extended this approach by learning a spectrogram representation which is sparse and convolutional.

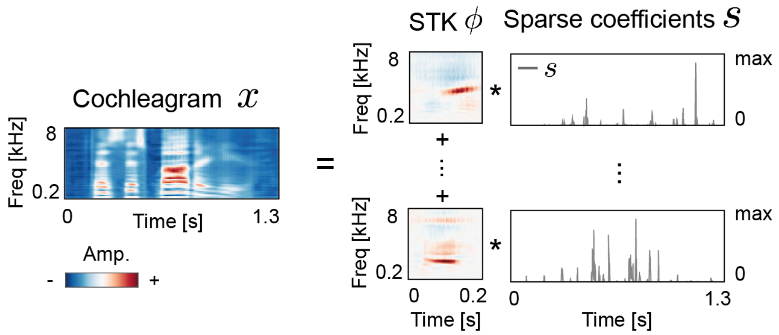

Figure 2: **Explanation of first layer of the model.** A cochleagram x of arbitrary length is represented as a sum of spectrotemporal kernels ϕ convolved with time-courses of corresponding coefficients s. Coefficients s are non-negative and have sparse distributions (i.e. remain close to zero most of the time).

A schematic of the first layer is depicted in Fig 2. We modeled the cochleagram $(x_{t,f})$ as a linear combination of spectrotemporal kernels ϕ convolved with their activation time-courses s and distorted by additive Gaussian noise $\xi_{t,f}$ with variance σ^2.:

$$\hat{x}_{t,f} = \left[\sum_{i=1}^{N} \phi_{i,f} * s_i\right]_t \tag{1}$$

$$x_{t,f} = \hat{x}_{t,f} + \xi_{t,f} \tag{2}$$

Kernel activations $s_{i,t}$ are assumed to be independent, i.e. their joint distribution is equal to the product of marginals:

$$p(\mathbf{s}) = \prod_{i=1}^{N}\prod_{t=1}^{T} p(s_{i,t}|\lambda_i) \tag{3}$$

We assumed that each spectrotemporal kernel remains inactive for most of the time, i.e. that the distribution of its activations is sparse. Moreover, we imposed a non-negativity constraint on the coefficients s. This facilitates interpretations in terms of neural activity and improves the interpretability of the learned representation. These constraints were embodied in an exponential prior on the coefficients s:

$$p(s_{i,t}|\lambda_i) = \frac{1}{\lambda_i}\exp\left[-\frac{s_{i,t}}{\lambda_i}\right] \tag{4}$$

where λ_i is the scale parameter.

The first layer of the model specifies the following negative log-posterior probability of the data:

$$E_1 \propto \frac{1}{\sigma^2}\sum_{f=1}^{F}\sum_{t=1}^{T}(\hat{x}_{t,f} - x_{t,f})^2 + \sum_{i=1}^{N}\frac{1}{\lambda_i}\sum_{t=1}^{T} s_{i,t} \tag{5}$$

This negative log-probability can be viewed as a cost function to be minimized when inferring the value of coefficients s: while maintaining low-reconstruction error (first term on the right-hand side), the sparsity of representation should be maximized (the second term).

Dependencies between coefficients - a signature of mid-level structure

Although the sparse coding strategy outlined above learns features that are approximately independent across the training set, residual dependencies nonetheless remain. In part this is because not all dependencies can be modeled with a single layer of convolutional sparse coding. However, dependencies also result from the non-stationary nature of natural audio. For particular sounds the learned features exhibit dependencies [42], and thus deviate from their (approximately independent) marginal distribution. For example, a spoken vowel with a fluctuating pitch contour would require many harmonic STKs to become activated, and their activations would become strongly correlated on a local time scale. Such local correlations reflect higher-order structure of particular natural sounds. Statistically speaking, this is an example of marginally independent random variables exhibiting conditional dependence (in this case conditioned on a particular point in time or a type of sound).

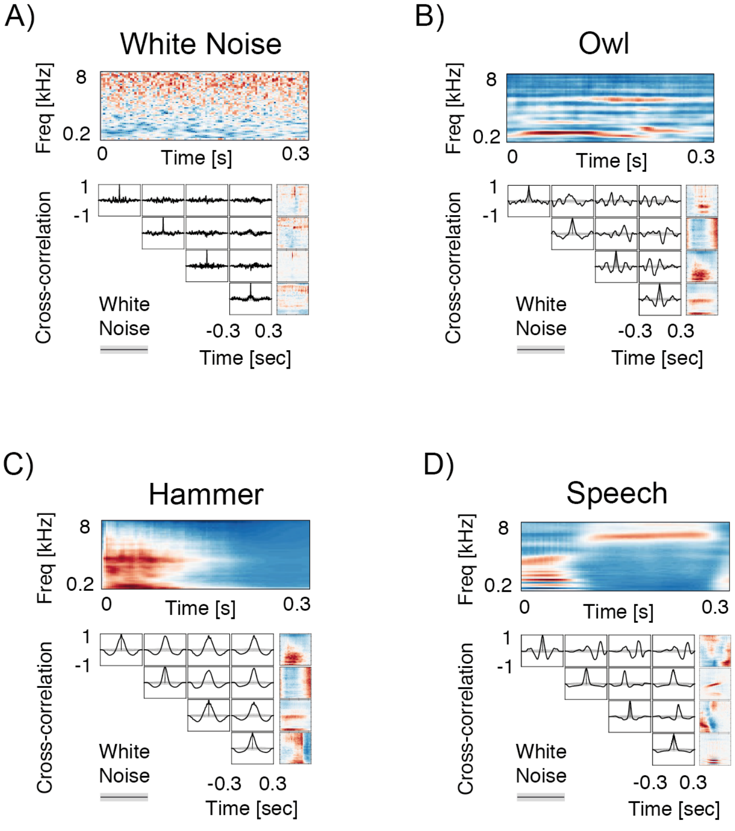

Figure 3: **Dependencies among spectrotemporal feature activations.** A) When encoding white noise (cochleagram depicted on top), sparse coefficients are uncorrelated on short time-scales. This is visible in the shape of the coefficient cross-correlation functions (black lines) corresponding to the four most strongly activated STKs (bottom, right column). Due to the lack of temporal structure, the cross-correlation between STKs is flat, and the autocorrelation of individual STKs is a Dirac delta function at 0. Here and in other panels, the mean was substracted from the (non-negative) coefficient trajectories prior to computing the cross-correlation. B) When encoding structured stimuli such as an owl vocalization, STK activations reveal strong local correlations - the cross-correlations deviate from those for white noise (thick gray lines). C) Same as B, for a hammer hit. D) Same as B, for a speech excerpt.

Fig 3 depicts such dependencies via cross-correlation functions of selected STK activations for a white noise sample and for three different natural sounds. Coefficient correlations (black curves in each subplot) vary from sound to sound, but in all cases deviate from those obtained with noise (gray bars within subplots), revealing dependence. These dependencies are indicative of "mid-level" auditory features, perhaps analogous to the correlations between oriented Gabor filters induced by an elongated edge. In this work, we exploited the fact that intermediate level representations can be learned by modeling dependencies among first-layer features [41–43].

Variability of STK activations

A second phenomenon evident in STK activations is that particular patterns of co-activation occur with some variability. This is visible in Fig 4, which depicts activations of selected STKs when encoding multiple exemplars of the same sound - the word "one" spoken twice by the same speaker (Fig 4A) and two exemplars of water being poured into a cup (Fig 4B).

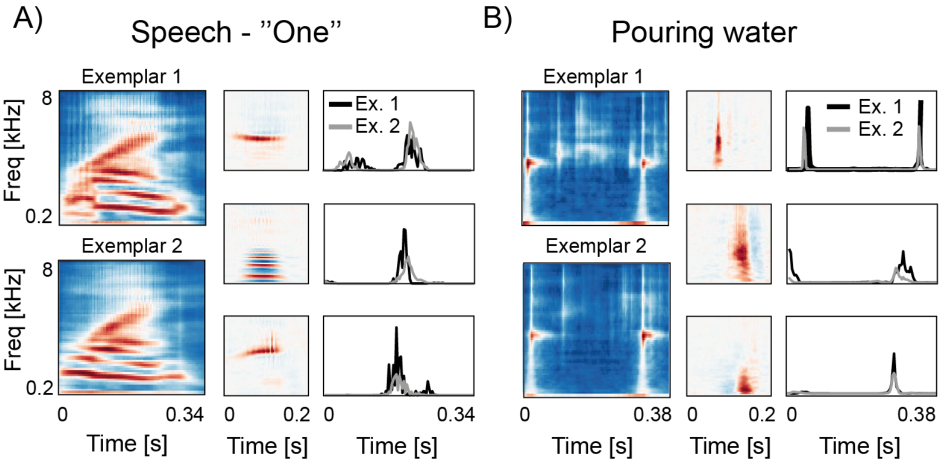

Figure 4: **Variability of spectrotemporal feature activations.** A) Coefficient trajectories of spectrotemporal features ϕ are depicted for two utterances of word "one" spoken by the same speaker. Although both coefficient trajectories (gray and black lines) exhibit the same global structure, they are not identical. B) Coefficient variability visualized in the same way for two different examples of water pouring into a cup.

It is apparent that the STK coefficient trajectories for the two exemplars in each case (black and gray lines) reflect the same global pattern even though they differ somewhat from exemplar to exemplar. The similarity suggests that the trajectories could be modeled as different samples from a single time-varying distribution parameterized by a non-stationary coefficient magnitude. When the magnitude increases, the probability of a strong STK activation increases. Retaining the (inferred) time-varying magnitude instead of precise values of STK coefficients would yield a representation more invariant to low-level signal variation, potentially enabling the representation of abstract regularities in the data. Such a representation bears an abstract similarity to the magnitude operation used to compute a spectrogram, in which each frequency channel retains the time-varying energy in different parts of the spectrum. Here we are instead estimating a scale parameter of the underlying distribution, but the process similarly discards aspects of the fine detail of the signal.

