

# UC Berkeley

## UC Berkeley Electronic Theses and Dissertations

### Title

Sparse Coding of Speech Data Predicts Properties of the Early Auditory System

### Permalink

<https://escholarship.org/uc/item/5rh069b2>

### Author

Carlson, Nicole

### Publication Date

2012

Peer reviewed|Thesis/dissertation

Sparse Coding of Speech Data Predicts Properties of the Early Auditory System

By

Nicole Liu Carlson

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Physics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael R. DeWeese, Chair

Professor Bruno A. Olshausen

Professor Ahmet Yildiz

Spring 2012

Sparse Coding of Speech Data Predicts Properties of the Early Auditory System

Copyright © 2012

by

Nicole Liu Carlson

## **Abstract**

Sparse Coding of Speech Data Predicts Properties of the Early Auditory System

by

Nicole Liu Carlson

Doctor of Philosophy in Physics

University of California, Berkeley

Professor Michael R. DeWeese, chair

I have developed a sparse mathematical representation of speech that minimizes the number of active model neurons needed to represent typical speech sounds. The model learns several well-known acoustic features of speech such as harmonic stacks, formants, onsets and terminations, but I also find more exotic structures in the spectrogram representation of sound such as localized checkerboard patterns and frequency-modulated excitatory subregions flanked by suppressive sidebands. Moreover, several of these novel features resemble neuronal receptive fields reported in the Inferior Colliculus (IC), as well as auditory thalamus and cortex, and my model neurons exhibit the same tradeoff in spectrotemporal resolution as has been observed in IC. To my knowledge, this is the first demonstration that receptive fields of neurons in the ascending mammalian auditory pathway beyond the auditory nerve can be predicted based on coding principles and the statistical properties of recorded sounds. In my second study, I look at linear filter estimation by creating spike-triggered averages for my model neurons. Surprisingly, whitening does not remove the effect of choosing different probe stimulus sets. This suggests that the type of probe stimulus is very important for uncovering the true receptive field of a neuron.

# Acknowledgments

First and foremost, I would like to thank my adviser Mike DeWeese for working with me for the past four years. Mike has always believed in my work even when it seemed like nobody else did, and had encouraging words whenever I found it hard to keep working.

Secondly, thank you to my unofficial co-adviser, Vivienne Ming. Vivienne has continued working with me even while having two children and founding a start-up. I am consistently amazed with how much she knows and all the time and advice she always has for me. Thank you so much for being an amazing role model.

In terms of technical help with my thesis, I would like to thank Jimmy Wang, Engin Bumbacher, and Liberty Hamilton for help with coding and advice about my project. A huge thank you to everyone in the Redwood Center for supporting me and assisting with all of my computer problems throughout the years. Additionally, thank you to my labmates in the DeWeese Lab for listening to countless lab meetings and for providing pepperoni pizza week after week.

I owe a huge debt to Bruno Olshausen and Ahmet Yildiz for being on my thesis committee. Thank you so much for donating your time to help me graduate.

I also want to recognize the National Science Foundation for providing me with a graduate fellowship. I almost certainly would have dropped out of grad school without the fellowship.

Thank you to every member of the Compass Project. Finding this group when I first came to Berkeley made me feel like I really belonged to a community, and I loved working with all of the physics ‘misfits.’ I’m so proud of all that we’ve accomplished throughout the years.

In terms of my personal life, thank you to all of my great friends both in the Bay Area and spread out across the world. I never would have survived grad school without having all of you giving me support when I was freaking out about my thesis or my code not working. I really appreciate all of the help that you have all given me.

Finally, a big hug goes out to my dog, Cooper. He gave me so much unconditional love and forced me to get out of bed every morning to walk him.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Previous Research</b>	<b>4</b>
2.1	Efficient Coding . . . . .	4
2.2	Sparse Coding in Vision . . . . .	5
2.3	Sparse Coding in Audition . . . . .	6
2.4	Experimental Work in Audition . . . . .	7
2.4.1	Auditory Pathway . . . . .	7
<b>3</b>	<b>Methods</b>	<b>14</b>
3.1	Sparse Coding . . . . .	14
3.1.1	Inference: Local Thresholding Circuit . . . . .	16
3.1.2	Sparsenet . . . . .	18
3.2	Learning . . . . .	18
3.3	Stimuli . . . . .	18
3.3.1	Whitening . . . . .	19
3.4	Analysis . . . . .	19
3.4.1	Presentation of Dictionaries . . . . .	20
3.4.2	Presentation of Experimental Data . . . . .	20
3.4.3	Usage . . . . .	20
3.4.4	Modulation Power Spectra . . . . .	20
3.5	Linear Filters . . . . .	20
3.5.1	Probe Stimulus Sets . . . . .	21

<b>4</b>	<b>Results/Discussion</b>	<b>22</b>
4.1	Cochleogram Dictionaries . . . . .	22
4.1.1	L0-sparse Cochleogram Dictionaries . . . . .	24
4.1.2	L1-sparse Cochleogram Dictionaries . . . . .	29
4.1.3	Sparsenet-trained Cochleogram Dictionaries . . . . .	34
4.1.4	Cochleogram timescale . . . . .	36
4.2	Spectrogram Dictionaries . . . . .	39
4.2.1	L0-sparse Spectrogram Dictionaries . . . . .	39
4.2.2	L1-sparse Spectrogram Dictionaries . . . . .	46
4.2.3	Sparsenet-trained Spectrogram Dictionaries . . . . .	50
4.3	Performance . . . . .	52
4.4	Modulation Power Spectra . . . . .	53
4.5	Analysis of Best Frequencies . . . . .	56
4.6	Comparison to Experimental Data . . . . .	59
4.7	Linear Filters . . . . .	64
4.7.1	The Model . . . . .	64
4.7.2	Standard Spike-Triggered Average . . . . .	64
4.7.3	Whitened Spike-Triggered Average . . . . .	66
4.7.4	Mean Absolute Error . . . . .	68
<b>5</b>	<b>Conclusions</b>	<b>70</b>
<b>A</b>	<b>List of Abbreviations</b>	<b>80</b>

# List of Figures

2.1	Auditory Pathway . . . . .	8
3.1	Schematic illustration of the coding model . . . . .	15
3.2	Local Thresholding Circuit . . . . .	17
3.3	Cost Functions for Sparsenet and the Local Thresholding Circuit . . . . .	17
4.1	Half-complete L0-sparse Cochleogram-Trained Dictionary . . . . .	23
4.2	Complete L0-sparse Cochleogram-trained dictionary . . . . .	26
4.3	Two-times overcomplete L0-sparse Cochleogram-trained dictionary . . . . .	27
4.4	Four-times overcomplete L0-sparse Cochleogram-trained dictionary . . . . .	28
4.5	Half-complete L1-sparse Cochleogram-trained dictionary . . . . .	30
4.6	Complete L1-sparse Cochleogram-trained dictionary . . . . .	31
4.7	Two-times overcomplete L1-sparse Cochleogram-trained dictionary . . . . .	32
4.8	Four-times overcomplete L1-sparse Cochleogram-trained dictionary . . . . .	33
4.9	Half-complete L1-sparse Cochleogram-trained dictionary (Sparsenet) . . . . .	34
4.10	Complete L1-sparse Cochleogram-trained dictionary (Sparsenet) . . . . .	35
4.11	Two-times Overcomplete L0-sparse Cochleogram-trained dictionary with altered timescale . . . . .	37
4.12	Four-times Overcomplete L0-sparse Cochleogram-trained dictionary with altered timescale . . . . .	38
4.13	Select elements from a Half-complete L0-sparse Spectrogram-trained dictionary . .	40
4.14	Half-complete L0-sparse Spectrogram-trained dictionary . . . . .	41
4.15	Complete L0-sparse Spectrogram-trained dictionary . . . . .	42
4.16	Two-times overcomplete L0-sparse Spectrogram-trained dictionary . . . . .	43

4.17	Select elements from a four-times overcomplete L0-sparse Spectrogram-trained dictionary . . . . .	44
4.18	Four-times overcomplete L0-sparse Spectrogram-trained dictionary . . . . .	45
4.19	Half-complete L1-sparse Spectrogram-trained dictionary . . . . .	46
4.20	Complete L1-sparse Spectrogram-trained dictionary . . . . .	47
4.21	Two-times overcomplete L1-sparse Spectrogram-trained dictionary . . . . .	48
4.22	Four-times overcomplete L1-sparse Spectrogram-trained dictionary . . . . .	49
4.23	Half-complete L1-sparse Spectrogram-trained dictionary (Sparsenet) . . . . .	50
4.24	Complete L1-sparse Spectrogram-trained dictionary (Sparsenet) . . . . .	51
4.25	Signal to noise ratios for various dictionaries . . . . .	52
4.26	Modulation power spectra of half-complete cochleogram dictionary and four-times overcomplete spectrogram dictionary . . . . .	53
4.27	Modulation Power Spectra and the Uncertainty Principle . . . . .	54
4.28	Selected elements ordered by Best Frequency . . . . .	56
4.29	Barplots of Median values of Best Temporal and Spectral Modulation Frequency. . . . .	58
4.30	Model comparisons to experimental receptive fields with single excitatory and inhibitory subfields . . . . .	60
4.31	Model comparisons to experimental receptive fields of checkerboards . . . . .	61
4.32	Model comparisons to broad-band experimental receptive fields . . . . .	63
4.33	Model checkerboard receptive field with standard spike-triggered averages . . . . .	65
4.34	Model checkerboard receptive field with whitened spike-triggered averages . . . . .	66
4.35	Model harmonic stack receptive field with whitened spike-triggered averages . . . . .	67
4.36	Model localized excitation/inhibition receptive field with whitened spike-triggered averages . . . . .	68
4.37	Model high-frequency excitation/inhibition patterned receptive field with whitened spike-triggered averages . . . . .	69
4.38	Mean absolute errors for standard and whitened spike-triggered averages . . . . .	69

# Chapter 1

## Introduction

The brain is constantly inundated by sounds rich with information. Our notable ability to interpret the sounds around us suggest the brain must have developed a method for processing and encoding these sounds. It has been postulated that the brain tries to transmit and encode information efficiently, so as to minimize the energy expended [1], reduce redundancy [2, 3, 4], maximize information flow [5, 6, 7, 8], or facilitate computations at later stages of processing [9], among other possible objectives. One way to create an efficient code is to enforce population sparseness, having only a few active neurons at a time. Sparse coding schemes pick out the statistically important features of a signal — those features that occur much more often than chance — which can then be used to efficiently represent a complex signal with few active neurons. Such a representation is robust against noise as well as revealing the underlying structure of the natural stimuli. Additionally, sparse codes are more energy efficient [10]. There is a trade-off between increasing the representational capacity of the network against the energy expended to keep neurons active. Maximizing the ratio of the representational capacity against the energy reveals that the optimum firing rate is low.

The principle of sparse coding has led to important discoveries in vision for the neural encoding of the visual sensory scene. Sparse coding of natural images revealed local, oriented edge-detectors that qualitatively match the receptive fields (RF, the stimulus which most strongly drives a neuron) of simple cells in primary visual cortex (V1) [11]. More recently, overcomplete sparse coding schemes have uncovered a greater diversity of features that more closely matches the full range of simple cell receptive field shapes found in V1 [12]. An encoding is called overcomplete if the number of neurons available to represent the stimulus is larger than the dimensionality of the input. This is a biologically realistic property for a model of sensory processing because information is encoded by increasing numbers of neurons as it travels from the optic nerve to higher stages in the visual pathway [13]. The same is true for the auditory pathway.

Despite experimental evidence for sparse coding in the auditory system [14, 15], there have been fewer theoretical sparse coding studies in audition than in vision. However, there has been progress, particularly for the earliest stages of auditory processing. Sparse coding of raw sound pressure level waveforms of natural sounds produced a “dictionary” of acoustic filters closely

resembling the impulse response functions of auditory nerve fibers [16, 17]. Acoustic features learned by this model were best fit to the neural data for a particular combination of animal vocalizations and two subclasses of environmental sounds. Intriguingly, they found that training on speech alone produced features that were just as well-fit to the neural data as the optimal combination of natural sounds, suggesting that speech provides the right mixture of acoustic features for probing and predicting the properties of the mammalian auditory system.

Another pioneering sparse coding study [18] took as its starting point speech that was first preprocessed using a model of the cochlea — one of several so-called cochleogram representations of sound. This group found relatively simple acoustic features that were fairly localized in time and frequency as well as some temporally localized harmonic stacks. These results were roughly consistent with some properties of receptive fields in primary auditory cortex (A1), but modeled responses did not capture the specific shapes of reported neuronal spectrotemporal receptive fields (STRFs; [19]). That study only considered undercomplete dictionaries, and it focused solely on a “soft” sparse coding model that minimized the mean activity of the model’s neurons, as opposed to “hard” sparse models that minimize the number of active neurons. The same group also considered undercomplete, soft sparse coding of spectrograms of speech [20], which did yield some STRFs showing multiple subfields and temporally modulated harmonic stacks, but the range of STRF shapes they reported was still modest compared with what has been seen experimentally in auditory midbrain, thalamus, or cortex. Another recent study considered sparse coding of music [21] in order to develop automated genre classifiers.

To my knowledge, there are no published studies of complete or overcomplete, sparse coding of either spectrograms or cochleograms of speech or natural sounds. I note that one preliminary sparse coding study utilizing a complete dictionary trained on spectrograms did find STRFs resembling formants, onset-sensitive neurons, and harmonic stacks [46] but they did not obtain novel acoustic features, nor any that closely resembled STRFs from the auditory system.

My goal is two-fold. First, I test whether an overcomplete, hard sparse coding model trained on spectrograms of speech can more fully reveal the structure of natural sounds than previous models. Second, I ask whether my model can accurately predict receptive fields in the ascending auditory pathway beyond the auditory nerve. I have found that, when trained on spectrograms of human speech, an overcomplete, hard sparse coding model does learn features resembling those of STRF shapes previously reported in the Inferior Colliculus (IC), as well as auditory thalamus and cortex. Moreover, my model exhibits a similar tradeoff in spectrotemporal resolution as previously reported in IC. Finally, my model has identified novel acoustic features for probing the response properties of neurons in the auditory pathway that have thus far resisted classification and meaningful analysis.

The second part of my research involved probing my theoretical basis functions in the same way experimentalists treat real neurons. A sensory neuron is characterized via its RF; an experimental construct defined as the stimulus that most strongly causes the neuron to fire. The simplest method of estimating a receptive field is through the spike-triggered average (STA) [22, 6, 23]; the average of all stimuli immediately preceding spikes. The STA is equivalent to the first-term in the Volterra kernel or Wiener kernel expansions, and this technique is also called reverse cor-

relation [24]. Additionally, this is the maximum likelihood estimate for the RF. The STA is valid if the relationship between stimulus and spiking is linear. This method has been used to estimate tuning curves in the auditory system and receptive fields in the visual system [25, 26, 27]. More recently, researchers have estimated STRFs in the auditory system using STAs and related measures [28, 29, 30, 31, 32, 33]. Technically, the STA is only accurate for white noise so this method is strongly stimulus dependent. A more sophisticated approach is the linear filter estimate or whitened spike-triggered average in which stimulus correlations are removed [34, 35, 36, 37].

Theoretically, these methods are only guaranteed to provide accurate predictions in the limit of infinite stimuli. There are many caveats about the experiments as well. An auditory recording experiment might only last for half an hour depending on the type of electrode, the type of anesthesia if any, or whether or not the animal was awake and behaving. The neurons sometimes fire at very low rates so if the stimulus that matches the receptive field of the neuron is not found, the estimate will be really noisy. This is a particular problem in the case of auditory neurons which typically have lower firing rates than vision neurons.

A well-known result is that different stimulus sets produce different receptive fields for the same neuron [19]. Historically, it has been difficult to determine whether these differences are actually due to the stimulus sets and if so which stimulus sets are actually uncovering the true RF structure or to the aforementioned experimental difficulties. I am in a unique position to test these two effects because my sparse coding model of auditory receptive fields provides me with ground truth.

I treated my model neurons as real neurons and “played” various probe stimulus sets to the neurons and used them to produce linear filter estimates. I compared STAs from the different probe stimulus sets with and without correcting for stimulus correlations (whitened STAs). For some cell-types, all stimuli can uncover the underlying auditory feature. Other cell-types have different estimations depending on the stimulus.

This unified work in the sparse coding of speech sounds has produced some remarkable comparisons to experimental data. I believe that my research has demonstrated that some features of the auditory system can be predicting by using the principles of efficient coding.

The dissertation is organized as follows. I discuss previous research (Ch. 2) with a literature review of both theoretical and experimental work. The methods chapter (Ch. 3) describes my model and analyses. I discuss my results in Ch. 4, and finish with conclusions and future work in Ch. 5.

