Spectro-Temporal Models of Inferior Colliculus Neuron Receptive Fields

S. Zayd Enam[1,*], Michael R. DeWeese[1,2],
1 Redwood Center for Theoretical Neuroscience, University of California, Berkeley, CA, United States of America
2 Helen Wills Neuroscience Institute, University of California, Berkeley, CA, United States of America
∗ **E-mail: zayd@berkeley.edu**

Abstract

Sparse codes for speech spectrograms qualitatively match properties of receptive fields of Inferior Colliculus (ICC) neurons. We find sparse codes of speech-spectrograms are well described by one of four models and we find that these models also fit ICC spectro-temporal receptive fields (STRF) well. Further, our models are able to express time-frequency inseparable receptive fields (e.g. frequency sweeps) that previous models were unable to satisfactorily describe. Our models allow the accurate characterization of high-dimensional STRFs with more natural parameterizations of the neuron's behavior.

Author Summary

In this paper we propose parametric model classes for the STRF of auditory neurons. A STRF represents the intensity of a neuron's response to different values of frequency and time. To determine STRF model classes we first fit model classes to the sparse codes of speech-spectrograms using non-linear least squares methods. These sparse codes were also found to qualitatively match the STRFs of neurons recorded from the Inferior Colliculus (ICC) of anesthetized animals [1]. In order to better quantitatively analyze these comparisons we compared the best-fit parameters of our models for the sparse codes and the best-fit parameters of our models when fit to STRFs generated from recordings of ICC neurons of anesthetized cats [2]. We then analyzed both data sets in this parameter space. Our classes (Checkerboards, Harmonic Stacks, Inhibited Harmonic Stacks, Frequency Sweeps and Phase Variant Checkerboards) accurately model the STRFs of the sparse coded neurons and the experimentally recorded ICC neurons. Our models are also able to capture time-frequency inseparable components of receptive fields that previous models could not. In order to determine the quality of our models we examined various goodness-of-fit measures and compared how well our models fit sparse codes of speech-spectrograms and ICC data.

Introduction

Previous work in visual neuroscience has shown that sparse codes for images of natural scenes can predict the physiological properties of spatio-temporal receptive fields of V1 neurons [3] [4]. V1 simple cell receptive fields have been shown to be modeled well by Gabor functions [5].The comparisons to physiological data are determined by fitting Gabor functions to the learned sparse codes and to the recorded neuron data. The two data sets can then be compared in this parameter space. These comparisons allow one to characterize more interesting aspects of a neuron's receptive field, including orientation and frequency statistics across populations. In recent work on a training higher level theoretical models, Karklin and Lewicki found a suitable hierarchical parameterization for natural images where they were able to classify and cluster the image patches based on which parts of a natural scene they were extracted from. [6]

Similarly, in the auditory domain STRFs are a popular method of characterizing the complex time frequency responses of auditory neurons to natural sounds [7]. Recent work has shown that overcomplete, 'hard' sparse codes for speech-spectrograms predict receptive fields of neurons in the ICC. [1]. However,

fitting models to these receptive fields to compare them to physiological data has proven to be more difficult than in the visual neuron case because 1) auditory receptive fields exhibit a wider diversity of receptive field shapes and 2) in order to capture the salient features characterizing the neuron, auditory receptive fields must be of larger dimensions (and often log-sampled in the frequency dimension). Indeed, past attempts at fitting shapes to auditory receptive fields involved cutting the receptive field into one dimensional slices along time and frequency and fitting 1-D Gabors to each slice (the resultant STRF is computed as the outer product of the the two 1-D Gabors) [2] [8]. This is because the size of the receptive fields makes fitting auditory receptive fields a difficult minimization problem over a large number of data points (e.g. $\approx 400 \times 200$ data points as opposed to 16×16 patches of V1 receptive fields) and a large number of simultaneously varying parameters of a non-convex model.

These 1-D outer product models are unsatisfactory because they are unable to meaningfully describe structure in receptive fields that is any more complicated than checkerboards. Previous work using the outer product model has attempted to decompose time-frequency inseparable receptive fields into time-frequency separable components through the use of singular value decomposition (SVD). The resulting decomposed dyads are fit using the above outer-product method. This method is suboptimal because the SVD step creates artificial time-frequency separable receptive fields which are not good parameterizations for any time-frequency modulation (even simple ubiquitous frequency sweeps). In this paper we show how auditory STRFs can be fit by time-frequency inseparable 2-D models in a tractable manner.