Second layer of model - encoding of STK combinations

The second layer of the model was intended to exploit the two statistical phenomena detailed in the previous section: conditional dependencies between STKs and their variation across exemplars. Similarly to the first layer, the second layer representation is formed by a population of sparsely activated basis functions. These basis functions capture local dependencies among STK magnitudes by encoding the joint distribution of STK activations rather than exact values of STK coefficients. The resulting representation is thus more specific than the first-layer code - instead of encoding single features independently, it signals the presence of particular STK combinations. It is also more invariant, generalizing over specific coefficient values.

Because the proposed representation is a population code of a distribution parameter, it bears conceptual similarity to previously proposed hierarchical models of natural stimuli that encoded patterns of variance [41,44], covariance [42] or complex amplitude [40,43]. The novelty of our model structure lies in being convolutional (i.e., it can encode stimuli of arbitrary length using the same representation) and in parameterizing distributions of non-negative STK coefficients, increasing the interpretability of the learned spectrogram features. The novelty of the model's application is to learn hierarchical representations of sound (previous such efforts have largely been restricted to modeling images).

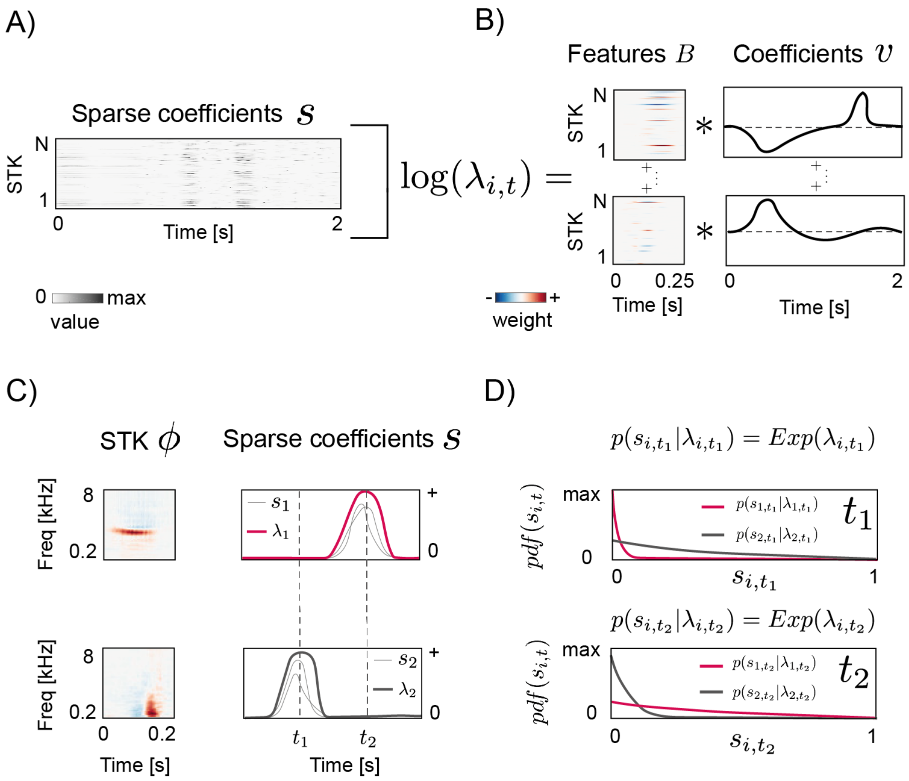

Figure 5: **Explanation of second layer of the model.** A) An array of STK activations s (a "STK-gram") serves as an input to the second layer. Rows correspond to first layer features ϕ_i and columns to time points. B) The second layer uses a population of features B to encode the logarithm of STK activation magnitudes. C) Coefficient trajectories s (thin grey lines) and their magnitudes λ (thick red and black lines) for two example STKs. D) Distributions of s at time points t_1 and t_2 are depicted in the right column.

The second layer of the model is depicted schematically in Fig 5A. We assume that STK activations s are samples from a non-stationary exponential distribution with time-varying scale parameter $\lambda_{i,t}$, relaxing the assumption of stationary λ_i made in learning the first layer:

$$p(s_{i,t}) = Exp(\lambda_{i,t}) = \frac{1}{\lambda_{i,t}} \exp\left[-\frac{s_{i,t}}{\lambda_{i,t}}\right] \tag{6}$$

When $\lambda_{i,t}$ is high, the distribution of $s_{i,t}$ becomes heavy tailed (Fig 5C, black line in the top row, red line in the bottom). This allows the coefficient s_i to attain large values. For small values of the scale parameter, the probability density is concentrated close to 0 (Fig 5D), red line in the top row, black line in the bottom), and coefficients $s_{i,t}$ become small. To model the magnitudes λ (which are non-negative, akin to variances), we took their logarithm, mapping their values onto the entire real line so that they could be represented by a sum of real-valued basis functions.

Patterns of STK magnitudes are represented in the second layer by a population of features B convolved with coefficients v:

$$\lambda_{i,t} = \exp\left[\sum_{j=1}^{M} B_{j,i} * v_j + \rho_i\right]_t \tag{7}$$

where ρ is a bias vector. Each second-layer basis function B represents a particular temporal pattern of co-activation of first-layer STKs. Their corresponding coefficients v are assumed to be sparse and independent:

$$p(\mathbf{v}) = \prod_{j=1}^{M} \prod_{t=1}^{T} p(v_{j,t}) \tag{8}$$

$$p(v_{j,t}) \propto \exp\Big(-\alpha |v_{j,t}| \Big) \tag{9}$$

where α controls the degree of sparsity.

As with the first layer, learning and inference are performed by gradient descent on the negative log posterior. Because the second layer units encode combinations of sparse first-layer coefficients, we placed a sparse prior on the L_1 norm of the basis functions B. The overall cost function to be minimized during learning in the second layer is then:

$$E_2 \propto \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{s_{i,t}}{\lambda_{i,t}} + \log(\lambda_{i,t}) + \alpha \sum_{j=1}^{M} \sum_{t=1}^{T} |v_{j,t}| + \beta \sum_{j=1}^{M} \sum_{i=1}^{N} \sum_{t_b=1}^{T_b} |B_{j,i,t_b}| \tag{10}$$

where β controls the strength of the sparse prior on B, and T_b is the temporal extent of each second-layer basis function. As in the first layer cost function (Eq 5), the first term on the right hand side of Eq 10 enforces a match of the representation to data (the magnitudes λ are pushed away from zero towards the observed coefficients s), while second and third terms promote sparsity of second-layer coefficients and basis functions, respectively.

Learning procedure

The two layers of the model were trained separately, i.e. the training of the second layer occurred after the first layer training was completed. In the first layer, training was performed with an EM-like procedure that iteratively alternated between inferring STK coefficients and updating STK features [29, 43]. Spectrotemporal features ϕ were initialized with Gaussian white noise. For each excerpt in the training set, coefficients $s_{i,t}$ were inferred via gradient descent on the energy function (5). Because inference of all coefficients s is computationally expensive, we adopted an approximate inference scheme [45]. Instead of inferring values of all coefficients for each excerpt, we selected only a subset of them to be minimized. This was done by computing the cross-correlation between a sound excerpt and features ϕ_i and selecting a fixed number of the largest coefficients $s_{i,t}$. The inference step adjusted only this subset of coefficients while setting the rest to 0. Given the inferred coefficients, a gradient step on the spectrotemporal features ϕ was performed.

Each learning iteration therefore consisted of the following steps:

1. Draw a random sound excerpt from the training data set. Excerpts were 403 ms in length (129 time samples of the spectrogram, sampled at 320 Hz).

2. Compute the cross-correlation of all basis functions ϕ_i with the sound excerpt. Select the 1024 pairs of coefficient indices and time-points (i, t) that yield the highest correlation values.

3. Infer the values of the selected coefficients by minimizing Eq. 5 with respect to $s_{i,t}$ via gradient descent. Set the rest of coefficients to 0.

4. Compute the gradient step on the basis functions as the derivative of Eq. 5 with respect to basis functions ϕ using inferred coefficient values $\hat{s}$. Update basis functions according to the gradient step.

5. Normalize all basis functions to unit norm.

This procedure was terminated after 200000 iterations.

The second layer was then learned via the same procedure used for the first layer. In each iteration a 528 ms long (169 samples at 320 Hz) randomly drawn sound excerpt was encoded by the first layer, and the resulting matrix of coefficients s served as an input to the second layer. A subset of coefficients v was selected for approximate inference by computing the cross correlation between features B_i and the logarithm of the first-layer coefficients s (analogous to step 2 in the procedure described above for the first layer). The energy function E_2 was first minimized with respect to coefficients v followed by a gradient update to the basis functions B (analogous to steps 3 and 4 for the first layer). Entries in the bias vector ρ corresponding to each coefficient s_i were set to the expectation of the coefficient across the entire training set: $\rho_i = \mathbb{E}_t[s_{i,t}]$ (i.e., the estimate of the marginal scale parameter λ_i for the corresponding STK). Learning was again terminated after 200000 iterations.

Training data and spectrogram parameters

We trained the model on two different sound corpora. The first corpus was the TIMIT speech database [46]. The second corpus combined a set of environmental sounds (the Pitt sound database [32]) and a number of animal vocalizations downloaded from freesound.org. The environmental sounds included both transient (breaking twigs, steps, etc.) and ambient (flowing water, wind, etc.) sounds; the animal vocalizations were mostly harmonic.

We computed cochleagrams by filtering sounds with a set of 65 bandpass filters intended to mimic cochlear frequency analysis. Filters were equally spaced on an equivalent rectangular bandwidth (ERB) scale [47]), with parameters similar to that from a previous publication [48]. Center frequencies ranged from 200 Hz to 8 kHz. We computed the Hilbert envelope of the output of each filter and raised it to the power of 0.3, emulating cochlear amplitude compression [49]. To reduce dimensionality, each envelope was downsampled to 320 Hz.