# Chapter 2

## Previous Research

### 2.1 Efficient Coding

For almost sixty years, scientists have postulated that the brain tries to encode sensory information efficiently. In 1954, Attneave [2] asserted that a major function of the sensory system is to remove the redundancy of stimuli by looking for correlations in the stimuli. One analogy is that if lower-level neurons fire for letters than a higher-level neuron might encode for a single word. In this case, there is not much information in a single letter; the redundancy is that all of the letters are needed to understand the meaning of the word. He considered the information theoretic efficient way to transmit images as an example of this principle. Another pioneer, Barlow [3] argued that to understand the brain, we need to study the natural environment, and he organized the ideas into five principles [38]:

1. It is important to look at the interactions on the cellular level instead of a more macroscopic or microscopic scale.
  2. The sensory system is organized to completely represent the sensory input with as few active neurons as possible.
  3. The stimuli contain trigger features that are matched to redundant patterns of stimulation.
  4. Perception is due to the activity of a small number of high-level neurons from a much larger population.
  5. The more frequently the neuron fires, the more certainty that the trigger feature is present.
- Barlow codified this framework in terms of the capacity of a channel to transmit information.

These principles were implemented in neural networks starting in the 1990s. Typically, associative neural networks follow Hebb's rule [39] commonly stated as: 'fire together, wire together' meaning that if two neurons are active at the same time, then the connection between them is strengthened. This rule allows a neural network to store patterns, but many stimuli will cause large numbers of neurons to be active. Foldiak [9] created a network that was capable of storing patterns sparsely, having only a few neurons active. This was achieved by having an anti-Hebbian learning rule in which neurons that fired together inhibited one another, and the one with the strongest

activation inhibited the others the most. Importantly though, the inhibition was not complete making this different from a winner-take-all rule in which the most active neuron completely shuts down the other neurons. Foldiak's network could store a large number of patterns while balancing the battle between small memory capacity where each unit stores one item (also known as a grandmother cell), and the large capacity that arises when many cells are active. This first implementation of sparse coding had the inhibition put in ad hoc.

Levy and Baxter [10] studied energy efficient neural codes by examining coding efficiency as well as the metabolic cost of active neurons. There is an intrinsic cost of a neuron recovering from spiking compared to a neuron that stayed inactive. They examined a theoretical ratio of the information capacity of a set of neurons and the energy cost of having a certain fraction of active neurons. They found that even for different spike cost functions, the optimal active fraction is quite low. Importantly, their result was independent of the number of neurons. For binary neurons, they found that having 2-16% of the neurons active is the most energy efficient.

## 2.2 Sparse Coding in Vision

A study of the statistics of images of natural scenes [40] found that the amplitude spectra of any set of natural images falls as the reciprocal of the spatial frequency. Moreover, the response properties of visual cortical cells, Gabor functions (Gaussians multiplied by a sinusoids) were found to be well-suited to the statistics of natural images [40]. However, there was no explanation of why V1 RFs should be Gabor functions.

A review of the literature [41] on the statistics of natural images noted that the most important result is that images are scale invariant meaning that images contain structure at every scale. In other words, natural images contain higher-order structure and are not Gaussian (containing only second-order structure). A common approach to analyzing data is Principal Components Analysis (PCA). However, PCA inherently assumes Gaussian data. The principal components of images are non-localized basis functions which do not resemble the experimentally found V1 receptive field shapes. Based on efficient coding, the brain will take advantage of the higher-order structure so a sparse coding scheme should perform better than PCA.

The first principled explanation for Gabor filter shapes was sparse coding. Neural responses were predicted based on the structure of natural images (photographs) [11], [42]. Sparse coding creates a representation which has high fidelity balanced with having only a few active coefficients. An early method was a linear generative model called Sparsenet [11] which created an L1-sparse (minimizing the average activity of the units) complete dictionary. The dictionary elements resembled V1 RFs and were parameterized with Gabor functions. The parameter values of the model elements had a similar distribution of that from experimentally measured receptive fields in the mammalian visual cortex. Specifically, Sparsenet reproduced those receptive fields which resembled edge detectors; they are spatially localized, oriented, and bandpass filtered.

An extension of this work was a study of sparse overcomplete representations [43]; a representation in which there are more basis functions than the dimensionality of the input. Unlike

a complete representation, there is no longer a unique encoding of each input. An overcomplete basis can give a better approximation of the statistical distribution of the data as well as greater coding efficiency.

One experimental group [44] found that Sparsenet did not capture the full distribution of the experimental receptive fields. In particular, Sparsenet missed some types of receptive fields, such as those which were blob-like or which contained many subfields. A more recent model, Sparse Set Coding [12], created an L0-sparse three-times overcomplete dictionary whose properties more closely matched the full range of experimental shapes. This dictionary produced receptive fields which more strongly resembled real visual receptive fields (as examined by the distributions of Gabor function parameters). Additionally, another new algorithm was developed that allows an approximation to L0-norm sparse coding [45].

## 2.3 Sparse Coding in Audition

Initial work in auditory sparse coding focused on modeling the Auditory Nerve (AN) (Figure 2.4.1) [17], [16]. The initial study [16] used Independent Components Analysis (ICA) with a fixed number of time samples on either a mixture of environmental sounds and animal vocalizations or speech data to produce filters that qualitatively matched experimental AN reverse correlation filters, the average stimuli that precede neuronal spikes. Lewicki noted that only the mixture of natural sounds or speech produced filters that matched the physiological shapes. Training only on environmental sounds produced a wavelet-like basis, and only on animal vocalizations produced a Fourier basis. A subsequent study [17] extended this work by allowing the time length of each filter to change during the learning process. Smith and Lewicki utilized raw sound pressure waveforms and trained on natural sounds or speech. The resulting basis functions quantitatively matched experimental auditory reverse correlation filters. The research confirmed quantitatively that a certain ratio of transient and ambient environmental sounds and animal vocalizations produced kernels that matched those trained solely on speech. Examining plots of bandwidth versus center frequency showed that the model filter properties were almost identical to the experimental data. They also noted that their spike code was more efficient at encoding the data than a Fourier or wavelet basis.

One study [20] modeled A1 receptive fields by forming a sparse representation of speech data that had been pre-processed by a filterbank designed to mimic the properties of the early auditory system. They found that their STRFs were time-frequency localized with some qualitative agreement with A1 STRF properties, but no detailed matches. Later, Klein et al. [18] encoded speech data with an L1-sparse, half-complete basis using a cochleogram representation. A cochleogram is a frequency decomposition of a waveform in which the properties of the frequency filters mimic those of the cochlea. Their dictionary elements were generally highly localized in frequency and slightly less-localized in time. The typical extent of their STRFs was 100-200 ms for the temporal localization and 0.5-3 octaves in spectral localization. These receptive fields included harmonic stacks and formants, resonances of the human vocal tract (basically a modulation of a harmonic

stack) that produce characteristic shapes in a spectrogram. Again, there was some qualitative agreement with A1 STRFs but no detailed matches. Replication of this study was the starting point of my research.

One more recent work examined correlated subspaces of the higher-level features of sound [46]. They trained waveform, spectrogram, and cochleogram dictionaries. Different filters of the waveform dictionary exhibited phase invariance, shift invariance, or bandwidth invariance. The spectrogram dictionary displayed elements that coded for onsets or offsets, harmonic stacks, and formants.

## **2.4 Experimental Work in Audition**

I now review the anatomy of the mammalian auditory pathway before describing experimental work which characterizes the auditory system.

### **2.4.1 Auditory Pathway**

Sound enters the ear as a raw sound pressure waveform. The signal is transformed from air vibrations to vibrations in fluid. The mechanical vibrations are transmitted into an electrical signal through the hair cells of the cochlea (**Fig. 2.4.1**). These hair cells are all triggered by specific frequencies which preserves the information in the sound. In humans, there are approximately 3500 inner hair cells and 12,000 outer hair cells. This signal is carried by 30,000 auditory nerve fibers which perform a frequency decomposition of the sound. The information passes through the cochlear nucleus to the superior olive where information from both ears is combined. This feeds into 392,000 midbrain neurons (Inferior Colliculus (IC)). The midbrain connects to the thalamus (ventral division of the Medial Geniculate Body (MGBv)). The information is then passed into the primary auditory cortex (A1) which contains approximately 100 million neurons. An important feature is that the information is represented by increasing numbers of neurons at each stage in this process. Although I have described this is a linear pathway, there is feedback between almost all of the areas mentioned [47]

Neurons in IC, MGBv, and A1 have recently been characterized via STRFs with a variety of stimuli. Original work used simple stimuli like pure tones and white noise. To drive the neurons strongly, experimentalists developed more complicated stimuli like natural sounds and dynamic moving ripples (to be described below).

### **Features of Auditory Receptive Fields**

I first review the types of features found in RFs. Many groups have found vastly different and sometimes contradictory results.

To find evidence for the efficient coding hypothesis, Rieke [48] investigated if the auditory system preferentially encodes natural stimuli. For bullfrog auditory nerves, naturalistic stimuli

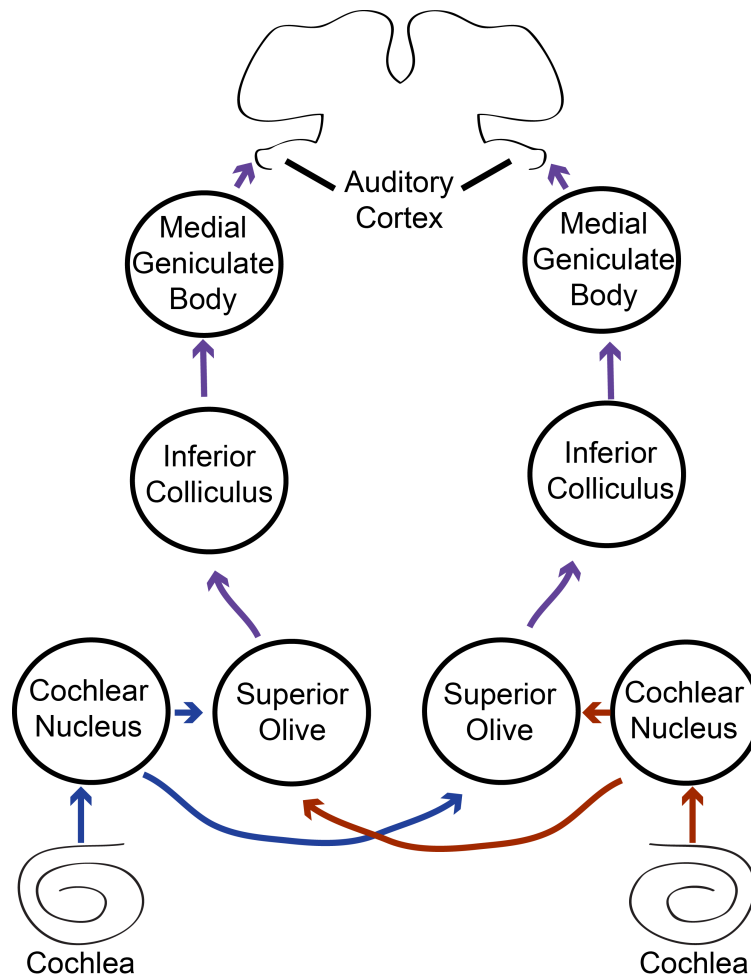


Figure 2.1: Diagram of Auditory Pathway. Sound enters the ear and is then transmitted into the cochlea. From the cochlea, the signal travels along the cochlear nerve to the cochlear nucleus and superior olive where sound is combined from both ears. The next stage is the Inferior Colliculus (IC) followed by the Medial Geniculate Body (MGB). After some processing, the signal is sent to the primary auditory cortex (AC) where higher-level processing occurs. Red and blue represent signal from the separate ears, and purple indicates that information from both ears has been combined.

but not white noise stimulated the auditory nerve fibers close to the physical limits for maximum information transmission. This suggests that the auditory system's properties are matched to the statistics of natural sounds. In a related study, one group [49] found that A1 neurons tended to exhibit 'envelope locking' where they match the co-modulations of different frequency bands. This means that they modulate their firing rates coherently with the temporal envelope. Unmodulated noise bands did not show the same result. Perceptually, co-modulated noise helps listeners to detect tones in noise. Their result suggests that A1 neurons can detect tones better with co-modulated noise. They argued that regularities in the auditory environment should be reflected in the way auditory neurons respond.

One initial A1 STRF study [33] discovered that neurons respond to 'edges' in the spectrogram representation. Their stimuli were random chords in time or asynchronous Poisson tone trains instead of more natural sounds. They claimed that the cortex decomposes the auditory scene into component parts like the visual cortex.

STRFs can change with specific task demands and salient sensory cues [50]. They trained ferrets to detect a target tone, and found that during the task, there were localized changes in the STRF shape that sometimes lasted for hours. The most common STRF change was facilitation by enhanced excitation or decreased inhibition of the target tone implying that it was additive facilitative gain. This study suggests cortical RFs are not fixed but may be constantly adapting and reorganizing.

A1 neurons responded to wideband sounds with either low or high spectral contrast (this refers to the sound level distribution across different frequencies) [51]. High contrast neurons are like linear filters, but low contrast neurons had nonlinear selectivity to spectral properties. This suggests that there are strong nonlinearities in A1.

Nelken [52] described the diverse properties of A1 neurons including their precision and the linearity. He argues that the role of A1 is to detect objects by organizing the distinctive features of an object. Because of this, he believes that IC should be considered the auditory analog of V1 because IC and thalamus seem to do feature detection in the auditory pathway.

Escabí and Read [53] reviewed properties of the auditory midbrain, thalamus, and cortex. They noted that cochleotopy (topographic organization of frequency sensitivity) is a fundamental aspect of the lemniscal auditory system from CN to IC. Neurons can be monotonic or non-monotonic to sound level dependence. Inhibitory processes are established early on in the Cochlear Nucleus and further refined at more central centers. Spectral tuning is shaped by flanking inhibition. Temporal modulation rates are higher in IC and lower in thalamus and cortex. There is a trade-off between spectral and temporal integration resolution.

One of the only studies to find high firing rates in A1 [54] found that this occurs when neurons are driven with their preferred stimuli. They argued that a subset of the neuronal population fires strongly during the entire stimulus while others only have transient responses. Most A1 neurons were driven more strongly by amplitude or frequency modulated tones than by pure tones and had well-defined best modulation frequencies.

In bat IC neurons, STRFs revealed selectivity for spectral motion [55]. Most neurons were se-

lective for downward Frequency-Modulated (FM) sweeps which reflects the fact that most species calls have downward-sweeping FM components. The most common STRF shape was a clear excitatory region surrounded by inhibition. The group used the STRFs to predict the calls to which neurons respond. The main contributor to direction selectivity was spectrum-time inseparability.

Nagel and Doupe [56] calculated STRFs in the primary auditory field L (analogous to A1 in mammals) of unanesthetized zebra finches. At low sound intensities, the neurons behaved like simple detectors for specific spectral and temporal modulation frequencies. At high intensities however, the neurons responded to differences in sound energies along their preferred direction. Unlike the previous study, they found few cells with direction selectivity, which might reflect the over-representation of sweeps in bats due to the nature of their communications.

To find experimental evidence for A1 using a sparse representation, one group [15] did cell-attached recordings in A1. For most sounds only 5% of neurons were active and that their activities followed a log-normal distribution instead of the typical exponential one. This matches a model in which neurons are silent most of the time and the sound is represented with a small subset of highly active neurons. Even though only a few were active, this can lead to as much as a 50% increase in the mean activity. They used tones, sweeps, white noise bursts, and natural sounds. This method didn't have the bias of choosing highly active neurons, but rather neurons based on their 'patchability'.

Schecter et al. [57] looked at two neuron subpopulations in IC. Lagged cells which had inhibition preceding excitation, and non-lagged cells which have excitation first. They predict that lagged cells have inhibitory feedback mediated by cortical feedback projects. The distributions of latencies of the two groups overlapped. They also noted that combining the output of these two groups in A1 could produce direction selective cells.

Lesica and Grothe [30] examined dynamic spectrotemporal feature selectivity in the auditory midbrain. They investigated the effect of high and low SPL on gerbil IC neurons with the sound of rain. They found that for a given stimulus, the STRF provides an accurate characterization of the feature selectivity of a neuron. At higher SPL, neurons displayed more inhibition. They hypothesized that the change was due to an operation of a static nonlinear system.

Auditory thalamic neurons respond to more complex features than midbrain neurons [58]. The auditory thalamic circuitry plays an important role in generating novel complex representations for specific features found in natural sounds.

I report on one study in an area lower than IC as it will later relate to my results. Clopton and Backoff [59] discovered harmonic stacks in the STRFs of guinea pig Dorsal Cochlear Nucleus (see Fig. 5 of their paper), a feature which has yet to be reported in higher levels.

## **STRF Estimation Methods and the Linearity of Auditory Neurons**

A common caveat given about a STRF is that it only reflect the linear part of a neuron's response function. It is well-known that the linear model is not a complete description of the neuron, but some portion of the neuron's response can be explained. Over time, many groups have refined STRF

estimation methods as well as assessing the linearity of auditory neurons. The STRF is equivalent to the STA if 1) the stimulus space includes all stimuli that make the neuron fire, 2) there is random and uniform sampling of the space and 3) the multiple dimensions used to represent the stimuli are independent.

STRFs from the avian auditory forebrain found by natural sounds are better response predictors than those from other sound sets [29].