Results

Our models can be split into four major classes of receptive fields each which of can be represented as the product of analytical functions. Here we will give an overview of the model classes and then we will discuss the possible ways that

Model Types

1. Checkerboard

Our first and simplest class is a 2-D checkerboard with a Gaussian window. While this model's most canonical shape is the actual checkerboard, it is also able to characterize many of the other shapes identified by Carlson et al. These include onset and terminating shapes amongst some other unnamed shapes identified [1]. The model can be characterized with the with following function over time t and stimulus frequency f:

$$R(t,f) = a \cdot \cos(2\pi\psi_t\, t + \phi_t) \cdot \cos(2\pi\psi_f\, f + \phi_f) \cdot \exp\left(-\frac{(t-t_c)^2}{2\sigma_t^2} - \frac{(f-f_c)^2}{2\sigma_f^2}\right) + c \tag{1}$$

In the above model the parameters ψ_t, ϕ_t and ψ_f, ϕ_f can be interpreted as the frequency and phase of the sinusoidal responses in time and stimulus frequency. The parameters t_c and f_c correspond to the center of the Gaussian window and σ_t and σ_f correspond to axis lengths of the window. The whole receptive field is then scaled by parameter a and has a bias parameter c. The canonical Gabor shape fitted to visual neuron receptive fields is simply a specific version of this model where f_x and ϕ_x are fixed to 0 [5]. An extension of this model is the replacement of the time and frequency variables with $(t-t_c)\cdot\cos(\theta) + (f-f_c)\cdot\sin(\theta)$ and $-(t-t_c)\cdot\sin(\theta) + (f-f_c)\cdot\cos(\theta)$ respectively. This allows the checkerboard to be oriented in directions non parallel to the time and frequency axis by adjusting the θ parameter. While the original model can be represented as the outer-product of two Gabors, extending the checkerboard with an orientation angle means this is no longer possible. We use this model and will discuss it further in our third class of neurons, frequency sweeps.