We set the number of features in the first layer to 128 and in the second layer to 100. Pilot experiments yielded qualitatively similar results for alternative feature dimensionalities. Each first layer feature encoded a 203 ms interval (65 time samples of the spectrogram).

Model Analysis

In this section we describe the methods used to analyze features learned by the model and to compare them with experimental data. The details here are not critical to understanding the central points of the paper.

Feature Selectivity Index

For comparison with neural data, we computed the feature selectivity index (FSI) proposed in [50]. The FSI is a number lying in the $[0, 1]$ interval. FSI values close to 1 imply that stimuli eliciting a response of a neuron (or, in our case, a model unit) are similar to each other (specifically, the stimuli are close to the mean stimulus eliciting a response). When the FSI value is close to 0, the corresponding neuron spikes at random i.e. stimuli preceding spikes are uncorrelated with the spike-triggered average. The relevance of the FSI for our purposes is that a neuron or model unit that exhibits invariance to some type of stimulus variation should have an FSI less than 1. By comparing the FSI across layers we hoped to quantify differences in the degree of representational abstraction.

The FSI computation procedure are described in detail in [50]. The only discrepancy between the use of the FSI here and its prior use in neurophysiological studies is that units in our model are continuously active, and do not discretely spike. We emulated the selection of stimuli eliciting spikes by selecting the stimulus excerpts yielding the highest activation of model layer units. In the first layer, we selected the 25 stimuli yielding highest positive activation. In the second layer, we separately computed FSI indices for stimuli eliciting positive and negative responses, using 25 stimuli per unit in each case. We then averaged the indices obtained for the two sets of stimuli for each unit.

Computing FSI for an i-th unit consists of the following steps:

1. Compute average of strongly activating stimuli (analogous to spike-triggered average - STA) - separately for positive and negative stimuli in the case of second-layer units.

2. Compute correlations $c_{S,i,n}$ between each strongly activating stimulus n and its respective STA.

3. Compute correlations $c_{R,i,n}$ between the STA and a randomly selected subset of stimuli n.

4. Compute the area $A_{S,i}$ under the empirical cumulative distribution function of correlations $c_{S,i}$, $A_{S,i} = \int_{-1}^{1} ECDF(c_{S,i}) dc_{S,i}$

5. Compute area $A_{R,i}$ under the empirical cumulative distribution function of $A_{R,i} = \int_{-1}^{1} ECDF(c_{R,i})dc_{R_i}$

6. The FSI of each unit is defined as: $FSI_i = \frac{(A_{R,i} - A_{S,i})}{A_{R,i}}$

Overlap of excitatory and inhibitory stimuli in the second layer

As will be described in the Results section, it became clear upon examining the model representations that there were interesting relationships between stimuli eliciting positive and negative responses in second-layer units. To quantify the overlap of the distributions of stimuli eliciting strong positive and negative responses, we estimated Bhattacharyya coefficients [51]. A Bhattacharrya BC coefficient measures the overlap of two probability distributions p and q and is defined as following:

$$BC(p,q) = \int \sqrt{p(x)q(x)}dx \tag{11}$$

The BC is a real number lying in the $[0,1]$ interval, where 1 corresponds to complete alignment of two distributions, while 0 implies that the support of the distributions do not overlap at all. Small BC values thus imply high stimulus separability - samples generated from non-overlapping distributions lie far apart and can be easily separated into distinct classes.

To estimate how strongly positive and negative stimuli of second-layer units are separated we computed the BC between distributions fitted to each stimulus class. First, we computed centers of mass of each stimulus in the time-frequency and modulation planes, representing each stimulus as a point in these two planes. We then fitted two-dimensional Gaussian distributions to the sets of points corresponding to positive and negative stimuli, separately for each plane, and computed the Bhattacharryya coefficient between the distributions for positive and negative stimuli. The Bhattacharryya coefficient of two Gaussian distributions with respective mean vectors μ_1, μ_2 and covariance matrices C_1, C_2 has the following closed form:

$$BC = \exp - \left[\frac{1}{8}(\mu_1 - \mu_2)^T C^{-1}(\mu_1 - \mu_2) + \frac{1}{2}\log\left(\frac{\det C}{\sqrt{\det C_1 \det C_2}}\right)\right] \tag{12}$$

where $C = \frac{C_1 + C_2}{2}$.

Results

First layer: Basic spectrotemporal features of natural sounds

The first-layer features learned from each of the two sound corpora are shown in Fig 6 A and B. These features could be considered as the model analogues of neural STRFs. The vast majority of features are well localized within the time-frequency plane, encoding relatively brief acoustic events. The STKs learned from speech included single harmonics and harmonic "stacks" (Fig 6 A - features numbered 1 and 2), frequency sweeps (feature 3), and broadband clicks (feature 4). The features learned from environmental sounds also included single harmonics and clicks (Fig 6 B - features 1, 2 and 4). In contrast to the results obtained with speech, however, harmonic stacks were absent, and a number of high-frequency hisses and noise-like features were present instead (feature 3).

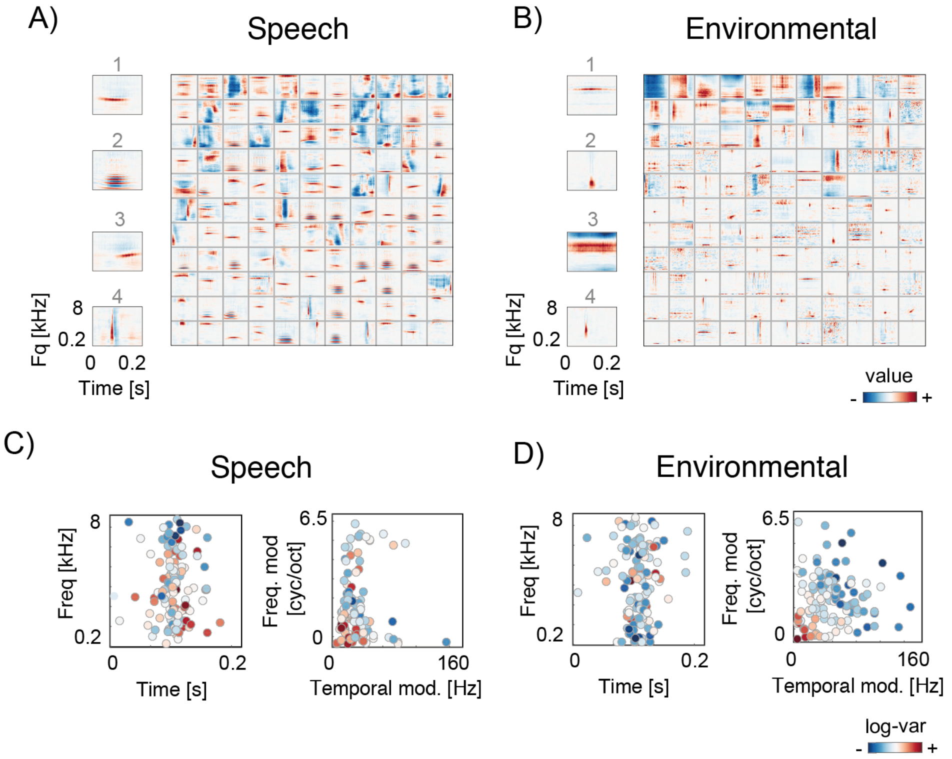

Figure 6: **Spectrotemporal kernels learned by the first layer.** A) A population of STKs learned from a speech database. Representative STKs for each corpus are magnified and numbered from 1-4 for ease of reference. B) Population of STKs learned from environmental sounds. C) Speech-trained STK population plotted on time-frequency (left) and spectral-temporal modulation (right) planes. Each dot corresponds to a single STK, and its color encodes the log of its mean coefficient (averaged over entire training dataset). D) The STK population trained on environmental sounds, represented as in (C).

The properties of the learned dictionaries are also reflected in distributions of feature locations in the time-frequency and spectrotemporal modulation planes, as visible in Fig 6C and D. Each dot position denotes the center of mass of a single STK, while its color signals the feature's average coefficient value over the stimulus set. For both speech and environmental sounds, the learned STKs uniformly the audio frequency spectrum (Fig 6 C and D, left panels). Due to the convolutional nature of the code, the energy of each feature is concentrated near the middle of the time axis. The modulation spectra of the learned STKs (Fig 6 C and D, right panels) are somewhat specific to the sound-corpus. Features trained on a speech corpus were more strongly modulated in frequency, while environmental sounds yielded STKs with faster temporal modulations. STKs learned from both datasets exhibit a spectrotemporal modulation tradeoff: if a STK is strongly temporally modulated its spectral modulation tends to be weaker. This tradeoff is an inevitable consequence of time-frequency conjugacy [52], and is also found in the STRFs of the mammalian and avian auditory systems [17, 53].

Second layer: Combinations of spectrotemporal features

STKs captured by the first layer of the model reflect elementary features of natural sounds. By contrast, the features learned by the second layer capture how activations of different STKs cluster together in natural sounds, and thus reflect more complex acoustic regularities. We first present several ways of visualizing the multi-dimensional nature of the second-layer representation, then make some connections to existing neurophysiological data, and then derive some neurophysiological predictions from the model.

Visualizing second-layer features

The second-layer features encode temporal combinations of STK log-magnitudes. Each feature can be represented as a $N \times T_b$ dimensional matrix, where rows correspond to STKs and columns to time-points. An example second-layer unit is shown in Fig 7A (left panel). A positive value in the i-th row and t-th column of a feature B_j encodes a local increase in the magnitude of the i-th STK. A negative value encodes a decrease in magnitude.

An alternative visualization is to examine the spectrotemporal structure of the STKs that have large weights in the second-layer feature. The same feature B_j is depicted in this way in the center panel of Fig 7A, which displays the four STKs with highest average absolute weights for this particular feature. To the right of each STK are its weights (i.e., the corresponding row of the B_j matrix). It is apparent that the weights increase and decrease in a coordinated fashion, and thus likely encode particular dependencies between the STKs.