Qiu et al [60] modeled IC STRFs as time-frequency separable Gabor functions. The parameterizations were good descriptions for about 60% of the neurons, and most of their cells were either purely separable or weakly inseparable. However, the model could not fully account for multiple excitatory spectral peaks such as those that would be found in harmonic stacks. None of the neurons had highpass response properties as predicted by Singh and Thenissen [61]. The stimulus they used were dynamic moving ripples. Each ripple is a spectrotemporal modulation frequency. This probes the relevant range of temporal and spectral modulations in a random fashion. They look like moving gratings in spectrogram space. These sounds fully span the temporal and spectral modulation space of natural sounds, and thus any sound can be expressed as a linear sum of ripples.

Machens et al. [62] examined the linearity of auditory neuron receptive fields, and found that only 11% of the response power of the subthreshold membrane potential can be predicted by the STRF. This implies that A1 neurons are highly nonlinear. They used natural sounds including animal sounds, environmental sounds, and Jimi Hendrix. The paper also developed a new STRF estimation technique in which STRFs had the correlations removed and were regularized with two constraints. Regularization penalizes the STRF parameters when they stray from zero or when neighboring ones are far from each other. The study also showed subthreshold STRFs are typically more extended temporally and spectrally than spiking STRFs.

An A1 STRF is a robust linear predictor of the cell [63]. They looked at three stimuli: dynamic moving ripples, spectrotemporal white noise, and Temporally Orthogonal Ripple Combinations (TORCs, these consist of several spectrotemporal modulation frequencies and span the entire modulation space like DMR). They looked at 250ms STRFs, and found that STRFs from all three of these stimuli were ‘remarkably similar’. For predictions, TORCs were the best. DMR had high Signal-to-Noise Ratio (SNR), but required many stimulus presentations to be calculated.

Another recent study [64] looked at methods of STRF estimation and compared four different factors in the estimation: 1) the choice of logarithmic or linear filter frequency spacing, 2) the time-frequency scale, 3) the stimulus amplitude compression (for example, putting all of the values through a logarithm function), and 4) the inclusion of adaptive gain control (AGC). AGC simulates the way that the ear keeps cochlear outputs at set levels. The latter two were found to be the most important factors. One thing to note is that the estimated STRF will depend on the stimulus unless the cell has no odd-order nonlinearities. They looked at STRFs from adult male zebra finch songs and samples of synthetic noise. Overall, the logarithmic amplitude scale was the best as it more closely matched a spectrogram. The AGC was always important.

Atencio et. al [65] categorized STRFs by looking for the stimulus dimensions with the maxi-

mum mutual information between the stimulus and response. They found that the first maximally informative dimension (MID) matched well with the STA. Taken together, the joint nonlinearity between the first two MIDs showed that they were synergistic. This method works for non-Gaussian stimuli, and they claimed that MIDs are more likely to be stable across stimuli than STAs.

## Modulations

Another area of study are the modulations of STRFs. The Modulation Power Spectra (MPS) is found by taking the 2D Fourier Transform of a STRF. It is often analyzed in terms of the peak or best spectral or temporal modulations as well as its separability. This refers to an MPS that can be described as a spectral modulation function (SMF) multiplied by a temporal modulation function (TMF).

Miller et al. [32] used DMR while recording simultaneously in the auditory thalamus and cortex in anesthetized cats. They compared STRF population properties of the two areas. The upper cutoff for spectral modulations was 1.3 cycles per octave (cyc/oct) in thalamus and 1.37 cyc/oct in A1. For temporal modulations, the mean rate in the thalamus was 32.4 Hz and 16.6 Hz in A1. The upper cutoff for temporal modulations was thalamus 62.9 Hz and 37.4 Hz in A1. The populations showed no directional bias towards positive or negative asymmetries. A neuron is often characterized by its best frequency (BF), the frequency that elicits the strongest response. These can be found using tuning curves or taken as the maximal value of a STRF. The best frequencies of both areas showed similar ranges, and there was a slight preference for stimulus energy onsets. The thalamus was more sharply tuned for temporal modulations whereas cortical cells were more low-pass. This work demonstrated that the cortex does not just inherit the properties of the thalamus

The MPS of natural sounds [61] has also been examined to make predictions about the auditory system. Because of the uncertainty principle, sounds cannot have rapid temporal and spectral modulations simultaneously. However, natural sounds have a characteristic shape even beyond this limit. Natural sounds are typically low-passed in temporal modulations and low for spectral modulations. The researchers postulated the STRFs should reflect these properties and have mostly low best spectral and temporal modulations. Most power is in the lower modulation frequencies and the power decays along the modulation axes following a power law.

Another study [66] looked at the ability of midbrain STRFs to encode birdsong versus modulation-limited noise (ML-noise). ML-noise is white noise where the spectral and temporal modulations are limited by the maximal spectral and temporal modulations of the song data. Of the neurons, 91% had different STRFs depending on which stimulus was used to create the STRF. Natural-like sounds had responses with higher information rates. Frequency tuning was broader and temporal tuning more precise during song processing. The neurons that were similar under both stimuli had extremely tight frequency tuning, although the best frequencies were not stimulus dependent. At the population level, the STRFs were able to capture the majority of the response behavior in midbrain neurons.

A more recent study looked at the comprehension of speech by human listeners by altering

its modulations [67]. Comprehension was impaired when temporal modulations below 12 Hz and spectral modulations below 4 cyc/kHz were removed. Gender characteristics were most important between 3-7 cyc/kHz. This suggests that some neurons must respond to those frequencies for human comprehension.

Rodriguez et al. [31] looked at the spectral and temporal modulation tradeoff in the IC. They found that the tradeoff is topographically ordered in the IC, and that the properties were not just inherited from the AN. Faster temporal modulations were found for low BFs and slower for high BF. Low-frequency sites had short integration times and spectrally broad STRFs. For high frequency BF, the STRFs had longer integration times and were spectrally narrow. For the vast majority of IC units, the time-frequency resolution product was greater than for AN fibers and more than an order of magnitude from the theoretical limit from the uncertainty principle. AN fibers exhibit only low-pass TM tuning, and IC has both band-pass and low-pass tMTFs. Clearly, some other mechanism must be taking place to account for the changed properties in IC relative to AN.

## **Separability**

Another major subject of research is the separability of the MPS.

A more recent study [28] quantified the separability of receptive fields; that is whether the STRF can be expressed as a spectral transfer function multiplied by a temporal transfer function. If a STRF is separable, then the neuron responds equally to upward and downward ripples. A related property is quadrant separability in which the MPS is separable in either the upward or downward spectral quadrant. Quadrant separable neurons have an asymmetric response to direction, and can code for direction selectivity. STRFs commonly had excitatory and inhibitory regions usually with side inhibition. Some neurons were inseparable, but there was no clear cutoff between separable and inseparable neurons; rather there was a continuum of separability. The best RFs typically had temporal modulations between 4-16 Hz and spectral modulations lower than 2 cycles per octave. Temporal but not spectral transfer functions were relatively symmetric so it appears that most inseparability is due to the spectral transfer function.

Auditory neurons respond to features that enhance the acoustic differences between classes of natural sounds [68]. For this group, the linear portions of responses ranged between 30-81%. The modulation tuning helps efficient coding in four ways: 1) The midbrain and forebrain neurons selectively filter spectral modulations such that the modulations relevant to natural sounds, the lower ones, are encoded. 2) The tuning attenuates the low frequencies that are redundant in natural sounds. 3) The maximal gain sensitivity of temporal modulation frequencies are those that vary the most among classes of natural sounds. 4) The ensemble temporal tuning whitens the temporal modulation power function which increases the bandwidth of the neural response to signals from within a natural sound class.

# Chapter 3

## Methods

This chapter describes the model used to encode speech data sparsely and the analysis I performed on my results.

An overview of the model is illustrated in **Fig. 3.1**, and more details of all steps will be given in subsequent sections. Raw sound pressure level waveforms of recorded human speech are first preprocessed by one of two simple models of the peripheral auditory system. The first of these preprocessing models is the spectrogram, which is the power spectrum of short segments of the original waveform at each moment in time. The alternative preprocessing step used is meant to more accurately model the cochlea [69, 70]; the original waveform is sent through a filter bank whose center frequencies are based on the properties of cochlear nerve fibers. Both models produce representations of the waveform as power at different frequencies over time. The spectrograms (or cochleograms) are then separated into segments of length 216 ms (250 ms). Because of the high dimensionality of the data, I perform principal components analysis and retain the first two hundred components to reduce the dimensionality. I then train a dictionary that can encode this data using the Local Thresholding Circuit (LTC), a sparse encoding algorithm [45]. The flexible algorithm allows us to enforce either L0 sparseness (minimizing the number of active dictionary units) or L1 sparseness (minimizing the absolute activity of all of the dictionary elements) during encoding by choice of thresholding function. Additionally, I explore the effect of dictionary over-completeness (with respect to the number of principal components) by training dictionaries which are half-complete, complete, or over-complete (two or four times). Following training, the various resulting dictionaries were analyzed for cell-types and compared to experimental receptive fields reported in the literature.

### 3.1 Sparse Coding

In sparse coding, the input  $y$  (spectrogram or cochleogram) is encoded as a matrix  $A$  multiplied by a vector of weighting coefficients  $s$ :  $y = As + \epsilon$  where  $\epsilon$  is the error. Each column of  $A$  represents one dictionary element or receptive field, the stimulus that most strongly drives the unit.

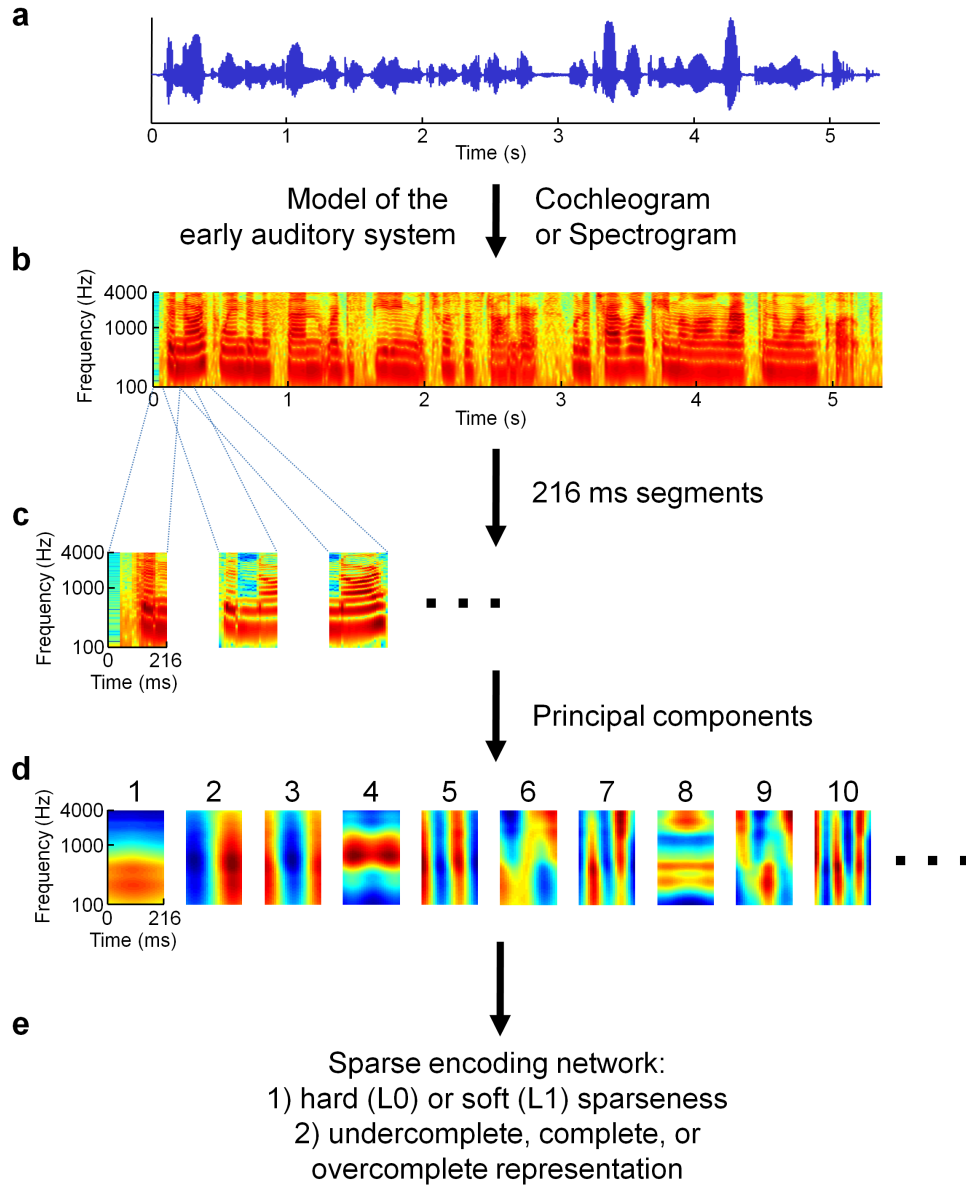


Figure-1

Figure 3.1: Schematic illustration of our sparse coding model. **(a)** Stimuli used to train the model consisted of examples of recorded speech. The blue curve represents the raw sound pressure waveform of a woman saying, “The north wind and the sun were disputing which was the stronger, when a traveler came along wrapped in a warm cloak.” **(b)** The raw waveforms were put through either a spectrogram or a “cochleogram” (not shown; see Methods for details). In either case, the power spectrum across acoustic frequencies is displayed as a function of time, with warmer colors indicating high power content and cooler colors indicating low power. **(c)** The spectrograms were then divided into overlapping 216 ms segments. **(d)** Subsequently, principle components analysis (PCA) was used to project each segment onto the space of the first two hundred principal components (first ten shown). **(e)** These projections were then input to a sparse coding network in order to learn a “dictionary” of basis elements analogous to neuronal receptive fields, which can then be used to form a representations of any given stimulus (*i.e.*, to perform inference).

If there are more columns in  $\mathbf{A}$  than elements in  $\mathbf{y}$ , this will be an overcomplete representation. I defined the degree of overcompleteness relative to the number of principal components. I learned the dictionary and inferred the coefficients by descending an energy function that minimizes the mean squared error of reconstruction under a sparsity constraint.

$$E(t) = \frac{1}{2} \|\mathbf{y}(t) - \mathbf{A}\mathbf{s}(t)\|^2 + \lambda \sum_m C(s_m(t)). \quad (3.1)$$

Here  $\lambda$  controls the relative weighting of the two terms and  $C$  represents the sparsity constraint or cost function (**Fig. 3.3**).

The sparsity constraint requires the column vector  $\mathbf{s}$  to be sparse by some definition. I focus on the L0-norm, minimizing the number of non-zero coefficients in  $\mathbf{s}$  (or equivalently the number of active neurons in a network). Another norm I have investigated is the L1-norm, minimizing the absolute activity of all of the neurons.

### 3.1.1 Inference: Local Thresholding Circuit

I performed inference of the coefficients with a recently developed algorithm, the Local Thresholding Circuit [45], which minimizes close approximations of either the L0- or L1-norms (**Fig. 3.2**). Each basis function  $\mathbf{A}_i$  is correlated with a computing neuron defined by an internal variable  $u_i$  as well as the output coefficient  $s_i$ . All of the units begin with the coefficients set to zero. These values change over time depending on the input. A unit  $u_i$  increases by an amount  $b_i$  if the input overlaps with the receptive field of the neuron:  $b_i(t) = \langle \mathbf{A}_i, \mathbf{y}(t) \rangle$ . The neurons evolve as a group following dynamics in which the units compete with one another to represent the input. The units inhibit each other with the strength of the inhibition increasing as the overlap of their receptive fields and the output coefficient values increase. This internal variable is then put through a thresholding function  $T_\lambda$  to produce the output value:  $s_i = T_\lambda(u_i)$ .

In vector notation, the full dynamic equation of inference is:

$$\begin{aligned} \dot{\mathbf{u}}(t) = f(\mathbf{u}(t)) &= \frac{1}{\tau} [\mathbf{b}(t) - \mathbf{u}(t) - (\mathbf{A}^T \mathbf{A} - I)\mathbf{s}(t)], \\ \mathbf{s}(t) &= T_\lambda(\mathbf{u}(t)). \end{aligned} \quad (3.2)$$

The variable  $\tau$  sets the time-scale of the dynamics.