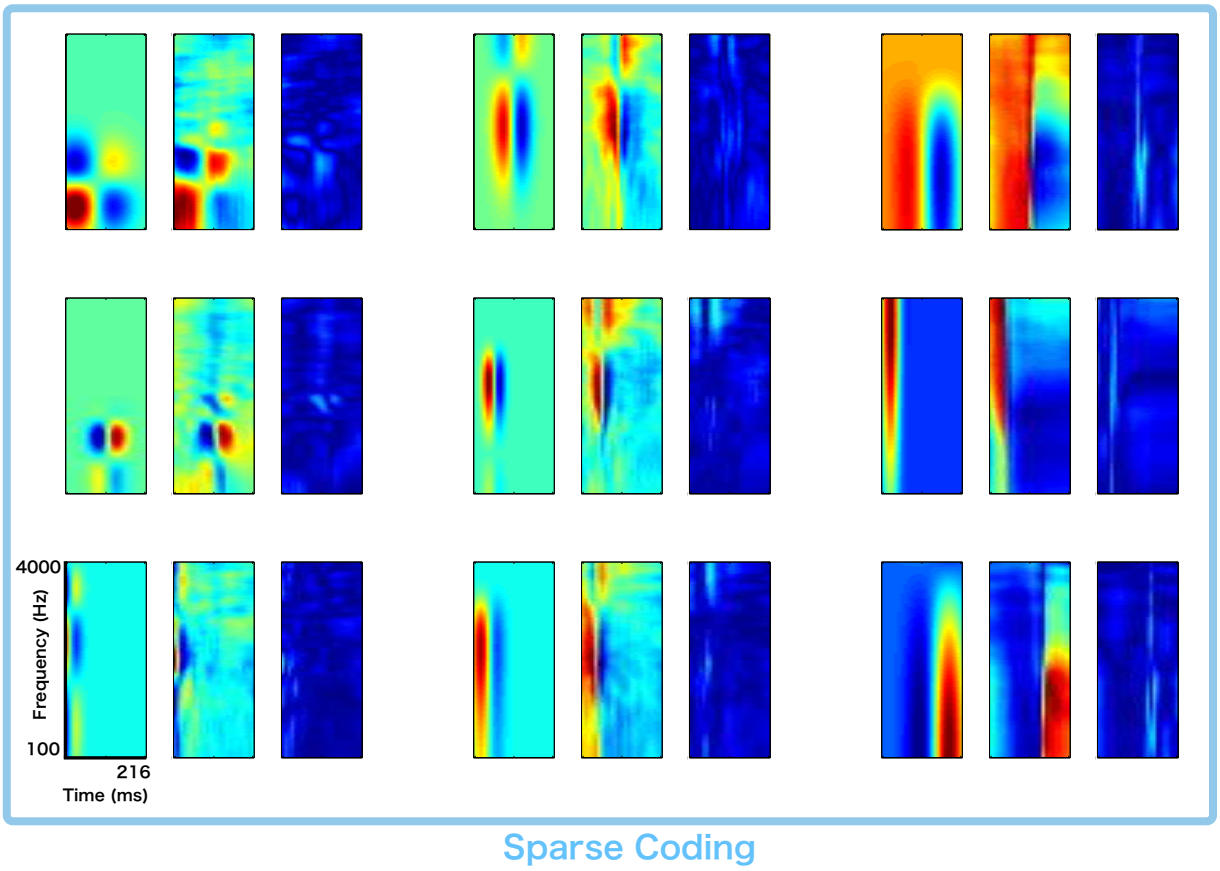

Figure 1. Checkerboard pattern receptive fields fit on checkerboard, onset and terminating shape sparse codes of speech spectrograms: In the left box of each panel we display the fitted model (with time from 0 to 216 msec plotted along the x axis and log-frequency from 100 to 4000 Hz plotted along the y axis), the middle box displays the sparse coded speech-spectrogram basis function that was fit and the right box displays the absolute value of the resultant residual error when the two are subtracted.

Often times it is common to represent a STRF with a log-sampled frequency axis in order to capture behavior across a wide range of frequencies. In this case we simply take the f variable to a power and the model we fit becomes:

$$R(t,f) = a \cdot cos(2\pi\psi_t\, t + \phi_t) \cdot cos(2\pi\psi_f\, 10^f + \phi_f) \cdot \exp\left(-\frac{(t-t_c)^2}{2\sigma_t^2} - \frac{(10^f - f_c)^2}{2\sigma_f^2}\right) + c \qquad (2)$$

2. Harmonic Stacks and Inhibited Harmonic Stacks

The next model class we describe models simple and inhibited harmonic stacks. Harmonic stacks represent neural activation across multiple frequency bands for a single window of time. These frequency bands are separated by gaps with no activation. A special type of harmonic stacks, inhibited harmonic stacks flank a simple harmonic stack with one that is of negative amplitude. This inhibited stack can be present

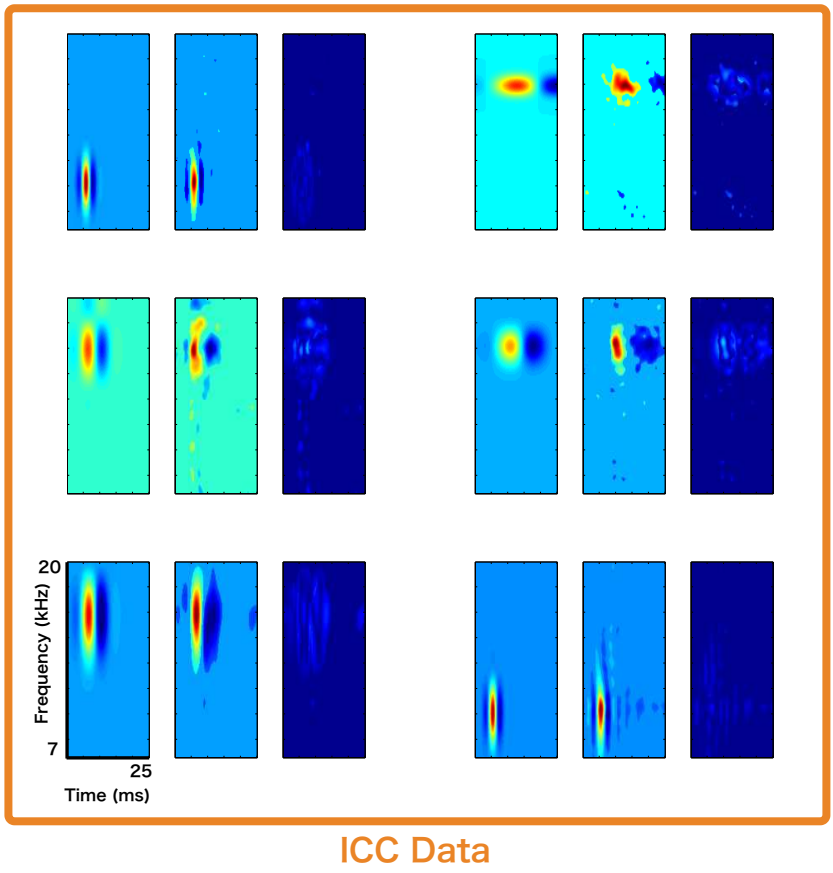

Figure 2. Checkerboard pattern receptive fields fit on receptive fields of Inferior Colliculus Neurons: In the left box of each panel we display the fitted model (with time from 0 to 216 msec plotted along the x axis and log-frequency from 100 to 4000 Hz plotted along the y axis), the middle box displays the sparse coded speech-spectrogram basis function that was fit and the right box displays the absolute value of the resultant residual error when the two are subtracted.