To summarize the full distribution of STKs contributing to a second-layer unit, we adopted the visualization scheme illustrated in the right panel of Fig 7A. We plot the center of mass of each STK in the modulation (top row) and time-frequency (middle row) planes, as in Fig 6C and D. The dot for a first-layer STK is colored red or blue, depending on the sign of their time-averaged weight, with the average absolute value of the weight signaled by the intensity of the color. The bottom row of the panel depicts the temporal pattern of STK magnitudes - line colors correspond to dots in the top and middle rows of the panel. Although the weights of most STKs maintain the same sign over the temporal support of the second-layer unit, there was not constraint enforcing this, and in some cases the weight trajectories cross zero.

Representative examples of second-layer basis functions are depicted in this way in Fig 7B. We separated them into two broad classes - "excitatory" units (columns 1 and 2 in Fig 7), which pool STKs using weights of the same sign, and therefore encode a pattern of coordinated increase in their magnitudes, and "excitatory-inhibitory" units (columns 3 and 4 in Fig 7) which pool some STKs with positive average weights and others with negative average weights. We note that excitatory-only and inhibitory-only units are functionally interchangeable in the model, because the encoding is unaffected if the sign of both the STK weights and coefficients are reversed.

As is apparent in Fig 7B, second-layer basis functions tend to assign weights of the same sign to STKs with similar frequency or modulation characteristics. We quantified this trend by computing distances between pairs of STKs with large weights of the same sign compared to pairs chosen randomly. Specifically, we measured the distance between STK centers of mass on the time-frequency or modulation planes between each STK and its nearest neighbor within the set of STKs with weights above a threshold (5% of the maximum for the unit). For comparison we computed the same distances but for randomly selected STK subsets of the same size. We found this distances between pooled STKs to be significantly lower than if pooling was fully random ($p \ll 0.001$; obtained via t-test).

Example units pooling similar STKs are depicted in panels A2 and C2. They encode joint increases in the magnitude of high-frequency and low-frequency STKs, respectively. The unit in panel B3 encodes an increase in magnitude of temporally modulated features (clicks) together with a simultaneous decrease of activation of a strongly spectrally modulated (harmonic) STK.

From inspection of Fig 7B it is also evident that some second-layer units pool only a few STKs (e.g. A1, C1, D1) while others are more global and influence activations of many first-layer units (e.g. A2, A3, C3).

Second layer features encode patterns of STK dependencies

To better understand the structure captured by the second-layer units, we considered their relationship to the two kinds of dependencies in STK activations that initially motivated the model. In sections , we observed that STK activations to natural sounds exhibit strong local cross-correlations as well as variations of particular temporal activation patterns. The second layer of the model was formulated in order to capture and encode these redundancies.

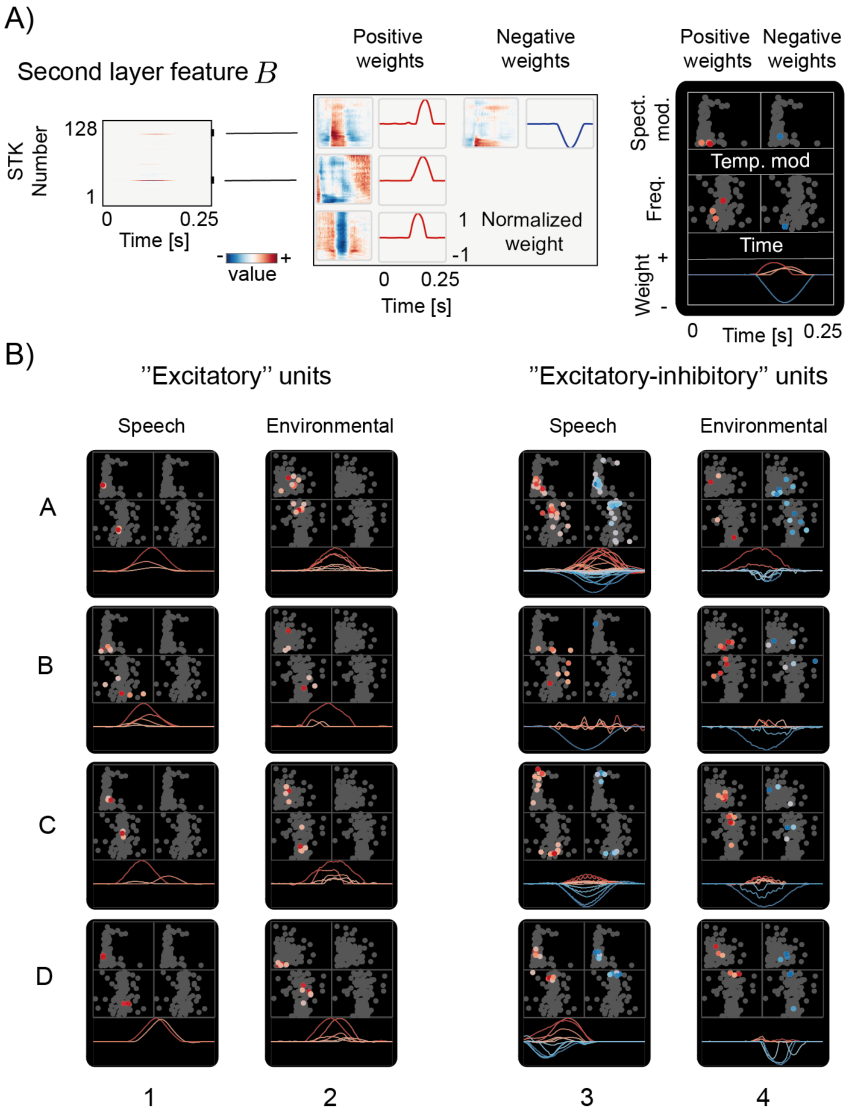

Figure 7: **Second layer model features.** A) Feature visualizations. Left panel - the first layer STK weights for an example second layer feature (representing magnitudes of each STK over a time window). Middle panel - STKs whose weights in the same second-layer feature deviate most strongly from 0 are displayed along with their weight profile over time. Right panel - STKs are plotted as dots in the modulation and time-frequency planes, with the dot location indicating the STK center of mass in the plane, and the color indicating the weight sign and magnitude (red denoting positive and blue denoting negative). STKs are divided into those with positive and negative weights for clarity. Bottom row visualizes temporal trajectories of STK weights for the feature. B) Examples of learned second-layer features. First two columns (labeled 1 and 2) depict units with positive ("excitatory") weights only. Last two columns (labeled 3 and 4) depict units that pool features with both positive and negative weights.

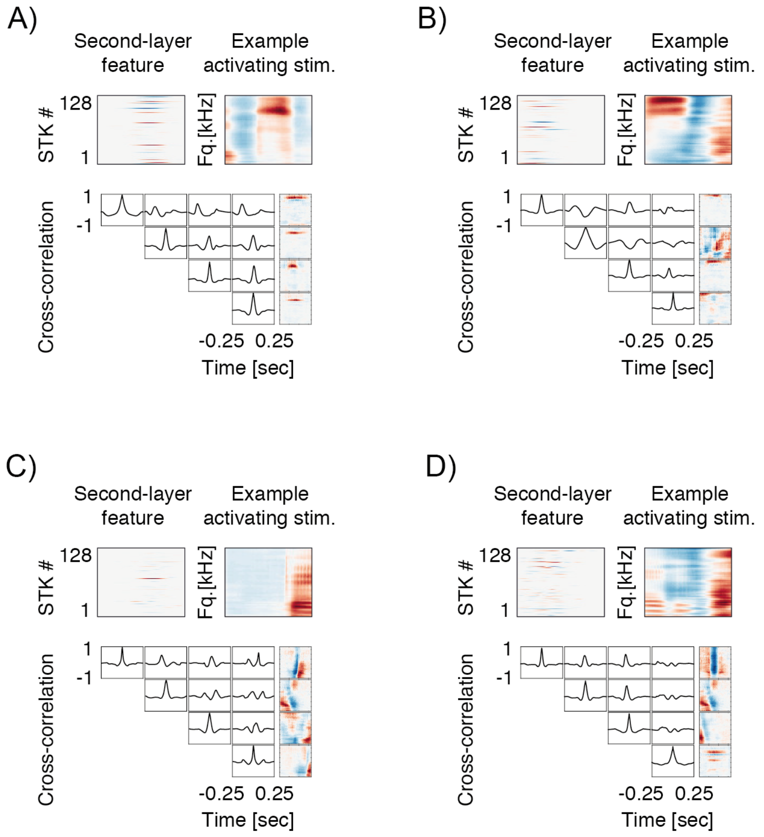

Figure 8: **Second-layer units respond to specific STK cross-correlation patterns.** A) A second-layer unit (top row, left column) and an example sound excerpt eliciting a strong positive response in the unit (top row, right column). Bottom shows cross-correlation functions of coefficient trajectories for four STKs. The STKs selected were those with the largest weights for this second-layer unit. Cross-correlations were averaged across 25 stimuli eliciting the strongest positive response of the second-layer unit across a large subset of the TIMIT corpus. B, C, D) same as A for three other second-layer units.

To first test whether the second layer captures the sorts of residual correlations evident in the first layer output, we measured cross-correlations between STK activations conditioned on the activation of particular second-layer units. Fig. 8A-D depicts four example second-layer units (top row, left column) along with an example stimulus that produced a strong positive response in the unit (selected from the TIMIT corpus). The bottom section of each panel depicts cross-correlation functions of activations of four STKs, averaged over 25 stimulus epochs that produced a strong positive response of the second-layer unit. The cross-correlation functions deviate substantially from 0, as they do when conditioned on excepts of natural sounds (Fig. 3). These correlations reflect the temporal pattern of STK coefficients that the second-layer unit responds to.