The thresholding function  $T_\lambda$  is determined by the sparsity constraint  $C$ . It is specified via the following equation:

$$\lambda \frac{dC(s_m)}{ds_m} = u_m - s_m = u_m - T_\lambda(u_m). \quad (3.3)$$

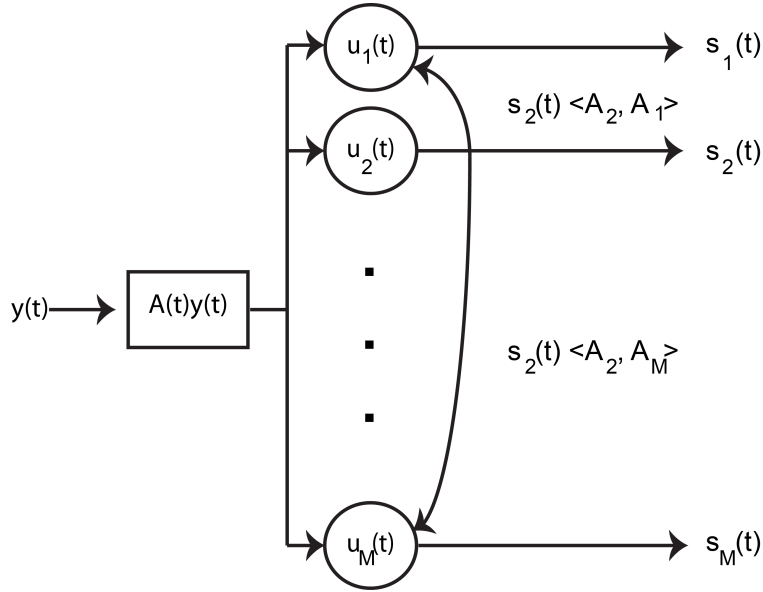


Figure 3.2: Diagram of the Local Thresholding Circuit. The input  $y$  is sent into the  $u$  level of the network. Internal coefficients  $u$  (initialized to zero) build up as their overlap with the signal  $b_i(t) = \langle \mathbf{A}_i, \mathbf{y}(t) \rangle$  increases. These internal values are transformed via a thresholding function to the external ones  $s$  provided that they pass a certain threshold. The units inhibit one another with the strength given by  $s \langle \mathbf{A}^T, \mathbf{A} \rangle$ . Here I show the second unit inhibiting the other two as an example.

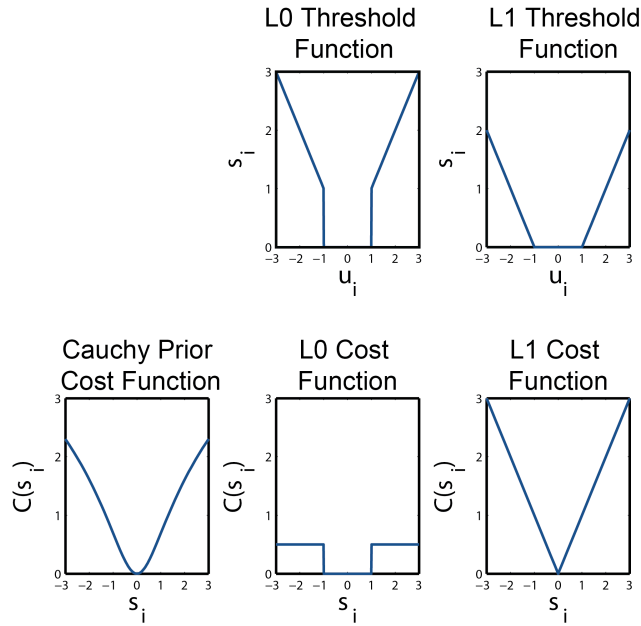


Figure 3.3: Various Cost (Sparsity Constraints) and Thresholding functions of the Local Thresholding Circuit. Left: Cauchy Prior Cost Function, a soft form of sparseness. This is the type of sparseness used in Sparsenet. Middle: L0 Threshold and Cost Functions. Right: L1 Threshold and Cost Functions. Note that the latter two have values of  $u_i$  that produce zero-values of  $s_i$ . With the Cauchy prior cost function, the values of  $s_i$  are always non-zero.

### 3.1.2 Sparsenet

I have also trained dictionaries using the original sparse coding network: Sparsenet. Sparsenet is similar to LTC, but there are no internal coefficients, and the form of sparseness is softer. They tested cost functions including an exponential, sigmoid, and absolute value. MATLAB code is available online from Bruno Olshausen's website ([www.redwood.berkeley.edu](http://www.redwood.berkeley.edu)). Here, I used a sigmoid function to approximate the L1-norm.

## 3.2 Learning

Learning is done via gradient descent on the energy function:

$$\begin{aligned} \mathbf{r}(t) &= \mathbf{y}(t) - \mathbf{A}\mathbf{s}(t), \\ \mathbf{A} &= \mathbf{A} + \eta_{\mathbf{A}}(\mathbf{r}(t)\mathbf{s}^T(t)) + \theta(\mathbf{A} - \mathbf{A}\mathbf{A}^T\mathbf{A}). \end{aligned} \tag{3.4}$$

The  $\theta$  term is a device for increasing orthogonality between basis functions [71]. This is equivalent to adding in a prior that the basis functions are unique.

## 3.3 Stimuli

I used two corpora of speech recordings from the handbook of the International Phonetic Association (<http://ww2.arts.gla.ac.uk/IPA/sndfiles.html>) and TIMIT [72]. These consist of people telling narratives in approximately thirty different languages. We resampled all waveforms to 16000 Hz, and then converted them into spectrograms by taking the squared Fourier Transform of the raw waveforms. I sampled at 256 frequencies logarithmically spaced between 100 and 4000 Hz. I monotonically transformed the output with the logarithm function, resulting in the log-power of the sound at specified frequencies over time. I sample following a log distribution of frequencies because humans perceive the pitch of sound following a relatively logarithmic distribution [73].

The data was then divided into segments covering all frequencies and 25 overlapping time points (16 ms each) representing 216 ms total. Subsequently, I performed principal components analysis on the samples to whiten the data as well as reduce the dimensionality. I retained the first 200 principal components as this captured over 93% of the variance in the spectrograms and lowered the simulation time. During analysis, the dictionaries were dewhitened back into spectrogram space.

I also trained with another type of input, cochleograms [69, 70]. These are similar to spectrograms, but the frequency filters mimic known properties of the cochlea via a cochlear model [69].

The cochlear model sampled at 86 frequencies between 73 and 7629 Hz. For this input, the total time for each sample was 250 ms (still 25 time points), and the first 200 principal components captured over 98% of the variance.

### 3.3.1 Whitening

Whitened data,  $\tilde{Y}$ , has a covariance matrix whose expectation is the identity.

$$E[\tilde{Y}\tilde{Y}^T] = I. \quad (3.5)$$

To transform my data into this space, I perform an eigen-decomposition of the covariance of my data,  $Y$ .

$$E[YY^T] = EDE^T. \quad (3.6)$$

where  $E$  is a matrix whose columns are the eigenvectors, and  $D$  is a diagonal matrix whose entries are the ordered eigenvalues of the covariance matrix. In order to whiten my data, I can define a whitening matrix  $W$  as follows:

$$\begin{aligned} W &= D^{-\frac{1}{2}}E^T, \\ \tilde{Y} &= WY. \end{aligned} \quad (3.7)$$

where  $D^{-\frac{1}{2}}$  is a diagonal matrix whose elements are the inverse square roots of the eigenvalues.

Similarly to transform back into the original space, I define a dewhitening matrix  $De$ :

$$De = ED^{\frac{1}{2}}. \quad (3.8)$$

I truncate the whitening and dewhitening matrices to only keep the first two hundred principal components as explained earlier. Therefore, I whiten and project down to 200 principal components in the same step.

## 3.4 Analysis

I now describe the analysis and characterization of my dictionaries.

### 3.4.1 Presentation of Dictionaries

All dictionary units were scaled to be between negative one and one when displayed. The coefficients in the encoding can take on positive or negative values during encoding. To reflect this, I looked at the skewness of each dictionary element. If the skewness was negative, the colors of the dictionary element were inverted when being displayed to reflect the way that element was actually being used.

### 3.4.2 Presentation of Experimental Data

Data from [30] was given to me in raw STRF format. Each was interpolated by a factor of three, but no noise was removed. Data from [31, 55, 60, 53, 62] were given to me in the same format as they were originally published.

### 3.4.3 Usage

I plot all dictionaries in order of their usage. To calculate this for each fixed dictionary, I used LCA to perform inference on the full set of training speech data. Then I ordered the elements by the number of times each was used for all of the data. Note that there are many different encodings one could use depending on how the parameter values are set. I generally kept the parameter values the same as they were during the actual learning.

### 3.4.4 Modulation Power Spectra

To calculate the modulation power spectra, I took a 2D Fourier Transform of each basis function. For each element, I plotted the peak of the temporal and spectral modulation transfer functions. For the cochleogram-trained basis functions, I approximated the cochleogram frequency spacing as being log-spaced to allow comparison with the spectrogram-trained dictionaries.

## 3.5 Linear Filters

To connect with experimental results, I also did linear filter estimations with my model dictionary elements once learning was complete. The most basic model is a standard spike-triggered average. This is the average stimulus before a spike.

Standard Spike-Triggered Average [6]:

$$\hat{\Phi} = \frac{p \cdot a'}{\sigma_p^2} \quad (3.9)$$

$\hat{\Phi}$  is the filter estimate of the receptive field.  $p$  is the probe stimuli.  $a$  are the algorithmic coefficients calculated from the LTC.  $\sigma_p$  is the stimulus variance.

An improved estimation method is the whitened Spike-Triggered Average. This removes the stimulus correlations from the estimate which means the estimate will reflect the underlying receptive field structure rather than structure in the stimulus.

$$\hat{\Phi} = (pp')^{-1}p \cdot a' \quad (3.10)$$

### 3.5.1 Probe Stimulus Sets

I now describe the different types of probe stimulus sets. Abbreviations in parentheses represent how stimuli are referenced in subsequent chapters and figures.

My first set was a holdout group of speech data (HS): This is a set of clips of vowels and phonemes that were not used in the original training. Since this set of speech has the same structure as the training data, it is mathematically guaranteed to give back the original receptive field [6].

Following Smith and Lewicki [17], I also used natural scenes (NS). This set consisted of recordings of animal vocalizations, environmental sounds, auditory scenes, and man-made objects such as blinds, zippers, planes etc.

I also used two types of music since music tends to be very complex and covers the input space fully. The first was a set of classical music (CM); Forty-six tracks from various composers including Bach, Beethoven, Chopin, Grieg, and Rachmaninoff. The second was indie rock music (IRM), one fifty-minute CD "Cloak and Cipher" by Land of Talk.

Three types of white noises were used. Because Gaussian stimuli are uncorrelated, they theoretically should be able to uncover the structure of the model receptive fields. The first type was waveform white noise (WWN), five second samples of mean-zero white noise generated in MATLAB that were attenuated at 20 dB to prevent clipping. Secondly, image white noise (IWN), Gaussian white noise images generated in MATLAB and treated as spectrograms. Finally, I used a stimulus set commonly used in experiments. White noise bursts (WNB) were 100 ms clips of broadband white noise that occurs randomly within a 250 ms window, generated in MATLAB.

My final two stimulus sets are often used in experiments to generate STRFs. Dynamic Moving Ripples (DMR) are a set of ripples in spectrogram space that spanned the space of temporal modulations between -65 and 65 Hz (discrete integers) and spectral modulations between 0 and 4 cycles/octave (continuously sampled). The most standard stimulus set is pure tones (PT), ten second clips of a single integer frequency randomly chosen between 0 and 4000 Hz, generated in MATLAB.

# Chapter 4

## Results/Discussion

I first discuss the types of dictionaries and the effects of the different model parameters. The second part details my comparisons with experimental data. Finally, I examine linear filter estimation.

### 4.1 Cochleogram Dictionaries

In general, training my network on cochleogram representations of speech resulted in smooth and simple shapes for the learned receptive fields of model neurons. Klein et al. [18] used a sparse coding algorithm that imposed an L1-like sparseness constraint to learn a half-complete dictionary trained on speech cochleograms. Their dictionary elements consisted of harmonic stacks at the lower frequencies and localized elements at the higher frequencies. To make contact with these results, I also trained a half-complete L0-sparse dictionary on cochleograms. The resulting dictionary (**Fig. 4.1**) consists of similar shapes to this previous work with the exception of one “onset element” in the upper left (this is the least used of all of the elements from this dictionary). Subsequent simulations revealed that the form of the dictionary is strongly dependent on the degree of overcompleteness. Even a complete dictionary exhibits a greater diversity of shapes than this half-complete dictionary (**Fig. 4.2**). I found this to be true both for L1-sparse dictionaries trained with LCA or with standard Sparsenet (**Figs. 4.6** and **4.10**). Since the representation of sound in the brain appears to be highly overcomplete, with many more neurons present at each subsequent stage of the ascending auditory pathway, the inability of the half-complete dictionary to produce the more complex and diverse receptive field shapes, like those measured in IC, MGBv, or A1, suggests overcompleteness in those regions is crucial to the flexibility of their auditory codes. I explore this further below.

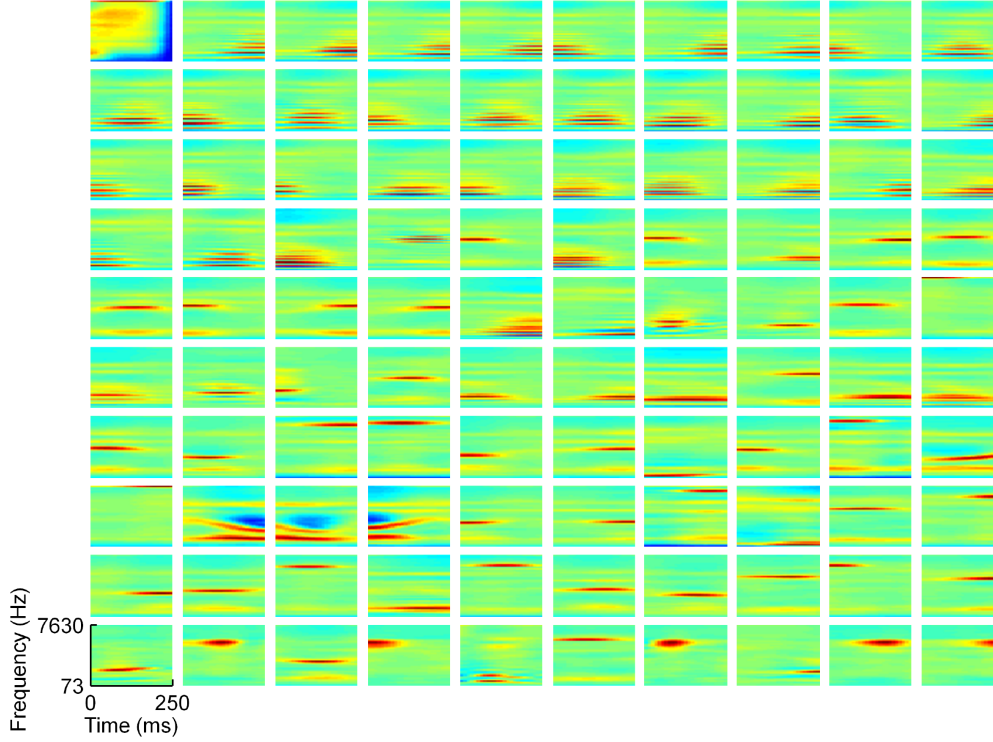


Figure 4.1: A half-complete sparse coding dictionary trained on cochleogram representations of speech exhibits a limited range of shapes. The full set of 100 elements from a half-complete,  $L_0$ -sparse dictionary trained on cochleograms of human speech resemble those found in a previous study [18]. Nearly all elements are extremely smooth, with most consisting of a single frequency subfield or an unmodulated harmonic stack. Each rectangle can be thought of as representing the spectro-temporal receptive field (STRF) of a single element in the dictionary (see Methods for details); time is plotted along the horizontal axis (from 0 to 250 ms), and log frequency is plotted along the vertical axis, with frequencies ranging from 73 Hz to 7630 Hz. Color indicates the amount of power present at each frequency at each moment in time, with warm colors representing high power and cool colors representing low power. Each element has been normalized to have unit Euclidean length. Elements are arranged in order of their usage during inference (*i.e.*, when used to represent individual sounds drawn from the training set) with usage increasing from left to right along each row, and all elements of lower rows used more than those of higher rows.

### 4.1.1 L0-sparse Cochleogram Dictionaries

Analogous to previous sparse coding studies in vision [12, 75], I find that the degree of overcompleteness is a crucial factor for determining the range and complexity of shapes of dictionary elements, whether they are trained on cochleograms or spectrograms. The number of types of STRFs increases when the degree of overcompleteness is increased (**Figs. 4.2, 4.3, 4.4**). For example, with more overcomplete dictionaries, some neurons have subfields spanning all frequencies or the full time-window within the cochleogram. Additionally, I find neurons that exhibit both excitation and suppression in complex patterns, though the detailed shapes differ from the shapes I find for the dictionaries trained on spectrograms. An L0-sparse complete dictionary (**Fig. 4.2**) displays elements similar to those of the half-complete dictionary. The sole onset element is still the least used during inference. This is followed by a series of harmonic stacks; many of which are time-shifted versions of another. There are formants interspersed among the most active elements. Formants are modulations “on top of” the underlying harmonic stack, which often contain pairs of subfields that diverge or converge over time in a manner that is not consistent with a pair of harmonics changing in time due to a temporary rise or fall of the fundamental frequency of the speaker’s voice. A new receptive field shape is highly localized excitation that is sometimes isolated or either preceded/followed by inhibition at the same frequency. This feature is reminiscent of experimental receptive fields. Four dictionary elements that are similar to onsets or offsets at almost all of the frequencies are also highly active elements. They occur at different latencies suggesting that it is easier to encode onsets when there are more basis functions.