pre or post the simple harmonic stack.

$$R(t,f) = a \cdot cos(2\pi\psi_t\, t + \phi_t) \cdot |cos(2\pi\psi_f\, f + \phi_f)| \cdot \exp\left(-\frac{(t-t_c)^2}{2\sigma_t^2} - \frac{(f-f_c)^2}{2\sigma_f^2}\right) + c \qquad (3)$$

In the simple harmonic stack case f_x and a are fixed to 0. Here our model is similar to the above checkerboard model except that we take the absolute value of the resulting sinusoid in the frequency dimension. This modification is necessary because the gaps between each positive lobe remain relatively constant and are smaller than the period of the positive lobes (this characteristic of the harmonic stack means it is not possible to model the structure by adding a simple out of phase sinusoid in the frequency dimension). Because of this it is also possible to alter the period the positive lobes and the width of the gaps as a function of frequency by adjusting the frequency parameter.

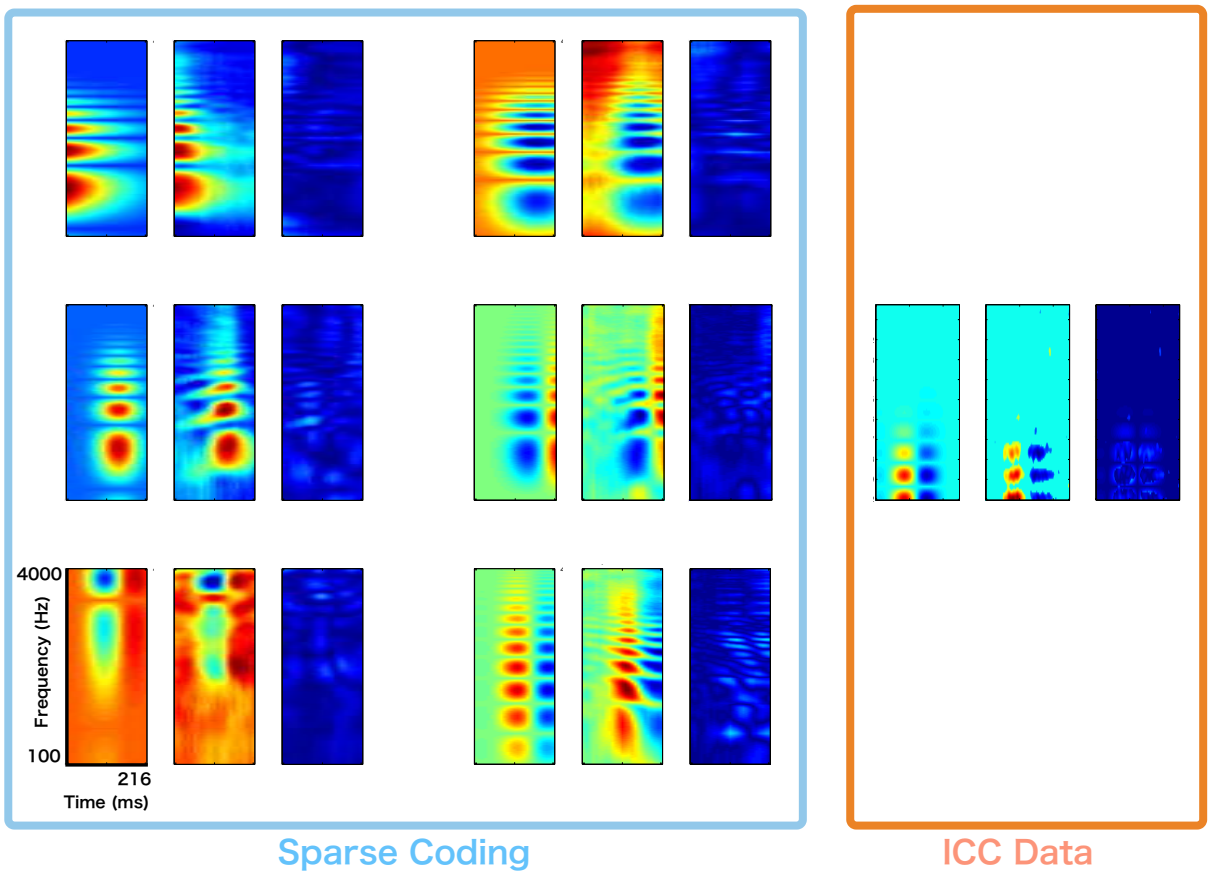

Figure 3. Harmonic Stack Receptive Field fits: Learned Neuron Example fits to Inhibited Harmonic Stacks. The left panels show example fits from the sparse coded speech-spectrograms and the right example is an example inhibited harmonic stack from the ICC dataset.