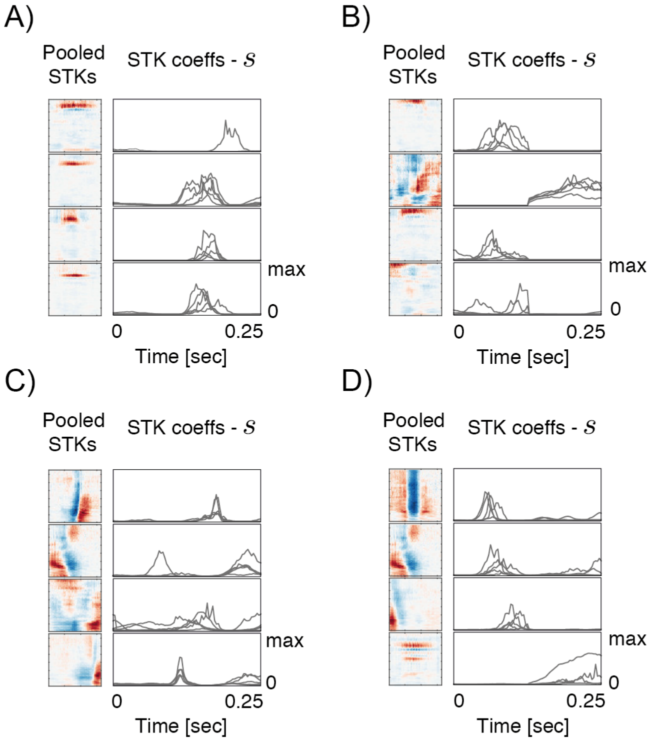

Figure 9: **Second-layer units generalize across variation in STK coefficients.**Panels A-D correspond to second-layer units depicted in Fig. 8 Each panel shows four STKs pooled with strongest weights (left column) by the corresponding second-layer unit. Next to each STK are their activation patterns (right column) for the 5 stimuli eliciting the strongest response in the second-layer unit. Despite some variability they share a global trend.

We next examined whether the second-layer units respond to STK activations fluctuating around particular global patterns (as depicted in Fig.4). Because the second layer of the model represents the magnitude, rather than the precise values of first-layer coefficients, it should be capable of generalizing over minor STK coefficient variation. Fig.9 plots STK coefficient trajectories for stimuli eliciting a strong response in the second-layer features from Fig. 8. The STK activation traces reveal variability in each case, but nonetheless exhibit a degree of global consistency, as we saw earlier for natural sound exemplars (Fig.4). These results provide evidence that the second layer is capturing the dependencies it was intended to model.

Comparison with neurophysiological data

Although our primary goal was to generate predictions of not-yet observed neural representations of sound, we first sought to test whether our model would reproduce known findings from auditory neuroscience. The first layer STKs replicated some fairly standard findings in the STRF literature, as discussed earlier. To compare the results from the second layer to experimental data, we examined their receptive field structure and the specificity of their responses, and compared each to published neurophysiology data.

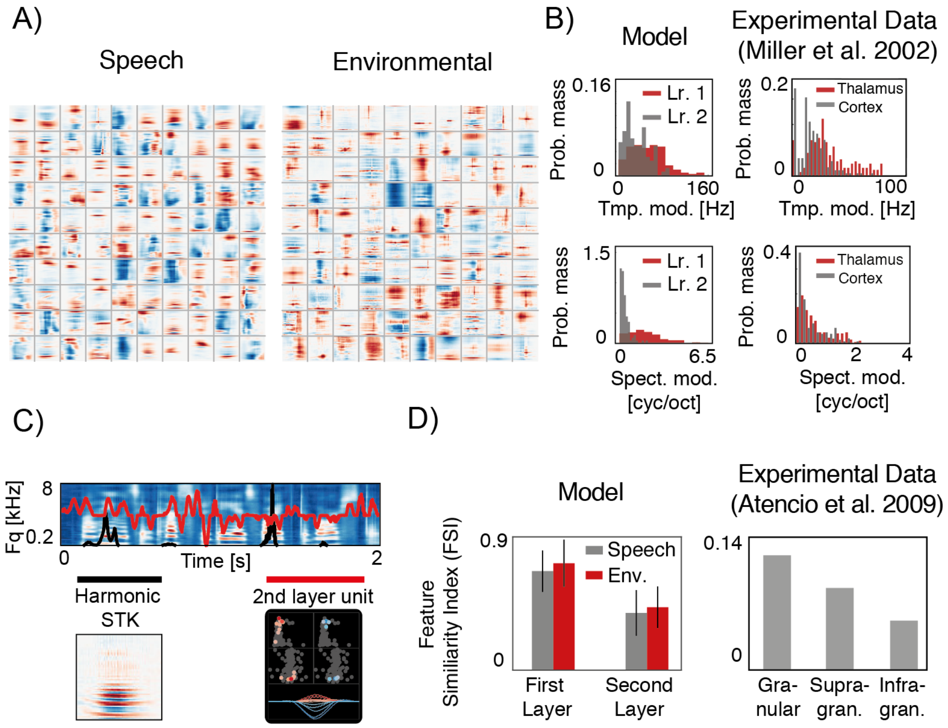

Figure 10: **Comparison with neurophysiological data.** A) Spectrotemporal receptive fields estimated for second layer units trained on speech and environmental sounds. B) Comparison of spectral and temporal modulation tuning of first and second layer receptive fields (panels in left column) to experimental measurements in auditory thalamus and cortex of the cat (panels in right column, courtesy of Lee Miller [53]). Higher processing stages both in the model and the auditory system exhibit tuning for coarser modulations in both frequency and time. Subpanels with experimental data reprinted with permission of original author. C) Example activation trajectories of first (black line) and second layer (red line) units to an excerpt of speech. The activation of the first layer unit is tightly locked to presence of a preferred stimulus. Responses of the second layer unit are less specifically locked to particular spectrotemporal structures. D) Comparison of tuning specificity, as measured with the Feature Selectivity Index (FSI), for first and second layers of the model trained either on speech or environmental sounds (gray and red bars, respectively), and for different layers of the auditory cortex of the cat (right panel replotted from [54]).

First, we estimated spectrotemporal receptive fields for units in both layers of the model. The receptive fields of the first-layer units are simply the STK of the unit. To estimate receptive fields of a second-layer unit, we drew inspiration from the spike-triggered average, generating a number of cochleagram samples from each basis function and averaging them. To generate samples from a $j-$th basis function, we set a coefficient v_j to 1 with all other $v_{i \neq j}$ set to zero. We then sampled STK activation trajectories from the distribution dictated by the second-layer feature's coefficient and weights, convolved them with the corresponding STKs and summed the results. We then averaged multiple such samples together. Although we could have computed something more directly analogous to a spike-triggered average, the average sample (possible only because we have the underlying generative model, unlike when conducting a neurophysiology experiment) has the advantage of alleviating the influence of stimulus correlations on the signature of the receptive field.

"Receptive fields" obtained in this way are depicted in Fig 10A. To compare the model units to neurophysiological data, we generated histograms of average spectral and temporal modulation frequency (center of mass in the modulation plane) of first- and second-layer receptive fields and plot them next to distributions of preferred modulation frequencies of neurons in the auditory thalamus and cortex of the cat [53] (Fig 10B). The same trend is evident in the model and the auditory system: the second-layer prefers features with slower/coarser spectral and temporal modulations relative to the first-layer, mirroring the difference seen between the cortex and thalamus. Lower modulation frequencies may result from combining multiple distinct STKs in downstream units. This analysis used features trained on speech, but environmental sounds yielded qualitatively similar results.

Neuronal tuning in early and late stages of the auditory system also tends to differ in specificity [1, 22]. Compared to the auditory brainstem, cortical neurons are less selective and respond to multiple features of sound [2,22], consistent with an increase in abstraction of the representation [1]. Suggestions of similar behavior in our model are apparent in the activations of first- and second-layer units to sound, and example of which is shown in Fig 10C. The first layer feature (black line) becomes activated only when it is strongly correlated with the stimulus. In contrast, activations of a typical second layer feature (red line) deviate from zero during many, seemingly different parts of the stimulus.

We quantified the specificity of tuning with the feature selectivity index (FSI), a measure introduced previously to quantify how correlated a stimulus has to be with a neuron's spike-triggered average to evoke a response [50]. An FSI equal to 1 implies that a neuron spikes only when a stimulus is precisely aligned with its STRF (defined as the spike-triggered average), whereas an FSI equal to 0 means that neural firing is triggered by stimuli uncorrelated with the STRF. We computed the FSI using the 25 cochleagram excerpts that most strongly activated each of the first- and second-layer units. The average FSI of each model layer is plotted in Fig 10D. Second-layer features are substantially less specific than first-layer features. A similar effect occurs across different cortical layers [54] (Fig 10D, right). Analogous differences seem likely to occur between thalamus and cortex as well, although we are not aware of an explicit prior comparison. The decrease in response specificity in our model can be explained by the fact that second layer units can become activated when any of the pooled first-layer features (or their combination) appears in the stimulus.

Experimental predictions

Inhibition-excitation patterns in mid-level audition

Our visualizations of second layer features in Fig 7B revealed that many represent the concurrent activation of many STKs. Examination of Fig 7B (columns 1 and 2) suggests that the first layer STKs that such units pool are typically highly similar either in their spectrotemporal or modulation properties. However, Fig 7B also shows examples of a distinct set of second-layer units in which increased activation of one group of STKs (again, typically similar to each other) is associated with a decrease in activity of another group of STKs. For example, when a harmonic feature becomes active when encoding a vowel, click-like features might become inactive. Such "opponency" was evident in a large subset of second layer units, and represents the main novel phenomenon evident in our model. To our knowledge no such opponent tuning has been identified in auditory neuroscience, but qualitatively similar opponent behavior is evident in visual neurons exhibiting end-stopping or cross-orientation inhibition. The results raise the possibility that coordinated excitation and inhibition could be a feature of central auditory processing, and we thus examined this model property in detail.