A two-times overcomplete dictionary with four hundred elements (**Fig. 4.3**) has the characteristic offset element found in the cochleogram dictionaries followed by the harmonic stacks as the least active dictionary elements. At this level of completeness, it does not seem as if there are additional basis function shapes. Instead, the same shapes from less complete dictionaries are seen at different delays or frequencies.

On the other hand, a four-times overcomplete dictionary (**Fig. 4.4**) has a myriad of more complicated shapes, and this is the first dictionary where the offset is not the least frequently used element. In fact, this dictionary has almost no broadband onset or offset elements. Perhaps, the number of basis functions makes it possible to encode an onset or offset with a few dictionary elements instead of just one. The increased precision in encoding the onset shape may compensate for having to use more neurons for one input. The plot of the usage shows that all of the elements are active for some sounds so the elements are not just encoding the noise, but for the actual signal in the data. Overcompleteness for cochleograms leads to specialization of the dictionary elements.

It is interesting to note that unlike the spectrogram dictionaries, the elements never strongly resemble experimental receptive fields. The cochleogram is not the correct representation to use with sparse coding to uncover the structure of speech. A cochlear model removes many of the redundancies found in the speech and has an implicit infinite precision assumption built-in. Sparse coding exploits the redundancies of the input to encode the data so there might not be enough structure left in the data after going through the cochlear model. In comparison, the spectrogram’s redundancy is more like that of the noise-reduction that finite-precision neural systems use. There-

fore, it is not surprising that the cochleogram model does a poor job of matching experimental receptive fields. I postulate that no cochlear model will do a good job of predicting experimental receptive fields (see Sec. 4.1.4 for further discussion).

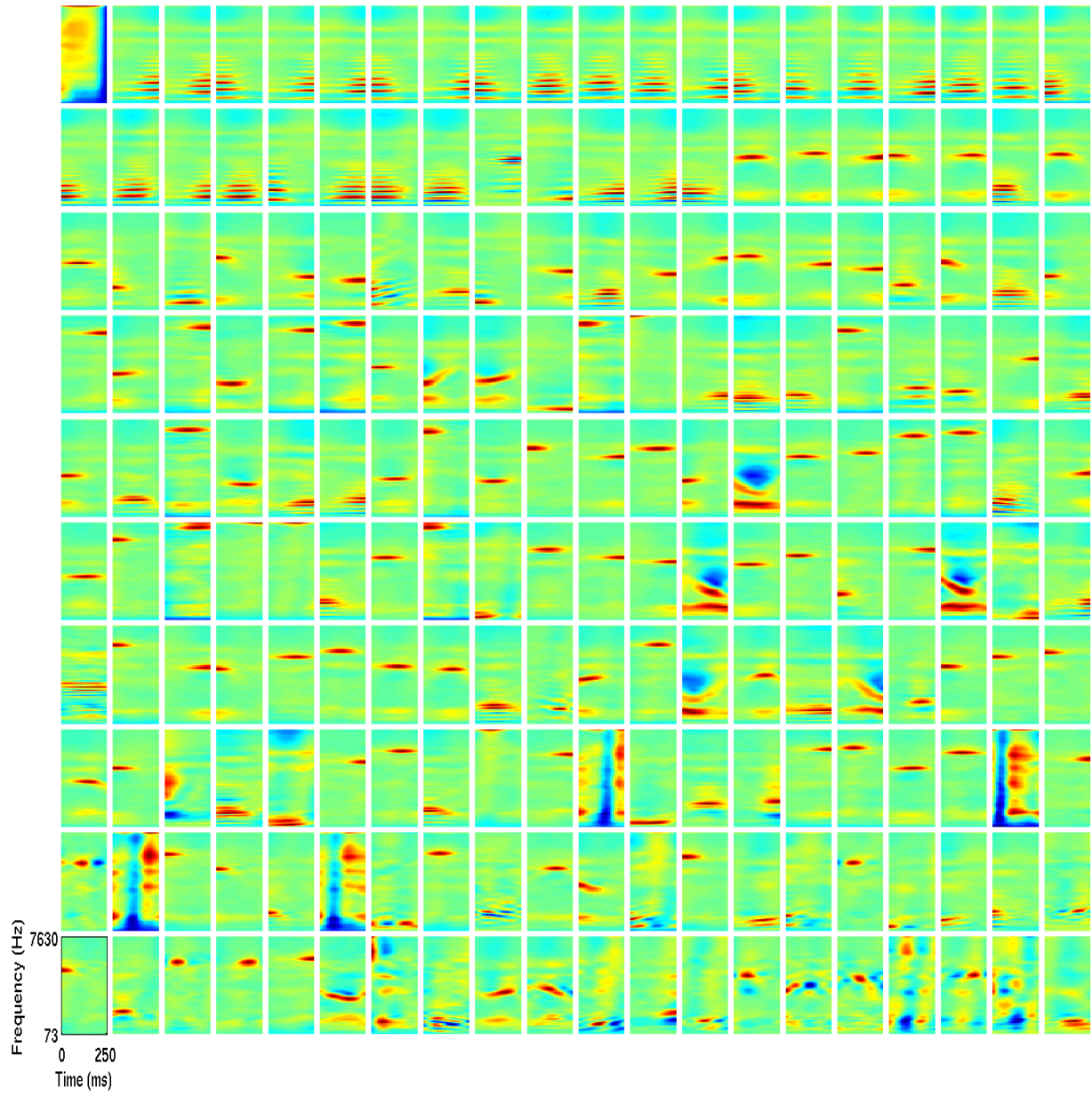


Figure 4.2: The full set of elements from a complete, L0-sparse dictionary trained using LCA [45] on cochleograms of speech. Same conventions as **Fig. 4.1**.



Figure 4.3: The full set of elements from a two-times overcomplete, L0-sparse dictionary trained with LCA [45] on cochleograms of speech. Same conventions as **Fig. 4.1**.

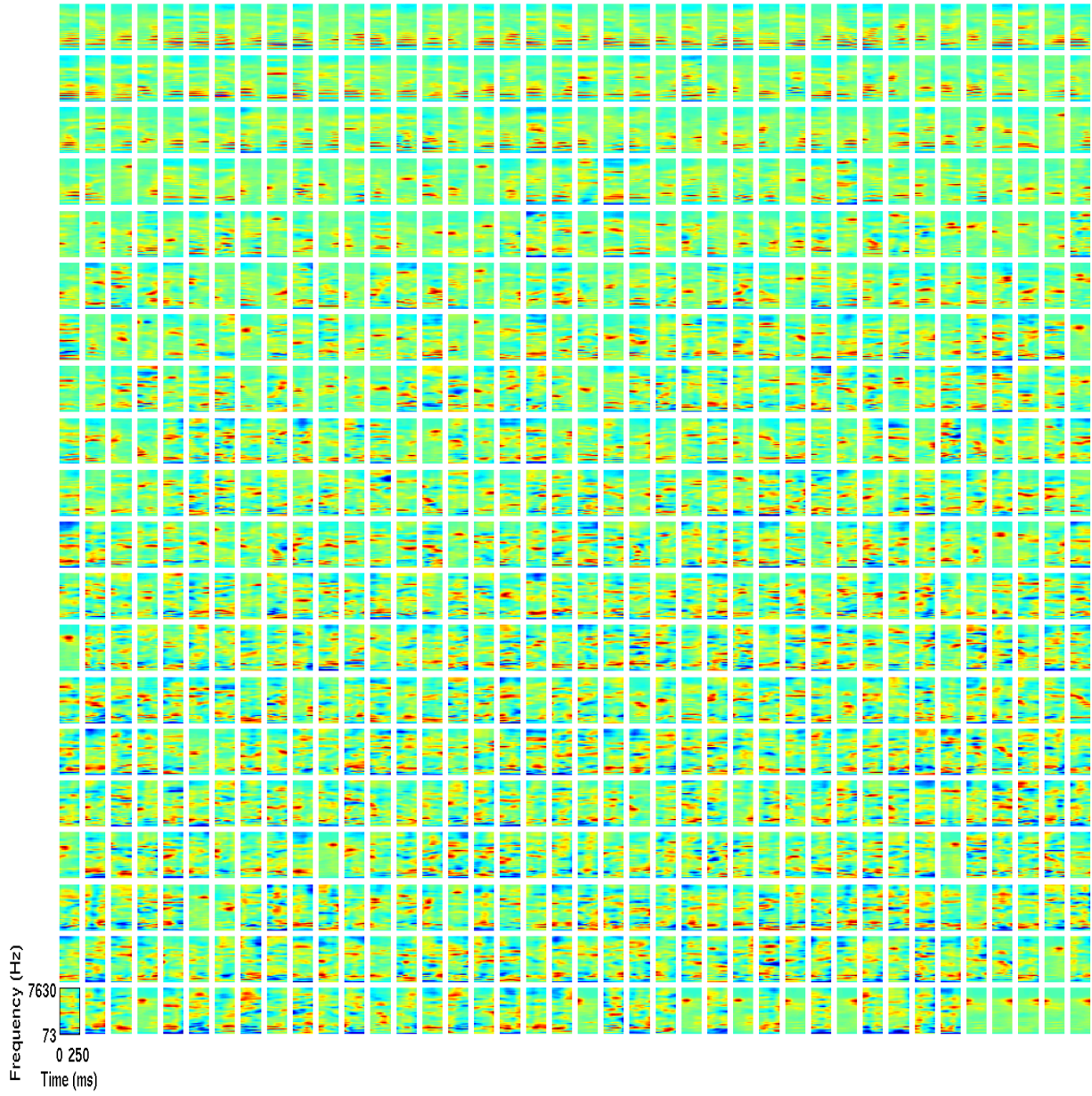


Figure 4.4: The full set of elements from a four-times overcomplete, L0-sparse dictionary trained with LCA [45] on cochleograms of speech. Same conventions as **Fig. 4.1**.

### 4.1.2 L1-sparse Cochleogram Dictionaries

I wondered to what extent the specific form of sparseness I imposed on the representation was affecting the particular features learned by my network. I tested two alternatives to L0-sparseness. The first is L1-sparseness with LCA. A half-complete dictionary (**Fig. 4.5**) looks very similar to the L0-sparse version except the offset element is not present in this dictionary. Thus, this dictionary appears qualitatively the same as the dictionary of Klein et al. [18] with harmonic stacks, localized frequency units, and a few formants. This group did not display their entire dictionary so it is not possible to compare the actual numbers of cell-types. This is not surprising since both models use the same type of sparseness.

A complete dictionary (**Fig. 4.6**) has the offset element and the “standard cochleogram” receptive field shapes. The more specialized and less smooth shapes are the most active units. There appear to be units with FM direction selectivity, a feature found in experimental STRFs although the individual shapes are not a perfect match.

The two-times overcomplete dictionary (**Fig. 4.7**) has two offset elements that are least active. This dictionary has the same shapes as less complete dictionaries but with different distributions. This dictionary also demonstrates one of the problems sometimes found in training. If the learning rate is too high, then some elements appeared twice in the dictionary, i.e. the elements were identical to MATLAB precision levels. Having a smaller learning rate and more iterations will negate this effect, but this dictionary took many simulations to train. This demonstrates that the dictionaries can sometimes be very sensitive to learning rate values.

The four-times overcomplete dictionary (**Fig. 4.8**) illustrates how the parameters can effect the encoding. The usage plots demonstrates that harmonic stacks are the most used elements and the complex shapes are the least used, unlike the other dictionaries. This can occur when the inference parameters are set to different values. Rather than change the parameters to look more like the other dictionaries, I left this dictionary as it was to illustrate this effect. Also, the cochleogram trained dictionaries did not match experimental data so less time was spent analyzing these dictionaries.

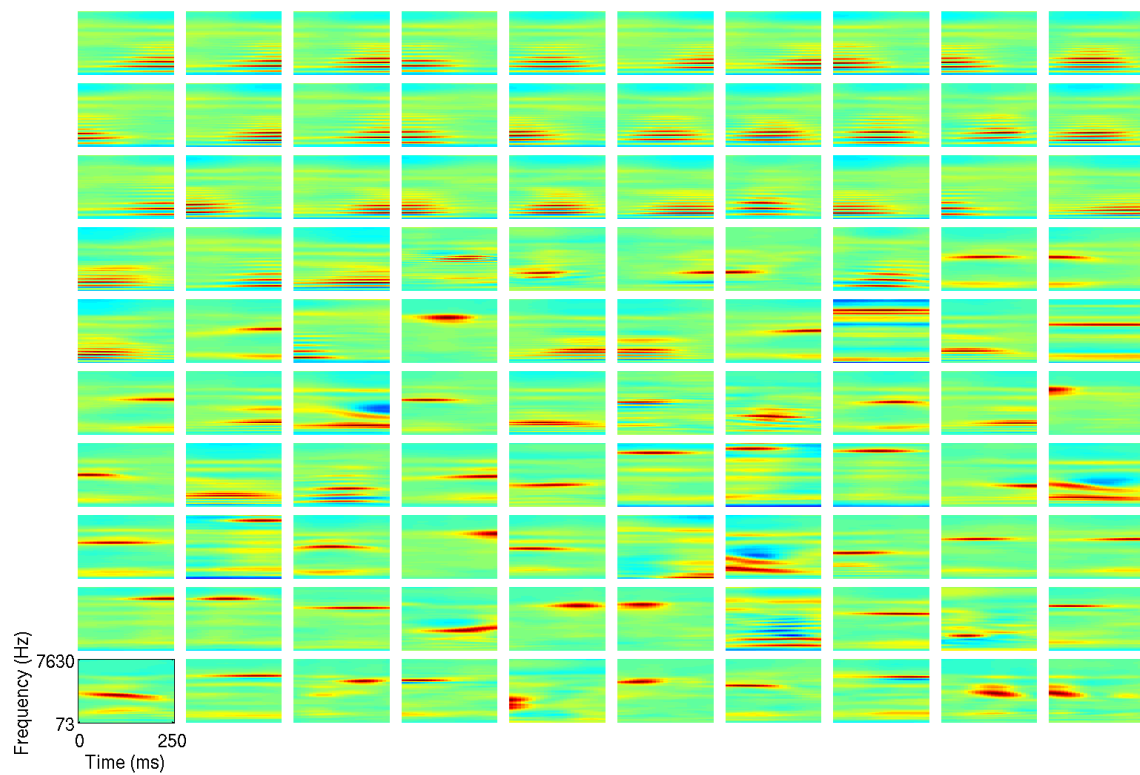


Figure 4.5: The full set of elements from a half-complete, L1-sparse dictionary trained with LCA [45] on cochleograms of speech. Same conventions as **Fig. 4.1**.

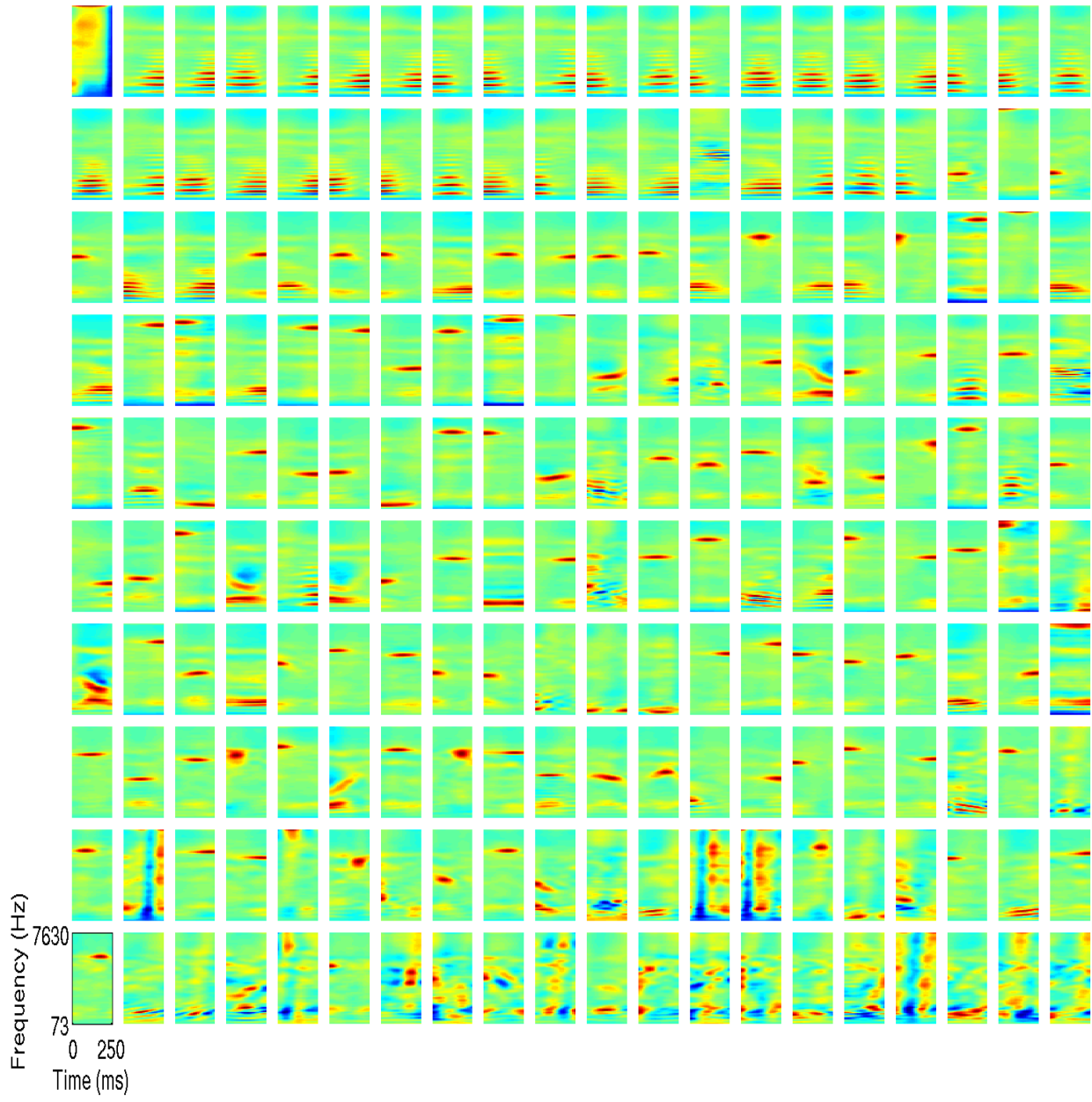


Figure 4.6: The full set of elements from a complete, L1-sparse dictionary trained with LCA [45] on cochleograms of speech. Same conventions as **Fig. 4.1**.