3. Frequency Sweeps

Frequency sweeps are a canonical example of time-frequency inseparable receptive fields. Existing SVD based methods to fit receptive fields break up the complex inseparable structure of the frequency sweeps into components that don't explain or effectively parameterize the 'sweep' characteristic of these receptive fields. Taking inspiration from the discussion of separable and inseparable components of STRFs in [9] our approach attempts to model the time frequency inseparable receptive fields by breaking the receptive field into two parts: 1) baseline activity that can be described by a time frequency separable model and 2) complex time-frequency inseparable activity. This effectively breaks the receptive field down to the 'sweep' component and other baseline activity. The 'sweep' component is often fit by an oriented Gabor function and the other components of the receptive field are modeled by one of earlier models.

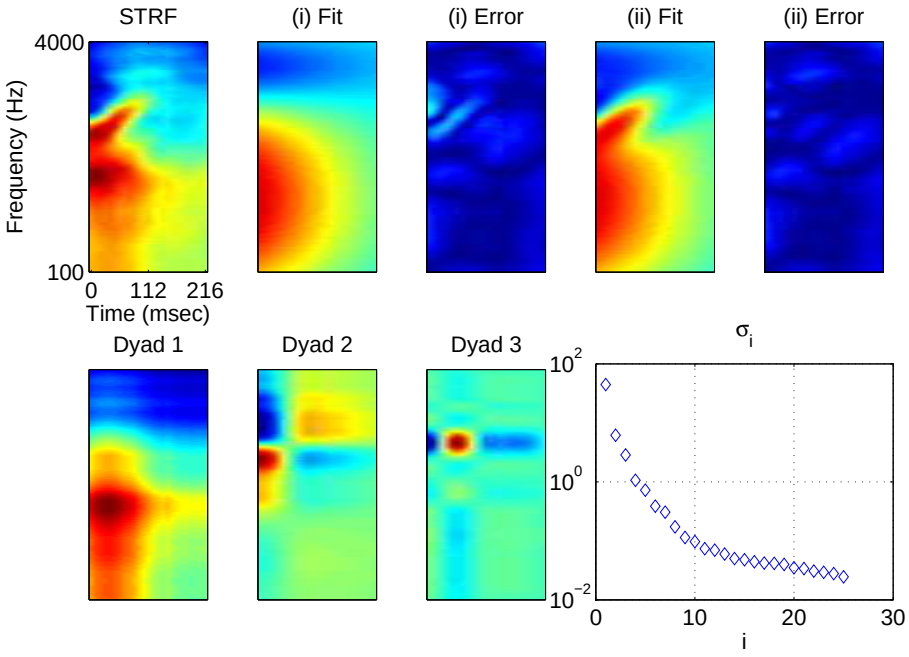

Figure 4. Time-frequency separable and inseparable components of a frequency sweep: Frequency sweeps are a common canonical shape found in auditory neuron STRFs [9] [10]. Existing models to fit this would first attempt to decompose the STRF to its statistically significant time-frequency separable dyads. The statistical significance can be estimated by a log plot of the magnitude of singular values. Each dyad would then be individually fit by a outer product of 1-D Gabors [2]. As can be seen by these dyads, this model does not capture the frequency sweep and it also does not parameterize the frequency sweep in any natural way. Instead, the SVD step creates artificial time-frequency separable components. In **(i)** we show the fit and absolute residual error of a 1-D Gabor method on the most significant dyad. In **(ii)** we show the result of our method of fitting time-frequency inseparable STRFs

.