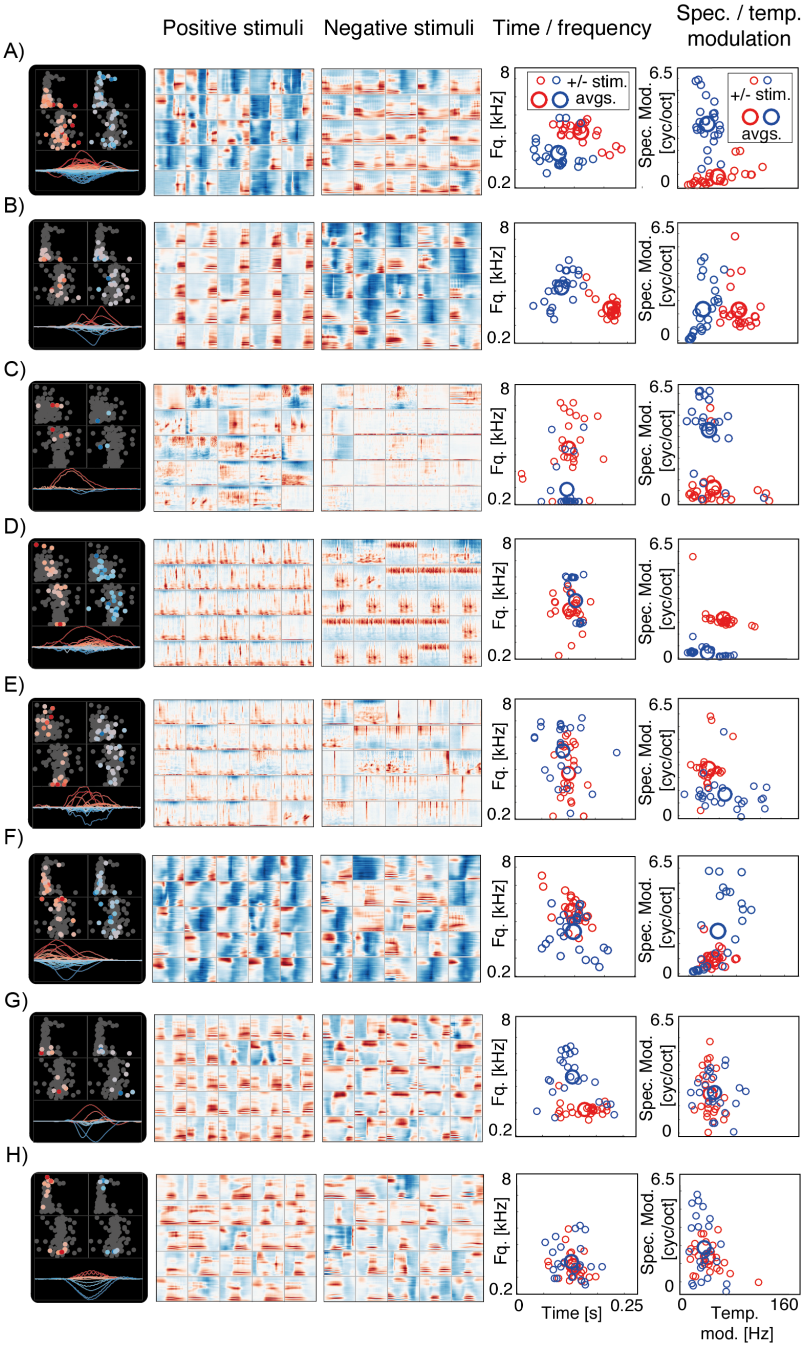

Figure 11: **Opponency in second layer units.** Each row corresponds to a particular second-layer unit. The leftmost column plots the center of mass of each first-layer STK in the modulation and time-frequency planes, along with the time courses of their weights in the second-layer unit (as in Fig 7B). The second and third columns from the left depict 25 stimuli eliciting strong positive and negative responses in the corresponding second-layer unit. In the fourth and fifth columns, positive and negative stimuli are visualized as red and blue circles, respectively, in time-frequency and modulation planes (the circle is located at the center of mass of the stimulus). Large circles correspond to centroids of positive and negative stimulus clusters.

Examples of opponent patterns encoded by second-layer features are visualized in Fig 11. One might imagine that one could simply examine the STKs pooled by each unit to determine the sort of stimuli eliciting strong positive or negative responses. But because the activation of each unit is the result of a non-linear inference process in which units compete to explain the stimulus pattern [29], it is often not obvious what a set of STKs will capture. Thus to understand which stimuli 'excite' or 'inhibit' second-level features, we inferred coefficients v by encoding the entire training dataset and selected two sets of 25 sound epochs that elicited the strongest positive and strongest negative responses, respectively, in each unit. These positive and negative stimuli are depicted in the second and third columns from the left, respectively. In the last two columns, the center of mass of each of these stimuli is plotted on the time-frequency plane (fourth column) and modulation plane (fifth column). Although the center of mass of each stimulus is admittedly a crude summary, the simplicity of the representation facilitates visualization and analysis of the clustering of positive and negative stimuli. We note also that the phenomenon of interest is that different sets of features have opposite effects on a second-layer unit - because second-layer coefficients and weights can be sign-reversed without changing the representation, the designation of stimuli (and STK weights) as 'exitatory' or 'inhibitory' is arbitrary.

In some cases, some natural function can be ascribed to the unit. For instance, positive activations of the unit shown in Fig 11B encode onsets of voiced speech, while its negative activations encode voicing offsets. By contrast, the unit shown in Fig 11H separates low from high pitch (which for speech distinguishes male from female speakers). For this unit, representations of positive and negative stimuli are mixed on the time-frequency and modulation planes, but positive stimuli contain lower harmonic stacks than negative stimuli, and listening to the stimuli revealed a clear difference in pitch/gender. The unit in in Fig 11G is positively activated by vowel excerpts separated by silence and suppressed by similar vowel excerpts separated instead by high-frequency noise.

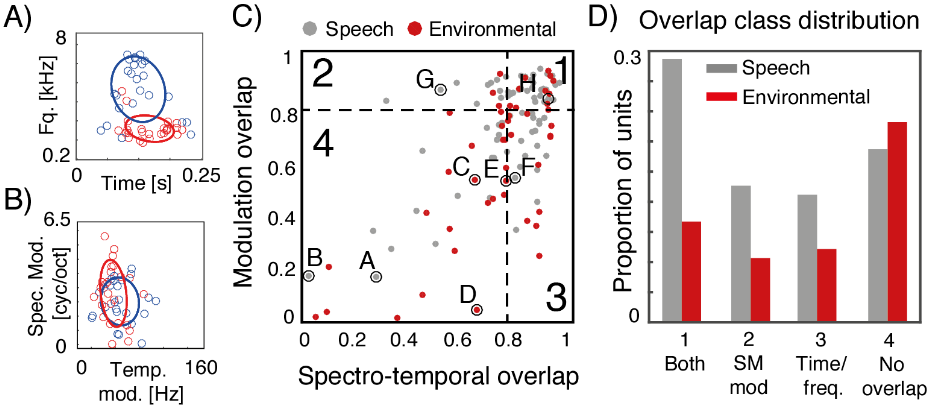

Figure 12: **Classification of opponent patterns in second-layer features.** A) Distribution of optimal positive and negative stimuli for an example second layer unit on the time-frequency plane. Ellipses mark covariance of Gaussian distributions fitted to positive or negative stimuli. B) Same as A but on spectro-temporal modulation plane. C) Separability of opponent stimulus classes as measured by their overlap in the time-frequency and modulation planes. Overlap was quantified by Bhattacharyya coefficients computed for the two Gaussians fitted to the positive and negative stimuli in each domain. Dashed lines correspond to the median of the Bhattacharyya coefficient across units. Numbers denote quadrants for reference in D. D) Distribution of units falling in each overlap quadrant for speech and environmental sound trained representations.

Perhaps the most salient property of the second-layer opponent stimulus classes is that they frequently segregate in at least one domain. To quantify this separation in the time-frequency and modulation domains, we fitted their distribution on each of the planes (with each STK represented by a point - its center of mass on the plane) with Gaussian distributions (Fig 12A and B). We then computed the Bhattacharyya coefficient (BC; a number summarizing the overlap between two distributions, closely related to the dot product of the densities) between the positive and negative Gaussians. Distributions with little overlap yield BC values close to 0, while BC values near 1 indicate high overlap. Each second-layer feature is thus characterized by two BC values measuring the overlap of positive and negative stimuli in the time-frequency and modulation planes. Fig 12C plots these values for each second-layer unit pooling STKs with opposite sign trained on speech (gray dots) or environmental sounds (red dots). To summarize these distributions, we plot the proportion of units for each training corpus falling into the four median-quadrants (Fig 12D). One property of the distributions is that more units with separable opponent stimuli emerged in a representation trained on environmental sounds compared to that trained on speech. This could reflect the diverse spectrotemporal characteristics of the environmental sounds corpus we used. Overall, however, opponent second layer units emerge irrespective of the training corpus, and they typically compute opponency between stimulus features that are distinct in at least one of the time-frequency or modulation planes. These opponent units encode STK activation patterns which are mutually exclusive and presumably do not co-occur in natural signals. The phenomenon is a natural one to investigate in the auditory cortex.

Discussion

Natural sounds are highly structured. The details of this structure and the mechanisms by which it is encoded by the nervous system remain poorly understood. Progress on both fronts is arguably limited by the shortage of signal models capable of explicitly representing natural acoustic structure. We have proposed a novel statistical model that captures an unexplored type of high-order dependency in natural sounds – correlations between the activations of basic spectrotemporal features. Our model consists of two layers. The first layer learns a set of elementary spectrotemporal kernels and uses them to encode sound cochleagrams. The second layer of the model forms a representation of co-occurence patterns of the first layer features.