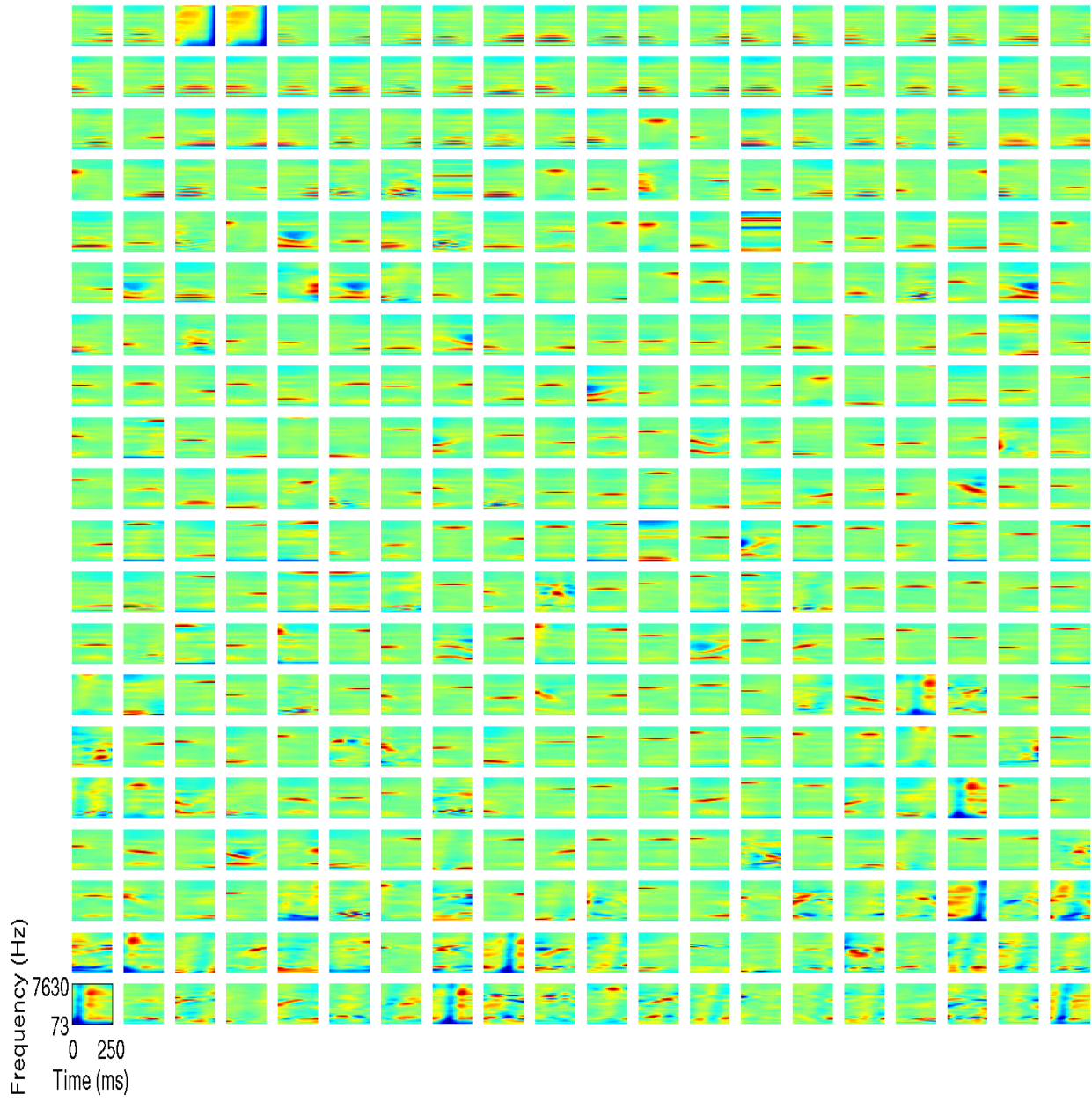


Figure 4.7: The full set of elements from a two-times overcomplete, L1-sparse dictionary trained with LCA [45] on cochleograms of speech. Same conventions as **Fig. 4.1**.

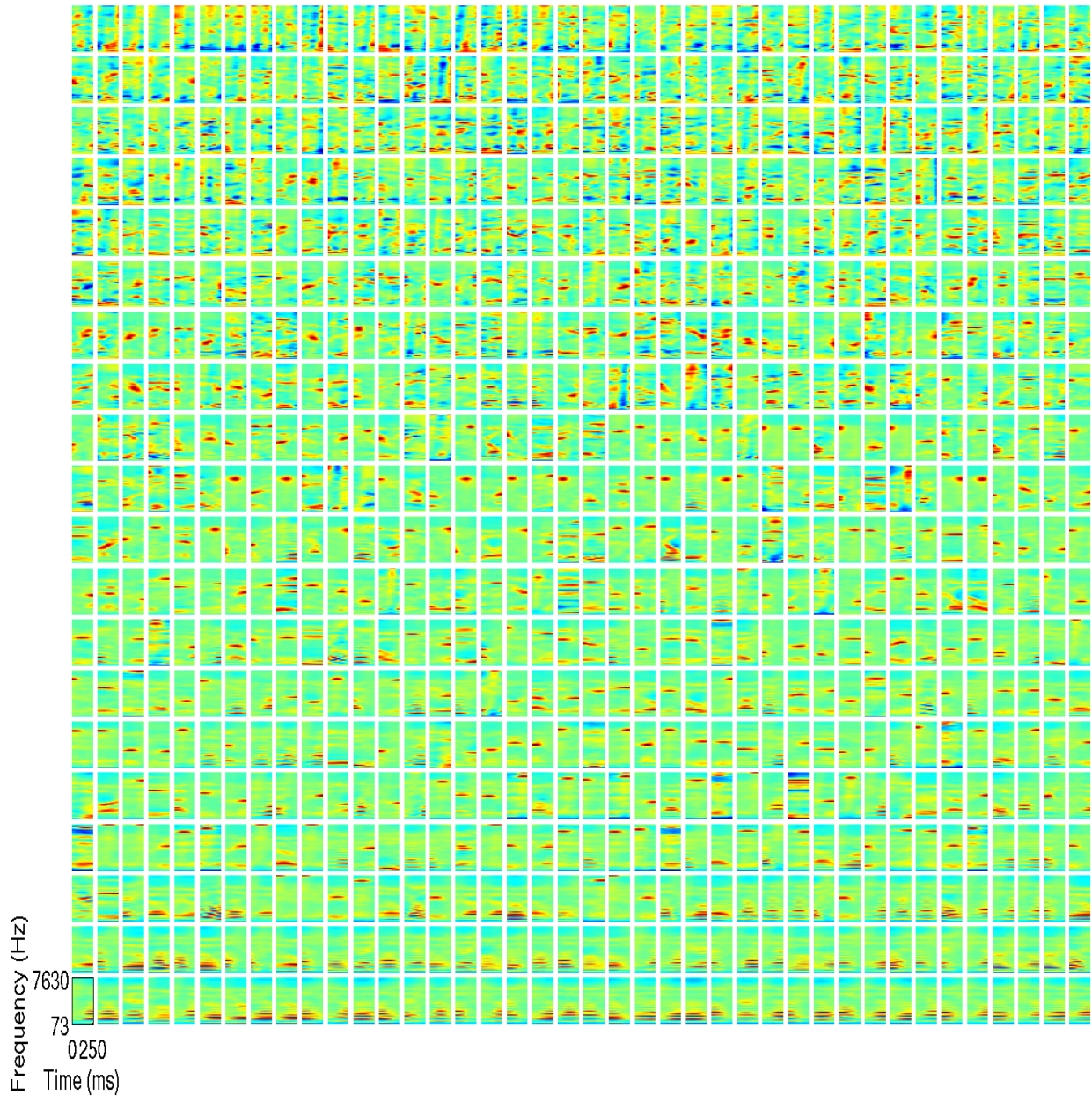


Figure 4.8: The full set of elements from a four-times overcomplete, L1-sparse dictionary trained with LCA [45] on cochleograms of speech. Same conventions as **Fig. 4.1**.

### 4.1.3 Sparsenet-trained Cochleogram Dictionaries

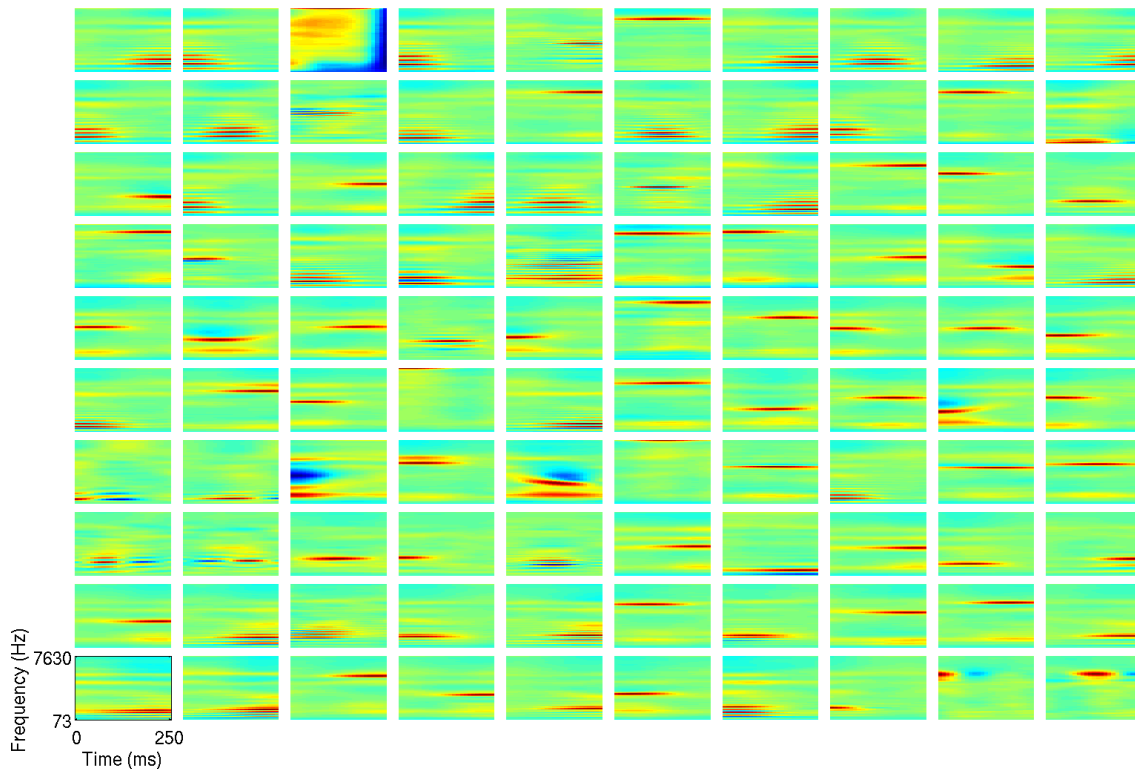


Figure 4.9: The full set of elements from a half-complete, L1-sparse dictionary trained with Sparsenet [11] on cochleograms of speech. Same conventions as **Fig. 4.1**.

The dictionaries trained with Sparsenet (**Figs. 4.9** and **4.10**) share the same characteristics as those dictionaries trained with LCA. The different cell-types are not as well separated. This is probably because Sparsenet uses an approximation to the L1-norm. This control shows that LCA is working as well as Sparsenet. Additionally, it is exciting that this type of training matches closely with the dictionary of Klein et al. [18].

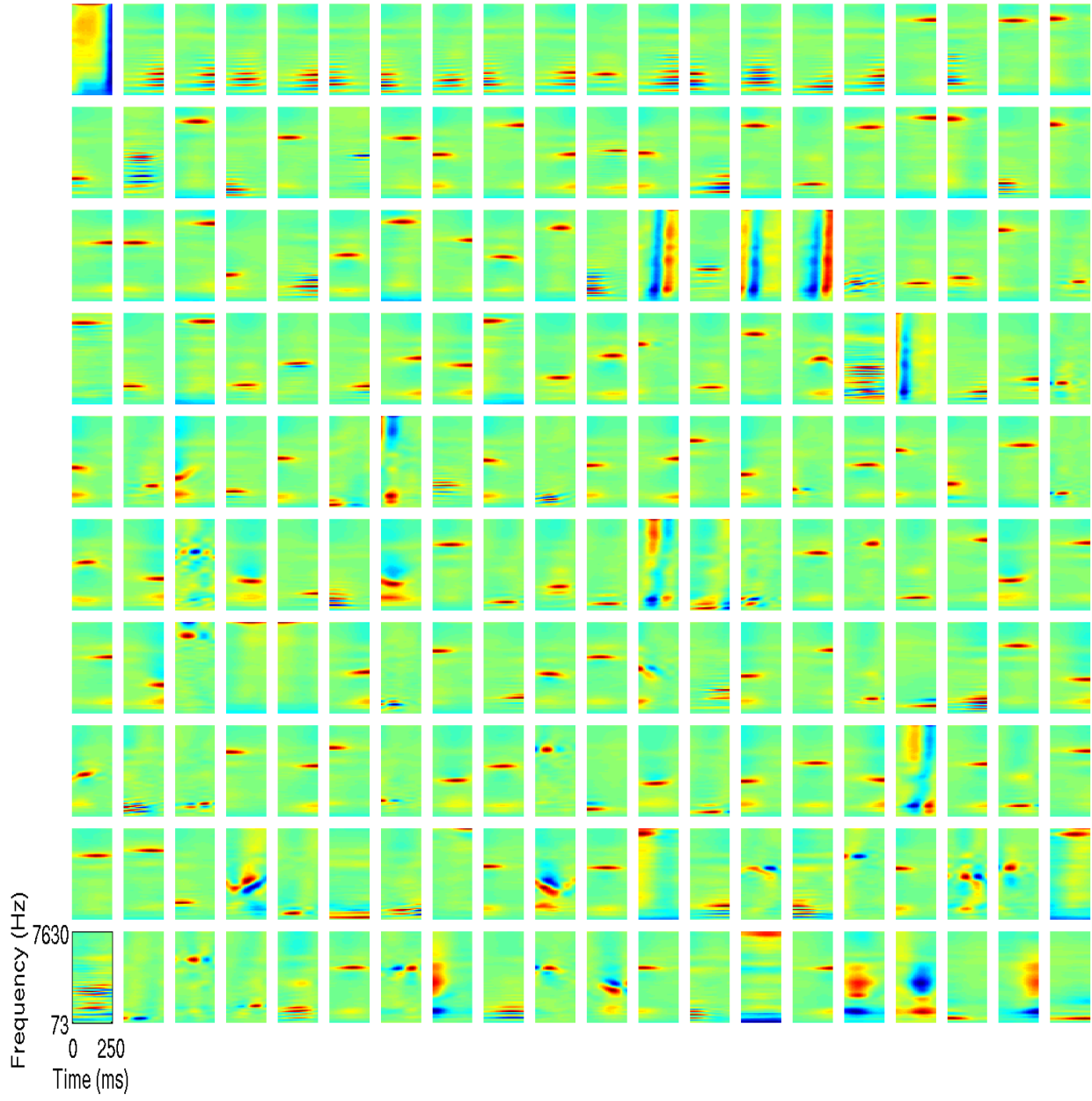


Figure 4.10: The full set of elements from a complete, L1-sparse dictionary trained with Sparsenet [11] on cochleograms of speech. Same conventions as **Fig. 4.1**.