4. Phase Variant Checkerboard

This last class is an extension of the standard Checkerboard model where the phase of the sinusoids is a function of the current time or frequency.

$$R(t,f) = a \cdot cos(2\pi\psi_t\, t + \phi_t(f)) \cdot cos(2\pi\psi_f\, f + \phi_f(t)) \cdot \exp\left(-\frac{(t-t_c)^2}{2\sigma_t^2} - \frac{(f-f_c)^2}{2\sigma_f^2}\right) + c \qquad (4)$$

These receptive fields are not that commonly seen in the sparse coded data set or in actual recorded STRFs but they are present. Again, for these models, the complex phase change is decomposed by an SVD step where it does not naturally represent the data. Here ϕ_x and ϕ_y represent functions dependent on stimulus frequency f and time t. For ϕ_x and ϕ_y we used a summation of shifted Heaviside step functions $\mathcal{H}$ with each time shift dependent upon the period of the corresponding sinusoid:

$$\phi_t(f) = \sum_i a \cdot \mathcal{H}(f - \frac{b}{\psi_t})$$

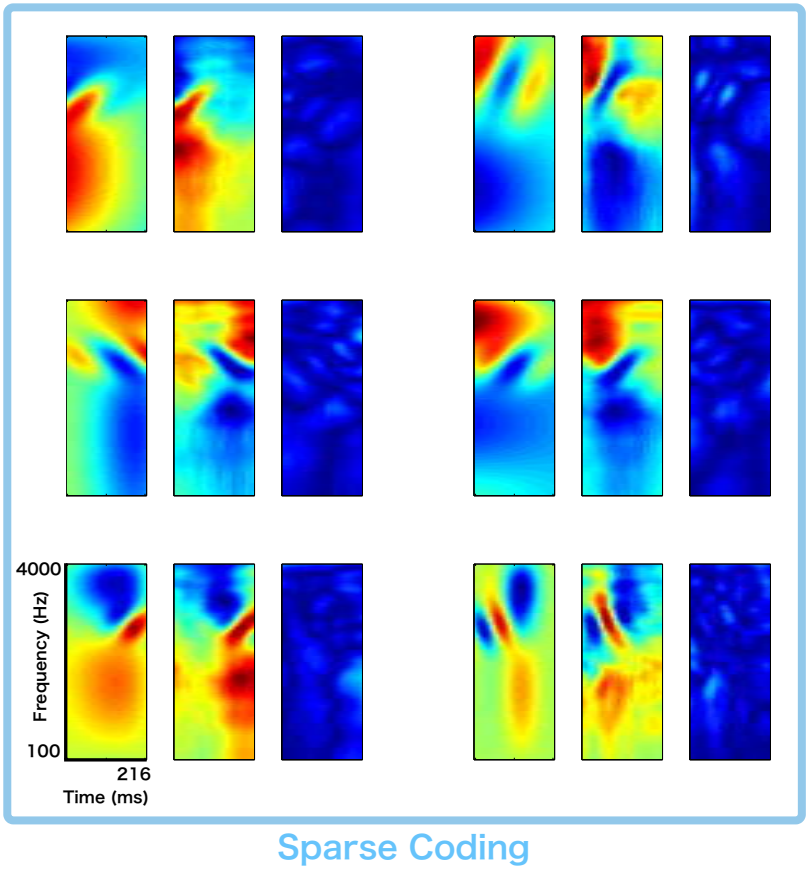

Figure 5. Frequency Sweep receptive fields fits of sparse codes of speech spectrograms: In the left box of each panel we display the fitted model (with time from 0 to 216 msec plotted along the x axis and log-frequency from 100 to 4000 Hz plotted along the y axis), the middle box displays the sparse coded speech-spectrogram basis function that was fit and the right box displays the absolute value of the resultant residual error when the two are subtracted.

Here a represents a scaling variable and b represents a constant multiplier of the shift frequency.

Quality of Fits

In order to determine how well this model fits the data we examine root-mean-square-error (RMSE) as a quality metric across each data set. We show that the root-mean-square-error of our model is low across all fits for both data sets and that our models are quantitatively better compared to previous models at fitting time-frequency inseparable STRFs.

Discussion

In this paper we show that neurons trained with a sparse coding network on speech-spectrogram data can be modeled well by three novel model classes (one of which is a generalization of the canonical Gabor function). Each model class is characterized with approximately 10 parameters. We also show