We adopted a generative modelling approach, inspired by its previous successes in the domain of natural image statistics. Previous hierarchical, probabilistic models of natural images were able to learn high-order statistical regularities in natural images [24, 25, 41–43, 55–58]. In many cases the representations learned by these models exhibit similarity to empirically observed neural codes in the visual system.

The hierarchical representations learned by our model provide predictions about neural representations of mid-level sound structure. Moreover, the model reproduces certain aspects of the representational transformations found through the thalamus and cortex. The results suggest that principles of efficient coding could shape mid-level processing in the auditory cortex in addition to the auditory periphery [32].

Statistical dependencies in natural sounds

Our model exploits statistical dependencies between first-layer spectrotemporal kernels. These dependencies have two likely causes. First, dependencies almost surely remain from limitations of the convolutional sparse coding model, in that first-layer coefficients are never completely marginally independent even after learning (due to insufficient expressive power of the code). Second, even if the first-layer coefficients were fully marginally independent they would exhibit local dependencies when conditioned on particular sound excerpts. Observations of such conditional dependence have been made previously, mostly in the context of modelling natural image statistics [41–43,55,59]. We similarly observed strong non-zero cross-correlations between STKs for particular natural sound excerpts, and found that the second-layer units captured these sorts of dependencies (Fig. 3,8).

We also observed that different instances of the same acoustic event (such as the same word uttered twice by the same speaker) yield similiar STK coefficient trajectories (Fig. 4), which can be thought of as samples from a nonstationary distribution with a time-varying magnitude parameter. By modeling this magnitude the model learned representations with some degree of invariance (Fig. 9).

Model results - Convolutional spectrotemporal features

The model first learned a sparse convolutional decomposition of cochleagrams. Although prior convolutional sparse coding algorithms have been applied in engineering (e.g. [45,60]), to our knowledge our model is the first instantiation in a neuroscience context. Convolutional representations are less redundant than "patch-based" codes and can represent signals of arbitrary length.

The learned spectrotemporal features span clicks, harmonics, combinations of harmonics, bandpass noise, frequency sweeps, onsets, and offsets. Some of these structures are present in previous learned codes of spectrograms, but due to the convolutional nature of the learned code, the model learns only a single version of each feature rather than replicating it at different time offsets. Although difficult to quantify, our impression is that the code is more diverse than that obtained with previous patch-based approaches [33, 35]. As in previous neurophysiological measurements of cortical STRFs, the model STKs tile the spectrotemporal modulation plane, subject to the contraints of the tradeoff between time and frequency [52].

As with features learned from sound waveforms [32, 34], we found the first-layer spectrogram features to depend on the training corpus (speech or a set of environmental sounds). Although certain spectrotemporal patterns (e.g. single harmonics or clicks) appeared in both features sets, others were corpus-specific. Corpus-specific structure was also evident in the distributions of features in the modulation plane. This corpus-dependence contrasts with the relatively consistent occurrence of Gabor-like features in sparse codes of images. These results raise the possibility that auditory cortical neuronal tuning might exhibit considerable heterogeneity, given that the full range of natural audio must be encoded by cortical neurons.

Model results – Second-layer features

The second layer of the model encodes a probability distribution of spectrotemporal feature activations. By jointly modelling the magnitudes of the first layer responses, the second layer basis functions capture patterns of spectrotemporal kernel covariation. Magnitude modeling was essential to learning additional structure – we found in pilot experiments that simply applying a second layer of convolutional sparse coding did not produce comparable results. This observation is consistent with previous findings that nonlinear transformations of sparse codes often help to learn residual dependencies [42, 55, 61].

The second layer of our model learned a representation of spectrotemporal feature co-activations that frequently occur in natural sounds. Typically, the spectrotemporal features clustered by second-layer units shared some property - frequency content, temporal pattern or modulation characteristics (Fig. 7). Our model also identified 'opponent' patterns – sets of spectrotemporal features that are rarely active simultaneously in natural sounds (Fig. 11). The opponent stimulus classes for a second-layer unit often segregated in at least one of the spectrotemporal modulation or time-frequency planes. In other cases the opponency separated some other readily apparent aspect of sound, such as male from female voices, or onsets from offsets.

Relation to auditory neuroscience

The structure learned by the model from natural sounds replicates some known properties of the auditory system. The modulation frequencies preferred by units dropped from the first to the second layer (Fig. 10B), as has been observed between the thalamus and cortex. Unit "tuning" specificity (i.e. the similarity between stimuli eliciting a strong response) also decreases from the first to the second layer (Fig. 10C,D). A similar specificity decrease has been observed between granular and supra- and infra-granular layers of the cortex [54]. Although we do not suggest a detailed correspondence between layers of our model and particular anatomical structures, these similarities indicate that some of the principles underlying hierarchical organization of the auditory pathways may derive from the natural sound statistics and architectural choices that constrain our model.

The combinations of spectrotemporal features learned by our model provide hypotheses for neural tuning that might be present in the auditory system. Some units combined only STKs with similar acoustic properties (Fig. 7). Others encode activations of 'opponent' sets of first-layer kernels. Such opponent kernel sets are those that do not typically become active simultaneously in natural audio. The results are somewhat analogous to excitation-inhibition phenomena in visual neurophysiology (such as end-stopping, length and width suppression or cross-orientation inhibition) emerging in models of natural images [42, 56, 62, 63]. The second-layer units of our model provide candidates of potentially analogous phenomena in the auditory cortex.

Our model results are also relevant to recent evidence that central auditory neurons in some cases are driven by more than one stimulus feature [15,22,54]. The second-layer units of our model pool up to dozens of single layer features with strong weights. Our model reproduces that trend and predicts that distinct dimensions may be combined in opponent fashion.

In addition to these potential similarities, there are several respects in which the model likely diverges from biological auditory systems. Most prominently, the model relies on acausal inference: the encoding is derived for the entire stimulus at once. This is clearly not something that a biological auditory system could instantiate because of the need to operate in real-time. Differences are also likely in the input data. The model is trained on a cochleagram which captures some but not all aspects of auditory nerve coding. Other models have successfully captured aspects of auditory perception with such input representations [48], but it remains possible that additional structures would become apparent with a richer input model.

Relation to other modeling approaches

The model described here represents one of several approaches to construction of hierarchical signal representations. The representations in our model are learned from natural sounds, and thus contrast with hand-engineered models that seek to replicate known or hypothesized features of sensory coding [48, 64]. One advantage of learning models from natural signals is that any similarities with known neural phenomena provide candidate normative explanations for these phenomena, e.g. that they arise from the demands of efficient coding or some other optimality constraint imposed by the model [29, 32]. Another potential advantage is that one might hope to learn structures that have not yet been observed neurally, but that could provide hypotheses for future experiments. The latter was the main motivation for our modeling approach.

The model is also unsupervised, and generative, specifying a joint probability distribution over the data and coefficients in both latent layers. An alternative approach to learning hierarchical representations of data is to use discriminative models, which can typically be thought of as optimizing a representation for performance of a single task. Prominent recent examples of such discriminative learning come from the field of deep neural networks [65]. Models of this class are typically straightforward to optimize and can yield high performance on classification tasks, but typically require large numbers of labeled training examples. The requirement of labeled training data is sometimes a practical limitation, and raises questions about the extent to which the learning procedure could relate to the brain. Generative models currently have the disadvantage of requiring custom optimization procedures. However, their generative nature facilitates certain applications (e.g. denoising) and allows samples to be straightforwardly generated, potentially for use in experiments. Moreover, they do not require labeled data, and learn representations that are independent of any particular task, much like the unsupervised learning believed to occur in sensory systems.

Conclusion

We have presented a hierarchical model of natural sounds. When trained on natural sound corpora, the model learned a representation of spectrotemporal feature combinations. The properties of the model layers resemble aspects of hierarchical transformations previously observed in the brain, suggesting that efficient coding could shape such transformations throughout the auditory system. The learned mid-level features provide hypotheses for auditory cortical tuning as well as a means to parameterize stimuli with which to probe mid-level audition.

Supporting Information

Gradients for learning and inference in the first layer Learning and inference in the first layer was achieved via gradient descent on negative log-posterior (Eq 5). Below we provide expressions of E_1 gradient with respect to $s_{i,t}$ and $\phi_{i,f,\tau}$ respectively

The non-negativity constraint on sparse coefficients s complicates the optimization process in learning and inference. To alleviate this complication, we introduced auxiliary coefficients z_i and assumed that non-negative sparse coefficients s_i are equal to squares of z_i:

$$s_i = z_i^2, z_i \in \mathbb{R} \tag{13}$$

We replace s_i with z_i^2 in equations below.

Gradients of first layer energy function (Eq 5) with respect to $z_{i,t}$ and $\phi_{i,f,\tau}$ was respectively:

$$\frac{\partial E_1}{\partial z_{i,t}} \propto -\frac{4z_{i,t}}{\sigma^2}\sum_{f=1}^{F}(\phi_{i,f} \odot e_f)_t + 2\lambda_i z_{i,t} \tag{14}$$

$$\frac{\partial E_1}{\partial \phi_{i,f,\tau}} \propto -2\Big(s_i \odot e_f\Big)_\tau \tag{15}$$

where $\odot$ denotes cross-correlation and $e_{f,t} = x_{f,t} - \hat{x}_{f,t}$ is the reconstruction error.

Gradients for learning and inference in the second layer Similarly, learning and inference in the second layer was achieved via gradient descent on the corresponding energy function (equation 10). Gradient expressions took the following form:

$$\frac{\partial E_2}{\partial v_{i,t}} \propto \sum_{n=1}^{N} \psi_n \odot B_{i,n} + \alpha\,\mathrm{sgn}(v_{i,t}) \tag{16}$$

$$\frac{\partial E_2}{\partial B_{j,i,t_b}} \propto (1 - \psi_i \odot v_j)_{t_b} + \beta\,\mathrm{sgn}(B_{j,i,t_b}) \tag{17}$$

where $\hat{\lambda}_{n,t} = \lambda_n \exp\Big[\sum_j^M v_j * B_j\Big]_t$ is the reconstruction of instantaneous scale parameter and $\psi_{n,t} = \frac{s_{n,t}}{\hat{\lambda}_{n,t}}$.