#### 4.1.4 Cochleogram timescale

A common criticism of the cochleogram dictionaries is that the timescale has been decimated by such a large number. Thus, it would not be possible for these dictionaries to uncover results similar to the spectrogram dictionaries or experimental data. To test this, I ran simulations of L0-sparse dictionaries that were four times as fine as my original set of cochleogram dictionaries. In the original set, twenty-five time points represent 250 ms. In my new simulations, fifty time points represent 125 ms. I exhibit both a two-times overcomplete dictionary (**Fig. 4.11**) and a four-times overcomplete dictionary (**Fig. 4.12**). The figures demonstrate that even these dictionaries do not exhibit shapes that resemble experimental STRFs. Instead, the STRF shapes resemble those found in the coarser cochleogram dictionaries, namely harmonic stacks at the lower frequencies, localized units at higher frequencies, some formants, and some “on/off” elements. The distribution of these cell-types is different from the original dictionaries, but the same was also true for L0-sparse vs. L1-sparse dictionaries. My conclusion from these simulations is that sparse coding on cochleograms will never be able to replicate experimental STRFs no matter how fine the timescale.

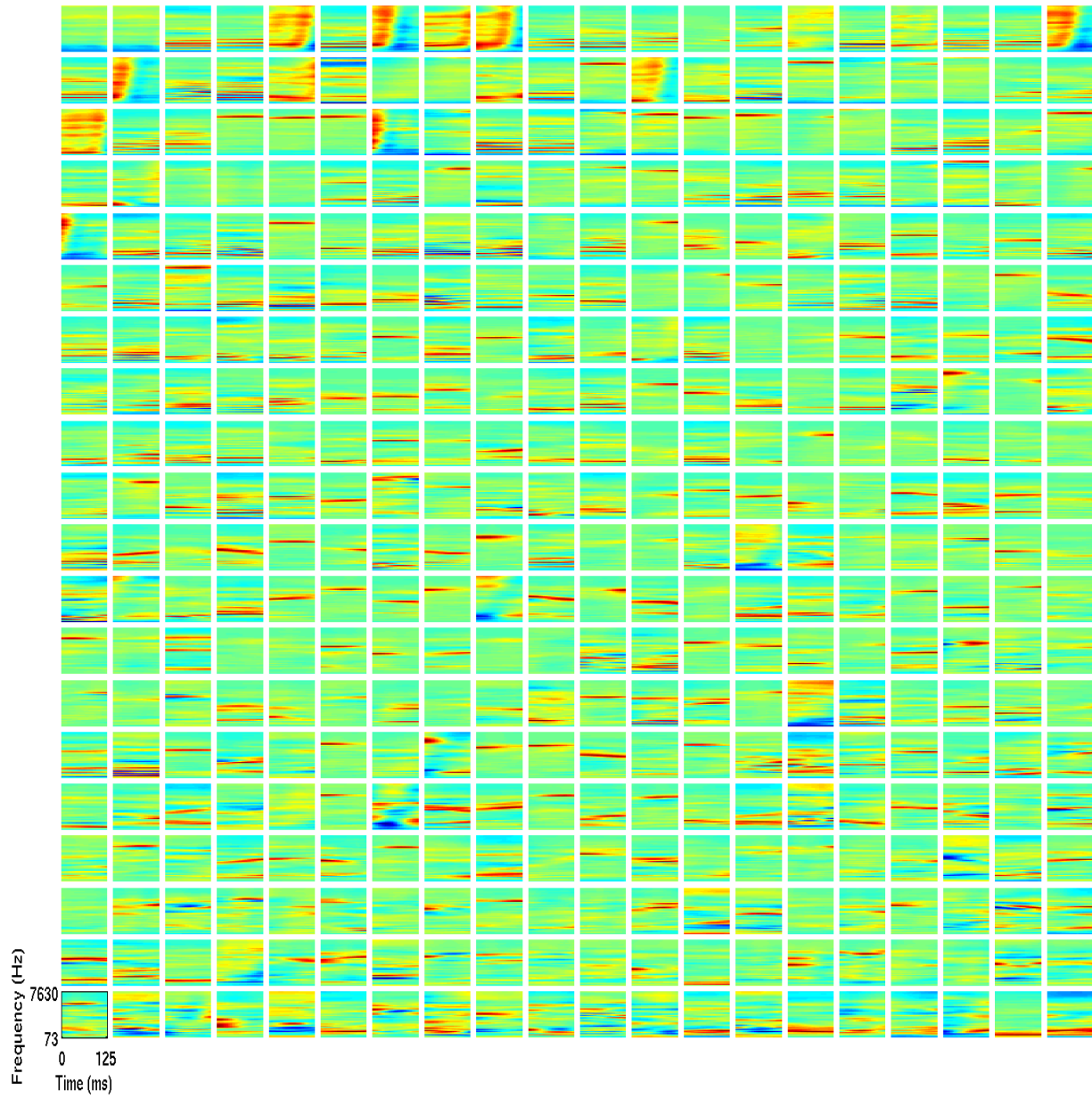


Figure 4.11: A two-times overcomplete L0-sparse coding dictionary trained on cochleogram representations of speech exhibits a limited range of shapes even with a finer timescale. Here, fifty time points represent 125 ms.

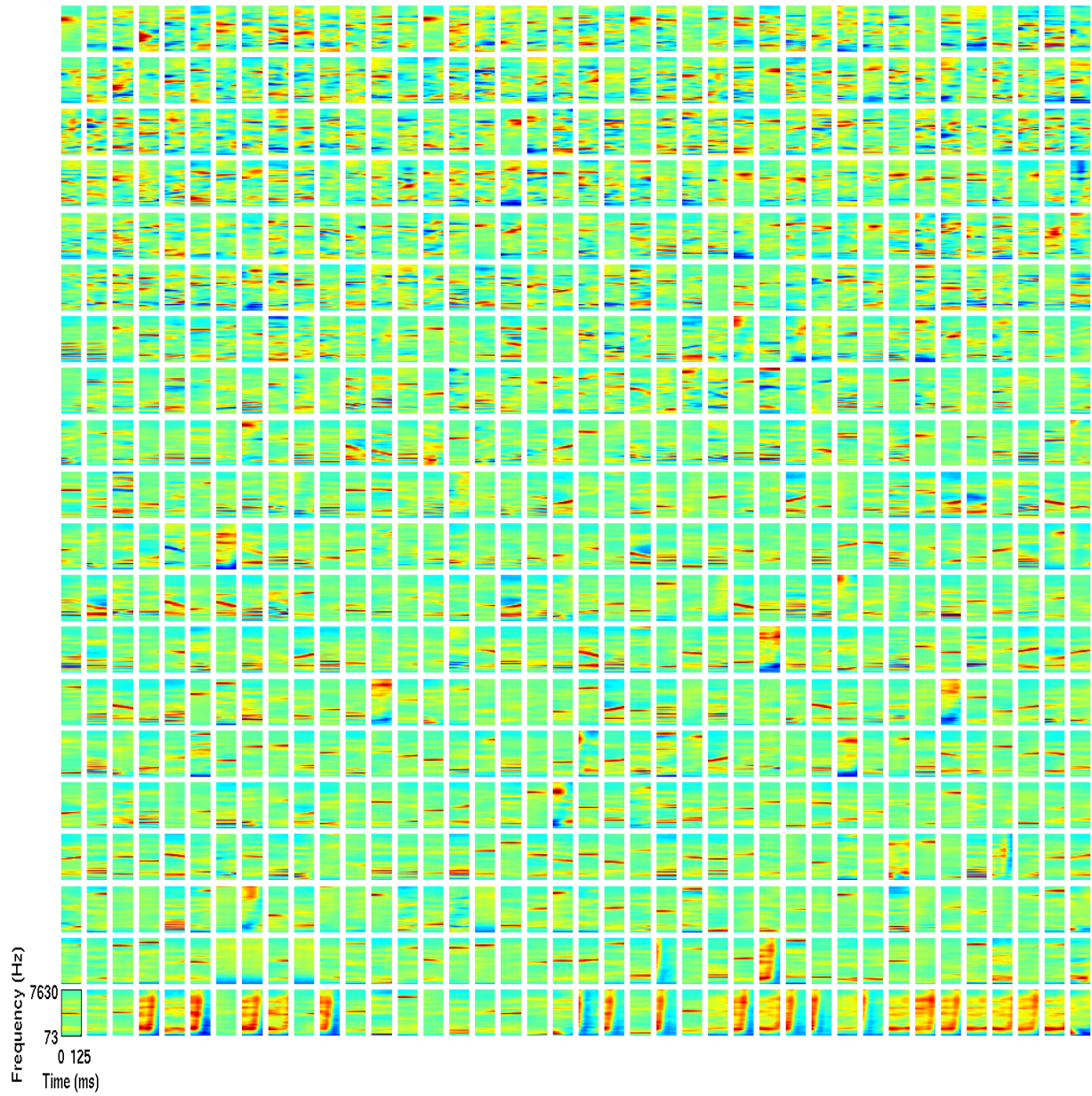


Figure 4.12: A four-times overcomplete L0-sparse coding dictionary trained on cochleogram representations of speech exhibits a limited range of shapes even with a finer timescale. Here, fifty time points represent 125 ms.

## 4.2 Spectrogram Dictionaries

### 4.2.1 L0-sparse Spectrogram Dictionaries

The spectrogram-trained dictionaries provide a richer and more diverse set of dictionary elements than those trained on cochleograms. I display representative elements of the different categories of shapes found in a half-complete L0-sparse spectrogram dictionary (**Fig. 4.13a-f**, the full dictionary is shown in **Fig. 4.14**) along with a plot of the usage of the elements during inference (**Fig. 4.13g**). Interestingly, I find that the different qualitative types of neurons separate according to their usage into a series of rises and plateaus. The least used elements are the harmonic stacks (**Fig. 4.13a**), which is unsurprising since only one of them needs to be active at any given time for a typical epoch of a recording from a single human speaker. Harmonic stacks are the least used for both the cochleogram and spectrogram dictionaries. The neighboring flat region consists of onset elements (**Fig. 4.13b**), which contain broad frequency subfields that change abruptly at one moment in time. These neurons were all used approximately equally often across the training set since it is equally probable that a stimulus transient will occur any time during the 216 ms time window.

The third region consists of more complex harmonic stacks that contain low power subfields on the sides (**Fig. 4.13c**), a feature sometimes referred to as “temporal inhibition” when observed in neural receptive fields; I will refer to this as “suppression” rather than inhibition to indicate that the model is agnostic as to whether these suppressed regions reflect actual direct synaptic inhibition to the recorded neuron, rather than a decrease in excitatory synaptic input, for example. The next flat region represents stimulus onsets, or ON-type cells, that tend to be more localized in frequency (**Fig. 4.13d**). The fifth group of elements is reminiscent of formants (**Fig. 4.13e**), which are resonances of the vocal tract that appear as characteristic frequency modulations common in speech. The final region consists of the most active neurons, which are highly localized in time and frequency and exhibit tight checkerboard-like patterns of excitatory and suppressive subfields, sometimes including diagonally oriented (time-frequency inseparable) subfields (**Fig. 4.13f**). These features are exciting because they are similar to experimentally measured receptive field shapes that to my knowledge have not previously been theoretically predicted, as discussed below.

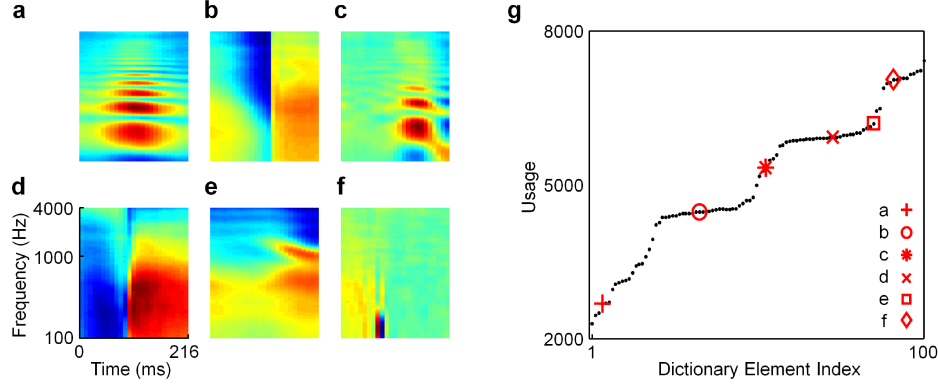


Figure 4.13: A half-complete, L0-sparse dictionary trained on spectrograms of speech exhibits a variety of distinct shapes that capture several classes of acoustic features present in speech and other natural sounds. (a-f) Selected elements from the dictionary that are representative of different types of receptive fields: (a) a harmonic stack; (b) an onset element; (c) a harmonic stack with flanking suppression; (d) a more localized onset/termination element; (e) a formant; (f) a tight checkerboard pattern (see Fig. 4.14 for the full dictionary). Each rectangle represents the spectro-temporal receptive field (STRF) of a single element in the dictionary; time is plotted along the horizontal axis (from 0 to 216 msec) and log frequency is plotted along the vertical axis, with frequencies ranging from 100 Hz to 4000 Hz. (g) A graph of the usage of the dictionary elements showing that the different types of receptive field shapes separate based on usage into a series of rises and plateaus; red symbols indicate where each of the examples from panels a-f fall on the graph. The vertical axis represents the number of stimuli that required a given dictionary element in order to be represented accurately during inference.

The complete dictionary (Fig. 4.15) possesses many of the same receptive field cell-types as the half-complete dictionary. The checkerboard shapes that are characteristic of IC begin appearing as the most active elements. While the shapes become more complicated and localized in this dictionary, they still do not resemble the cochleogram dictionaries. As mentioned earlier, the spectrogram representation retains more correlations (redundancy) than the cochleogram dictionary. This might explain why the two representations produce such different results. The majority of the FM direction selectivity receptive fields are selective for downward sweeps which matches experimental work [55] as well as the type of sweeps that are in speech.

The two-times overcomplete dictionary (Fig. 4.16) displays the same cell-types as less complete dictionaries. The checkerboard units extend for more cycles in time, which is a trend as overcompleteness increases. Such extended checkerboard shapes have not yet been found in real cells.

I show representative examples of essentially all distinct cell types found in a four-times overcomplete L0-sparse dictionary trained on spectrograms (Fig. 4.17, the full dictionary is shown in Fig. 4.18). With increased levels of overcompleteness, the learned features become more complex, exhibiting richer patterns of excitatory and suppressive subfields. Features in the half-complete dictionary do appear as subsets of the larger dictionaries (Fig. 4.17a, c, e, g, i), but the more complex and biologically interesting features are not unique to overcomplete training.

Novel features that were not observed in smaller dictionaries include: an excitatory harmonic stack flanked by a suppressive harmonic stack (Fig. 4.17b); a neuron excited by low frequencies

(**Fig. 4.17d**); a neuron sensitive to two middle frequencies (**Fig. 4.17f**); a localized but complex excitatory subregion followed by a suppressive subregion that is strongest for high frequencies (**Fig. 4.17h**); a checkerboard pattern with roughly eight distinct subregions (**Fig. 4.17i**); a highly temporally localized OFF-type neuron (**Fig. 4.17j**); and a broadband checkerboard pattern that extends for many cycles in time (**Fig. 4.17k**).

As in the case of the half-complete dictionary (**Fig. 4.13**), the different classes of receptive field shapes segregate as a function of usage even as more intermediary shapes appear (see **Fig. 4.18** for the full four-times overcomplete dictionary). However, the plateaus and rises seen for the usage plot for the half-complete dictionary (**Fig. 4.13g**) are no longer distinct for this four-times overcomplete representation (**Fig. 4.17m**).

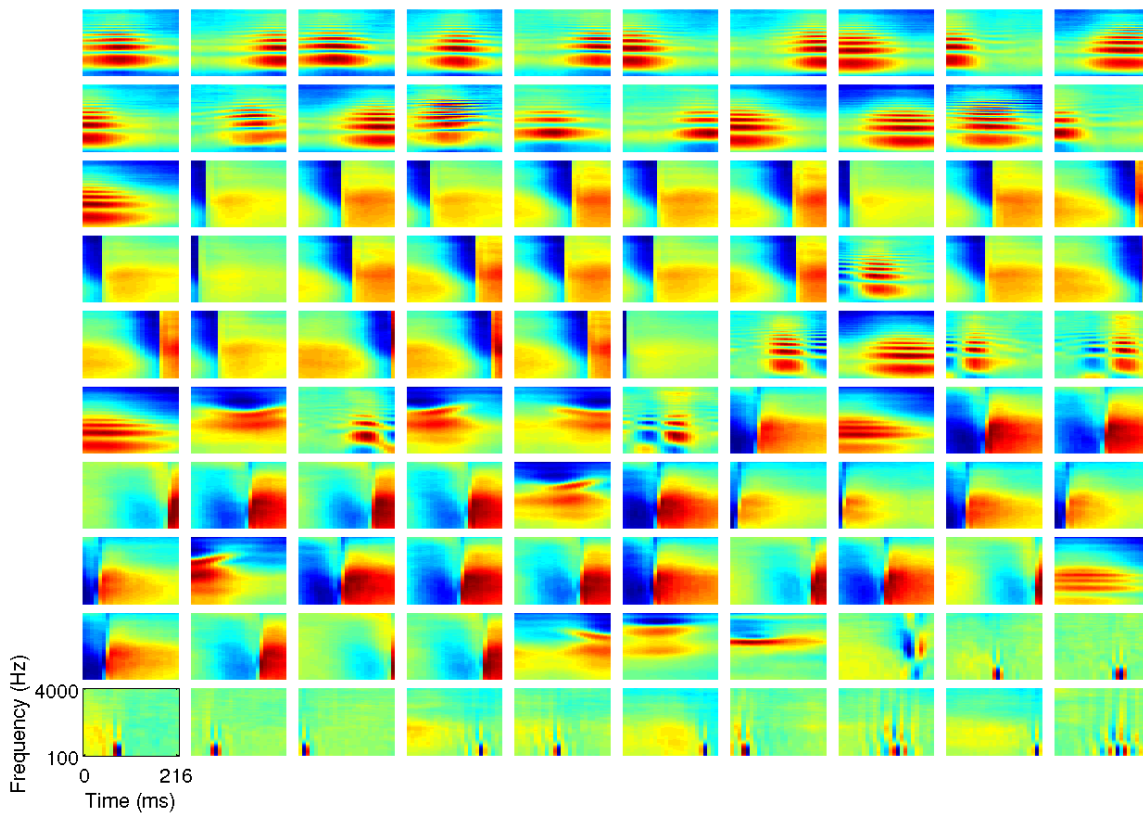


Figure 4.14: The full set of elements from a half-complete, L0-sparse dictionary trained with LCA [45] on spectrograms of speech. Each rectangle represents the spectrotemporal receptive field of a single element in the dictionary; time is plotted along the horizontal axis (from 0 to 216 msec), and log frequency is plotted along the vertical axis, with frequencies ranging from 100 Hz to 4000 Hz. Color indicates the amount of power present at each frequency at each moment in time, with warm colors representing high power and cool colors representing low power. Each element has been normalized to have unit Euclidean length. Elements are arranged in order of their usage during inference with usage increasing from left to right along each row, and all elements of lower rows used more than those of higher rows.