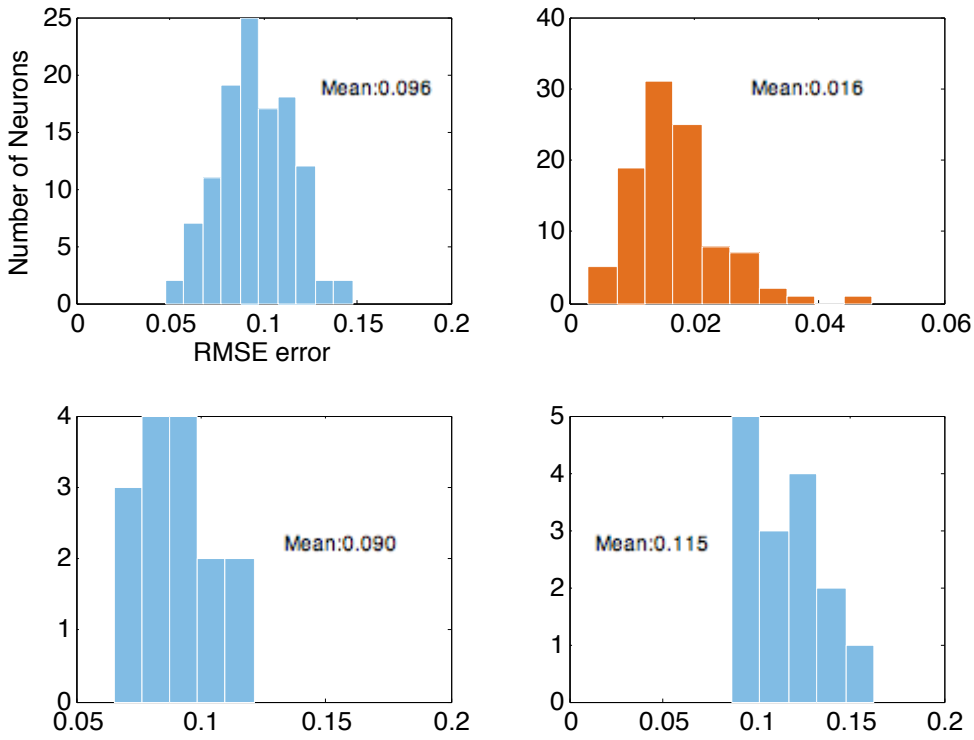

Figure 6. Histogram plots of RMSE of fits: In the top row we show a histogram of RMSE errors across the individual datasets where top-left plot is for the sparse-codes of speech-spectrogram and the top-right plot is for the is for the ICC data. In the bottom row we plot the advantage of time-frequency inseparable model. In the bottom-left we plot the RMSE error of fitting time-frequency inseparable models to frequency sweeps and in the bottom-right we show the the best that time-frequency separable models can do on the same receptive fields.

that the same models apply to biological neurons and model the spectro-temporal receptive fields of experimentally recorded Inferior Colliculus neurons well. Furthermore we show that within these model classes our parametrized best fits have similar distributions across both data sets. This provides evidence for the validity of the learned speech spectrogram neurons.

Our optimization tools solve a fitting problem for a very large and non-convex parameter space. The previous approach to fit visual or auditory receptive fields have either tackled much smaller receptive field sizes [5] [11] or have approached larger problems by slicing the STRF into individual time and frequency components and computing the resultant 2-D STRFs as the outer product of two one dimensional Gabors [2]. The second approach is sub optimal because the resulting model is only able to capture time-frequency separable components of the receptive field. In order to capture even simple time-frequency modulations such as frequency modulation the receptive field must be artificially converted into time-frequency separable subcomponents through the use of singular value decomposition. We have constructed 2-D receptive field models and have used effective optimization tools (parameter estimation, random restarts, simulated annealing) to find accurate best-fit parameters for much larger receptive fields.

Existing work on fitting receptive fields of auditory neurons uses models that can capture only time-frequency separable components of the receptive fields [2] [8]. Within our datasets we have many receptive

fields that have non time-frequency separable components. Previous methods to tackle this problem have attempted to model these inseparable components by decomposing the receptive field into time frequency separable components using Singular Value Decomposition. This model does not accurately characterize receptive fields as they do not capture the underlying dependent structure between time and frequency in these receptive fields. Additionally these models increase the number of parameters required to fit each receptive field as they are required to be broken up into multiple components that must each be fitted separately.

A direction for future research that we are currently exploring is the space of possible sounds that can be generated from these parameterized models.

Materials and Methods

Data

Recorded Inferior Colliculus Data

Our dataset consisted of 99 single unit recordings in the ICC of anesthetized cats. Sounds were presented binaurally with an independent sound to each ear. From these separate single unit recording, contralateral and ipsilateral STRFs were computed using spike-triggered averaging [12]. Dynamic moving ripple stimulus was presented which consists of amplitude and frequency modulated sinusoids that model features similar to those present in vocalizations. The STRFs consisted of either 400 samples in frequency and 276 samples in time or 230 samples in frequency and 400 samples in time. Sampled frequencies spanned from 0 Hz to 20 kHz with a time interval of 100 milliseconds.