Acknowledgments

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. We acknowledge a McDonnel Foundation grant to JHM.

References

[1] Chechik G, Nelken I. Auditory abstraction from spectro-temporal features to coding auditory entities. Proceedings of the National Academy of Sciences. 2012;109(46):18968–18973.

[2] Atencio CA, Sharpee TO, Schreiner CE. Receptive field dimensionality increases from the auditory midbrain to cortex. Journal of neurophysiology. 2012;107(10):2594–2603.

[3] Carruthers IM, Laplagne DA, Jaegle A, Briguglio JJ, Mwilambwe-Tshilobo L, Natan RG, et al. Emergence of invariant representation of vocalizations in the auditory cortex. Journal of neurophysiology. 2015;114(5):2726–2740.

[4] Bizley JK, Nodal FR, Nelken I, King AJ. Functional organization of ferret auditory cortex. Cerebral Cortex. 2005;15(10):1637–1653.

[5] Elie JE, Theunissen FE. Meaning in the avian auditory cortex: neural representation of communication calls. European Journal of Neuroscience. 2015;41(5):546–567.

[6] Russ BE, Ackelson AL, Baker AE, Cohen YE. Coding of auditory-stimulus identity in the auditory non-spatial processing stream. Journal of neurophysiology. 2008;99(1):87–95.

[7] Mesgarani N, David SV, Fritz JB, Shamma SA. Phoneme representation and classification in primary auditory cortex. The Journal of the Acoustical Society of America. 2008;123(2):899–909.

[8] Obleser J, Leaver A, VanMeter J, Rauschecker JP. Segregation of vowels and consonants in human auditory cortex: evidence for distributed hierarchical organization. Frontiers in psychology. 2010;1:232.

[9] Bendor D, Wang X. The neuronal representation of pitch in primate auditory cortex. Nature. 2005;436(7054):1161–1165.

[10] Overath* T, McDermott* JH, Zarate JM, Poeppel D. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. Nature Neuroscience. 2015;18:903–911. doi:10.1038/nn.4021.

[11] Norman-Haignere S, Kanwisher NG, McDermott JH. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron. 2015;88(6):1281–1296.

[12] Aertsen A, Johannesma P. The spectro-temporal receptive field. Biological cybernetics. 1981;42(2):133–143.

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

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

[15] Sharpee TO, Atencio CA, Schreiner CE. Hierarchical representations in the auditory cortex. Current opinion in neurobiology. 2011;21(5):761–767.

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

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

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

[19] Sahani M, Linden J; MIT Press. How linear are auditory cortical responses? 2003;15:125.

[20] Williamson RS, Ahrens MB, Linden JF, Sahani M. Input-specific gain modulation by local sensory context shapes cortical and thalamic responses to complex sounds. Neuron. 2016;91(2):467–481.

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

[22] Kozlov AS, Gentner TQ. Central auditory neurons have composite receptive fields. Proceedings of the National Academy of Sciences. 2016; p. 201506903.

[23] Harper NS, Schoppe O, Willmore BD, Cui Z, Schnupp JW, King AJ. Network Receptive Field Modeling Reveals Extensive Integration and Multi-feature Selectivity in Auditory Cortical Neurons. PLOS Comput Biol. 2016;12(11):e1005113.

[24] Lee H, Ekanadham C, Ng AY. Sparse deep belief net model for visual area V2. In: Advances in neural information processing systems; 2008. p. 873–880.

[25] Hoyer PO, Hyvärinen A. A multi-layer sparse coding network learns contour coding from natural images. Vision research. 2002;42(12):1593–1605.

[26] Barlow HB. Possible principles underlying the transformations of sensory messages. 1961;.

[27] Attneave F. Some informational aspects of visual perception. Psychological review. 1954;61(3):183.

[28] Srinivasan MV, Laughlin SB, Dubs A. Predictive coding: a fresh view of inhibition in the retina. Proceedings of the Royal Society of London B: Biological Sciences. 1982;216(1205):427–459.

[29] Olshausen BA, Field DJ. Sparse coding with an overcomplete basis set: A strategy employed by V1? Vision research. 1997;37(23):3311–3325.

[30] Bell AJ, Sejnowski TJ. The "independent components" of natural scenes are edge filters. Vision research. 1997;37(23):3327–3338.

[31] van Hateren JH, van der Schaaf A. Independent component filters of natural images compared with simple cells in primary visual cortex. Proceedings of the Royal Society of London B: Biological Sciences. 1998;265(1394):359–366.

[32] Lewicki MS. Efficient coding of natural sounds. Nature neuroscience. 2002;5(4):356–363.

[33] Klein DJ, König P, Körding KP. Sparse spectrotemporal coding of sounds. EURASIP Journal on Advances in Signal Processing. 2003;2003(7):1–9.

[34] Smith EC, Lewicki MS. Efficient auditory coding. Nature. 2006;439(7079):978–982.

[35] Carlson NL, Ming VL, DeWeese MR. Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. PLoS Comput Biol. 2012;8(7):e1002594.

[36] Terashima H, Okada M. The topographic unsupervised learning of natural sounds in the auditory cortex. In: Advances in Neural Information Processing Systems; 2012. p. 2312–2320.

[37] Młynarski W. The opponent channel population code of sound location is an efficient representation of natural binaural sounds. PLoS Comput Biol. 2015;11(5):e1004294.

[38] Mlynarski W. Efficient coding of spectrotemporal binaural sounds leads to emergence of the auditory space representation. Name: Frontiers in Computational Neuroscience. 2014;.

[39] Carlin MA, Elhilali M. Sustained firing of model central auditory neurons yields a discriminative spectro-temporal representation for natural sounds. PLOS Comput Biol. 2013;9(3):e1002982.

[40] Hyvärinen A, Hoyer P. Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. Neural computation. 2000;12(7):1705–1720.

[41] Karklin Y, Lewicki MS. A hierarchical Bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. Neural computation. 2005;17(2):397–423.

[42] Karklin Y, Lewicki MS. Emergence of complex cell properties by learning to generalize in natural scenes. Nature. 2009;457(7225):83–86.

[43] Cadieu C, Olshausen BA. Learning transformational invariants from natural movies. In: Advances in neural information processing systems; 2008. p. 209–216.

[44] Bumbacher E, Ming V. Pitch-sensitive Components Emerge from Hierarchical Sparse Coding of Natural Sounds. 2012; p. 219–229.

[45] Blumensath T, Davies M. Sparse and shift-invariant representations of music. Audio, Speech, and Language Processing, IEEE Transactions on. 2006;14(1):50–57.

[46] Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS, Dahlgren NL, et al. TIMIT acoustic-phonetic continuous speech corpus. Linguistic data consortium, Philadelphia. 1993;33.

[47] Glasberg BR, Moore BC. Derivation of auditory filter shapes from notched-noise data. Hearing research. 1990;47(1):103–138.

[48] McDermott JH, Simoncelli EP. Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. Neuron. 2011;71(5):926–940.

[49] Robles L, Ruggero MA. Mechanics of the mammalian cochlea. Physiological reviews. 2001;81(3):1305–1352.

[50] Miller LM, Escabı MA, Schreiner CE. Feature selectivity and interneuronal cooperation in the thalamocortical system. The Journal of Neuroscience. 2001;21(20):8136–8144.

[51] Bhattachayya A. On a measure of divergence between two statistical population defined by their population distributions. Bulletin Calcutta Mathematical Society. 1943;35:99–109.

[52] Singh NC, Theunissen FE. Modulation spectra of natural sounds and ethological theories of auditory processing. The Journal of the Acoustical Society of America. 2003;114(6):3394–3411.

[53] Miller LM, Escabí MA, Read HL, Schreiner CE. Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. Journal of neurophysiology. 2002;87(1):516–527.

[54] Atencio CA, Sharpee TO, Schreiner CE. Hierarchical computation in the canonical auditory cortical circuit. Proceedings of the National Academy of Sciences. 2009;106(51):21894–21899.

[55] Hyvärinen A, Hurri J, Hoyer PO. Natural Image Statistics: A Probabilistic Approach to Early Computational Vision.. vol. 39. Springer Science & Business Media; 2009.

[56] Hosoya H, Hyvärinen A. A hierarchical statistical model of natural images explains tuning properties in V2. The Journal of Neuroscience. 2015;35(29):10412–10428.

[57] Garrigues P, Olshausen BA. Learning horizontal connections in a sparse coding model of natural images. In: Advances in Neural Information Processing Systems; 2008. p. 505–512.

[58] Berkes P, Turner RE, Sahani M. A structured model of video reproduces primary visual cortical organisation. PLoS Comput Biol. 2009;5(9):e1000495.

[59] Karklin Y, Ekanadham C, Simoncelli EP. Hierarchical spike coding of sound. In: Advances in neural information processing systems; 2012. p. 3032–3040.

[60] Grosse R, Raina R, Kwong H, Ng AY. Shift-invariance sparse coding for audio classification. arXiv preprint arXiv:12065241. 2012;.

[61] Shan H, Zhang L, Cottrell GW. Recursive ica. Advances in neural information processing systems. 2007;19:1273.

[62] Coen-Cagli R, Dayan P, Schwartz O. Cortical surround interactions and perceptual salience via natural scene statistics. PLoS Comput Biol. 2012;8(3):e1002405.

[63] Zhu M, Rozell CJ. Visual nonclassical receptive field effects emerge from sparse coding in a dynamical system. PLoS Comput Biol. 2013;9(8):e1003191.

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

[65] Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences. 2014;111(23):8619–8624.