Sparse Encoded Neurons

Our data consisted of receptive fields trained using an 4x over-complete representation of speech-spectrograms constructed from the TIMIT dataset [1]. We randomly selected 99 receptive fields to use for statistical comparisons. For each of the neurons we ran fitting procedures for multiple classes and selected the class with the best fit (determined by a sum of square errors metric). The samples in frequency for the spectrograms were logarithmically spaced ($\log_{10}$) and as a result we modified our model to include this. The data had 256 samples in frequency and 25 samples in time with the corresponding frequencies ranging from 100 Hz to 4 kHz and a time interval of 216 milliseconds.

Parameter Estimation

Our models are highly non-convex and so were prone to many local minima issues. Along with this, starting parameters for optimization greatly affected the ability to determine good fits and so they were set using educated guesses based on pre-processing steps completed on each receptive field. The models shared a set of similar parameters including: f_x, f_y, ϕ_x, ϕ_y, sinusoidal frequencies and phases in the time and frequency axis of the spectrogram and $x_c, y_c, \sigma_x, \sigma_y$, center point and standard deviations of the Gaussian envelope respectively. Extending previous work on parameter estimation in Gabor fitting, we estimated the sinusoidal frequency parameters by determining the maximal power region in the 2-D Fourier transform of the receptive field. The center point parameters for the Gaussian envelope were determined by computing the maximum point of the Hilbert transform. σ_x, σ_y were estimated using the second and first singular value of negative inverse Hessian matrix of the Hilbert transform. The starting parameter for the constant multiplier a was selected based on the mean of the data. For cases of negative harmonic stacks (where $a < 0$) the mean for the two classes of harmonic stacks was linearly separable on normalized receptive fields and so was used to determine the sign of a.

Model Fitting

In order to tackle the problems of local minima we used simulated annealing along with random restarts for our optimization procedure. We setup a nonlinear least squares objective and function minimization was done using a Trust Region minimization algorithm in MATLAB. For each receptive field we conducted 100 separate instances of objective minimization with a set of estimated starting parameters (methods described above) or preset values for starting parameters. The preset values were selected based on manually tweaking the model parameters to get a generally representative shape for most receptive fields. For each minimization instance we add a randomness element to each starting parameter to allow it to vary from its initial value. The iteration that converged to parameters with the lowest sum of square error (SSE) was selected as the best fit parameters for our model. The process was then repeated for different model classes and the model class which resulted in the best fit (determined by the root-mean-square error (RMSE)) amongst the classes was chosen. In order to examine the quality of our fits we used quantitative measures (SSE, RMSE) but also examined the error residuals of our receptive fields to discern whether we were missing out on any structure in the original receptive field.

Acknowledgments

The authors are grateful to Vivienne Ming for her valuable feedback and to Joel Zylberberg for helpful discussions.

References

1. Carlson NL, Ming VL, DeWeese MR (2012) Sparse Codes for Speech Predict Spectrotemporal Receptive Fields in the Inferior Colliculus. PLoS Computational Biology 8: e1002594.
2. Qiu A, Schreiner CE, Escabí Ma (2003) Gabor analysis of auditory midbrain receptive fields: spectro-temporal and binaural composition. Journal of neurophysiology 90: 456–76.
3. Olshausen B, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code natural images. Nature .
4. Rehn M, Sommer FT (2007) A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. Journal of computational neuroscience 22: 135–46.
5. Jones JP, Palmer La (1987) An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. Journal of neurophysiology 58: 1233–58.
6. Karklin Y, Lewicki MS (2009) Emergence of complex cell properties by learning to generalize in natural scenes. Nature 457: 83–6.
7. Theunissen E, Sen K, Doupe AJ (2000) Spectral-Temporal Receptive Fields of Nonlinear Auditory Neurons. Journal of Neuroscience 20: 2315–2331.
8. Woolley SMN, Gill PR, Fremouw T, Theunissen FE (2009) Functional groups in the avian auditory system. The Journal of neuroscience : the official journal of the Society for Neuroscience 29: 2780–93.
9. Sen K, Theunissen FE, Doupe aJ (2001) Feature analysis of natural sounds in the songbird auditory forebrain. Journal of neurophysiology 86: 1445–58.
10. Miller LM, Escabí Ma, Read HL, Schreiner CE (2002) Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. Journal of neurophysiology 87: 516–27.

11. Zylberberg J, Murphy JT, DeWeese MR (2011) A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of V1 simple cell receptive fields. PLoS computational biology 7: e1002250.

12. Escabi Ma, Schreiner CE (2002) Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. The Journal of neuroscience : the official journal of the Society for Neuroscience 22: 4114–31.