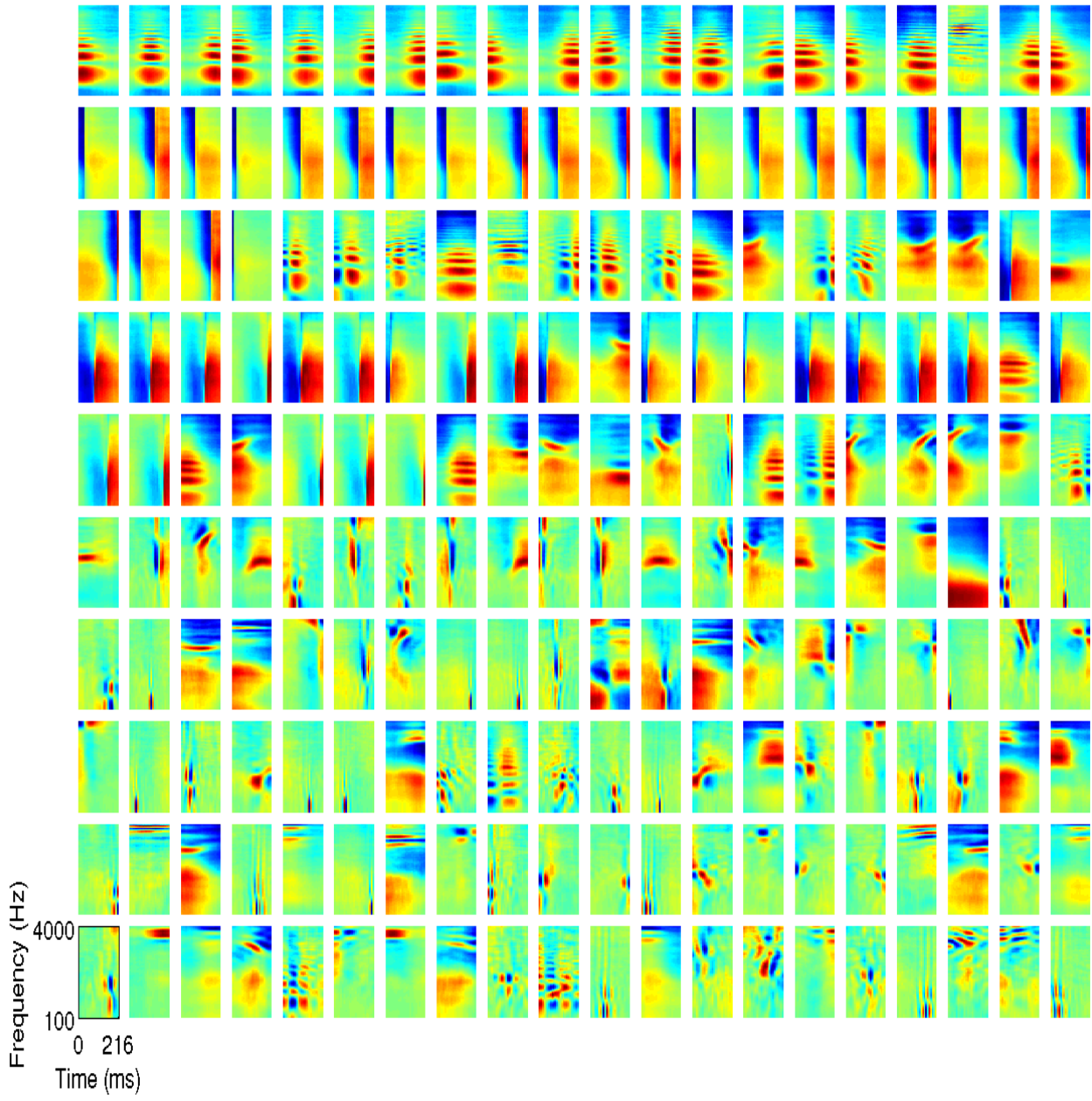


Figure 4.15: The full set of elements from a complete, L0-sparse dictionary trained with LCA [45] on spectrograms of speech. Same conventions as **Fig. 4.14**.

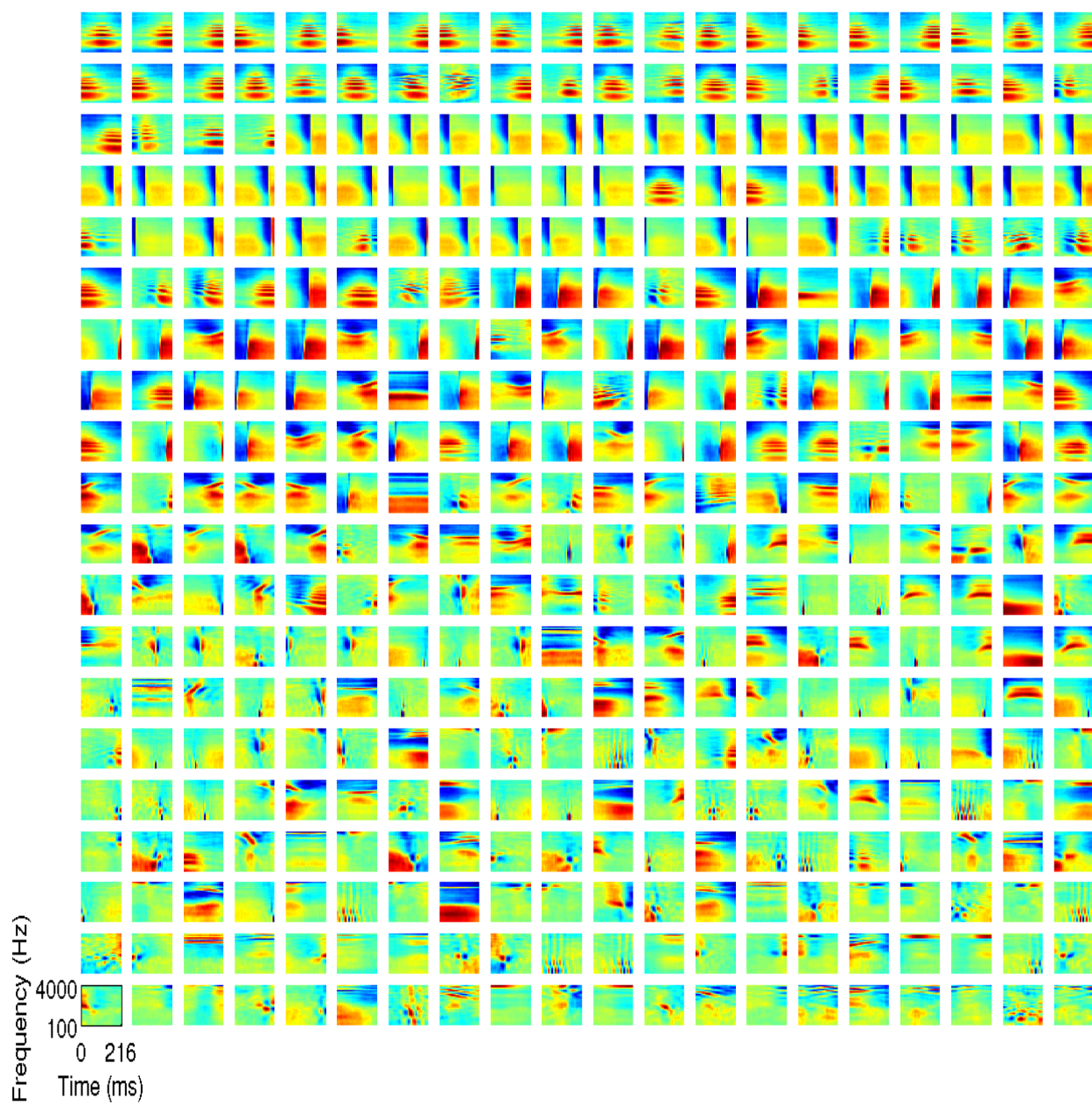


Figure 4.16: The full set of elements from a two-times overcomplete, L0-sparse dictionary trained with LCA [45] on spectrograms of speech. Same conventions as **Fig. 4.14**.

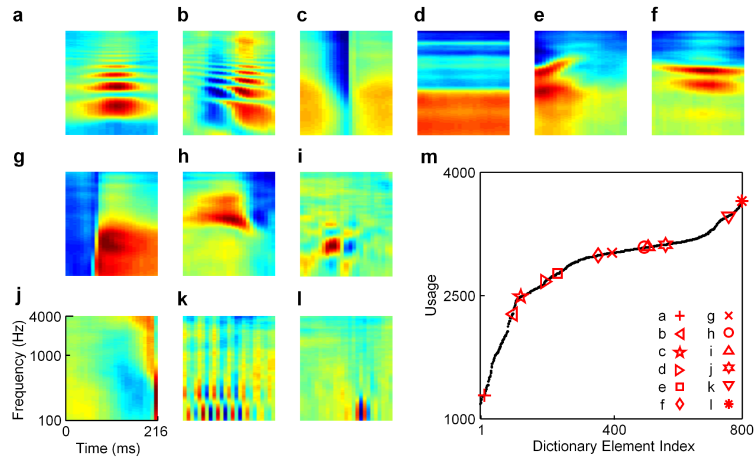


Figure 4.17: Representative elements from the four-times overcomplete L0-sparse spectrogram-trained dictionary. **a**, **c**, **e**, **g**, **j**, and **l** resemble those of the half-complete dictionary (see **Fig. 4.13**). Other neurons display more complex shapes than those found in less overcomplete dictionaries: **(b)** a harmonic stack with larger flanking suppressive subregions; **(d)** a neuron sensitive to lower frequencies; **(f)** a short harmonic stack; **(h)** a localized but complex pattern of excitation with flanking suppression; **(i)** a localized checkerboard with larger excitatory and suppressive subregions than those in panel **l**; **(k)** a checkerboard pattern that extends for many cycles in time. Several of these patterns resemble neural spectro-temporal receptive fields (STRFs) reported in various stages of the auditory pathway that have not been predicted by previous theoretical models (see text and **Figs. 4.30**, **4.31**, and **4.32**). **(m)** A graph of usage of the dictionary elements during inference. The different classes of dictionary elements still separate according to usage (see **Fig. 4.18** for the full dictionary) although the notable rises and plateaus as seen in **Fig. 4.13g** are less apparent in this larger dictionary.

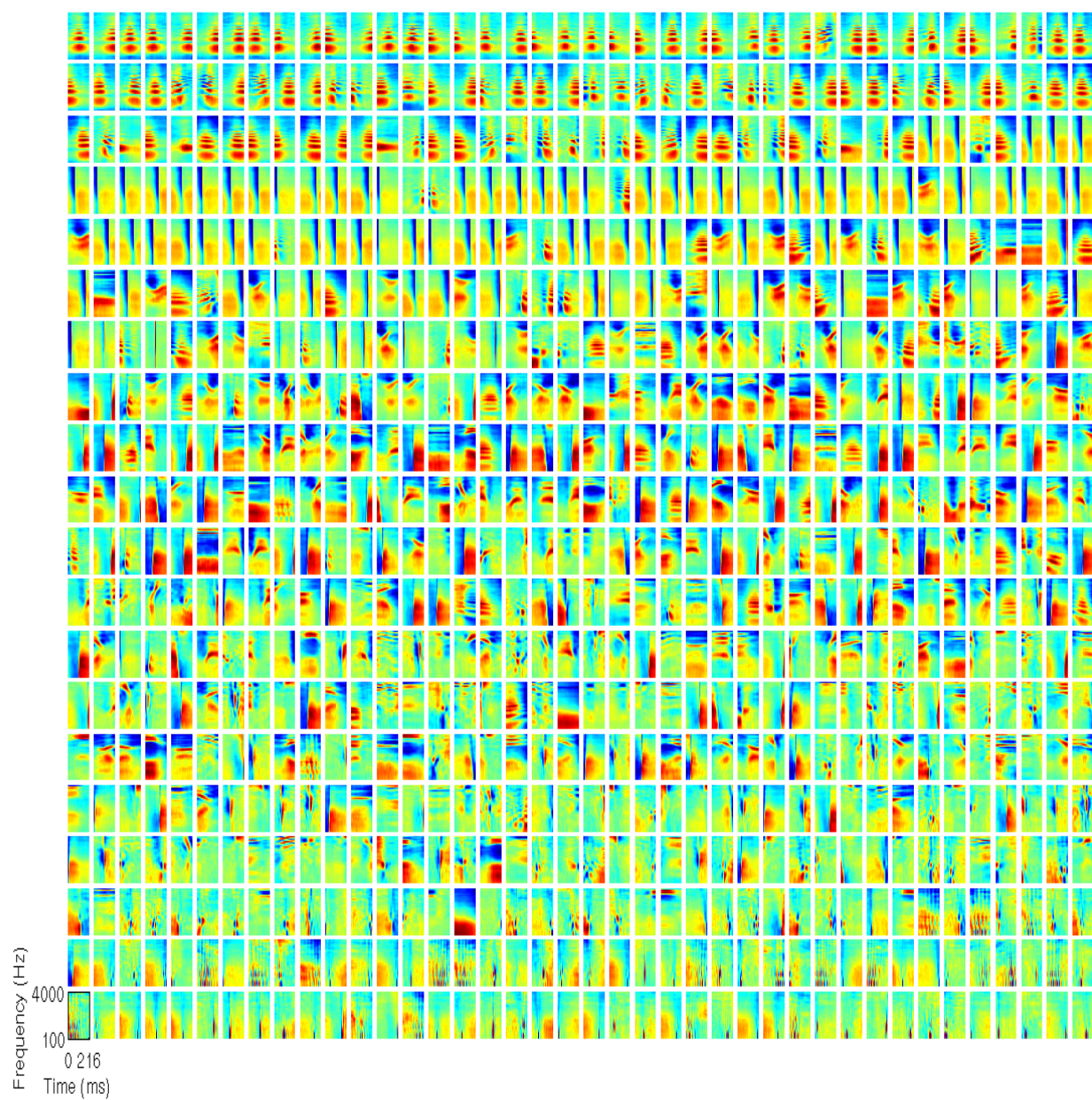


Figure 4.18: The full set of elements from a four-times overcomplete, L0-sparse dictionary trained with LCA [45] on spectrograms of speech. Same conventions as **Fig. 4.14**.

### 4.2.2 L1-sparse Spectrogram Dictionaries

Like the cochleogram representation, I tested the form of sparseness. I used the LCA algorithm [45] to find the soft sparse solution (*i.e.*, one that minimizes the L1 norm), and obtained similar results to what I found for the hard sparse cases, with increasing overcompleteness resulting in greater diversity and complexity of learned features (see **Figs. 4.19-4.22**).

The L1-sparse half-complete spectrogram trained dictionary (**Fig. 4.19**) is qualitatively similar to the L0-sparse dictionary (**Fig. 4.14**) in terms of the cell-types. The distributions are different, e.g. there are three checkerboard elements instead of just one. However, no new shapes appeared, only the relative proportion of each cell-type changed.

Similarly, the complete dictionary (**Fig. 4.20**) has more checkerboard shapes than its L0-sparse counterpart. The checkerboard shapes also last for more cycles in time than elements from the half-complete dictionary as well as the complete L0-sparse dictionary.

The two-times overcomplete dictionary (**Fig. 4.21**) and four-times overcomplete dictionary (**Fig. 4.22**) followed the same trends as their L0-sparse counterparts; more complex shapes appear.

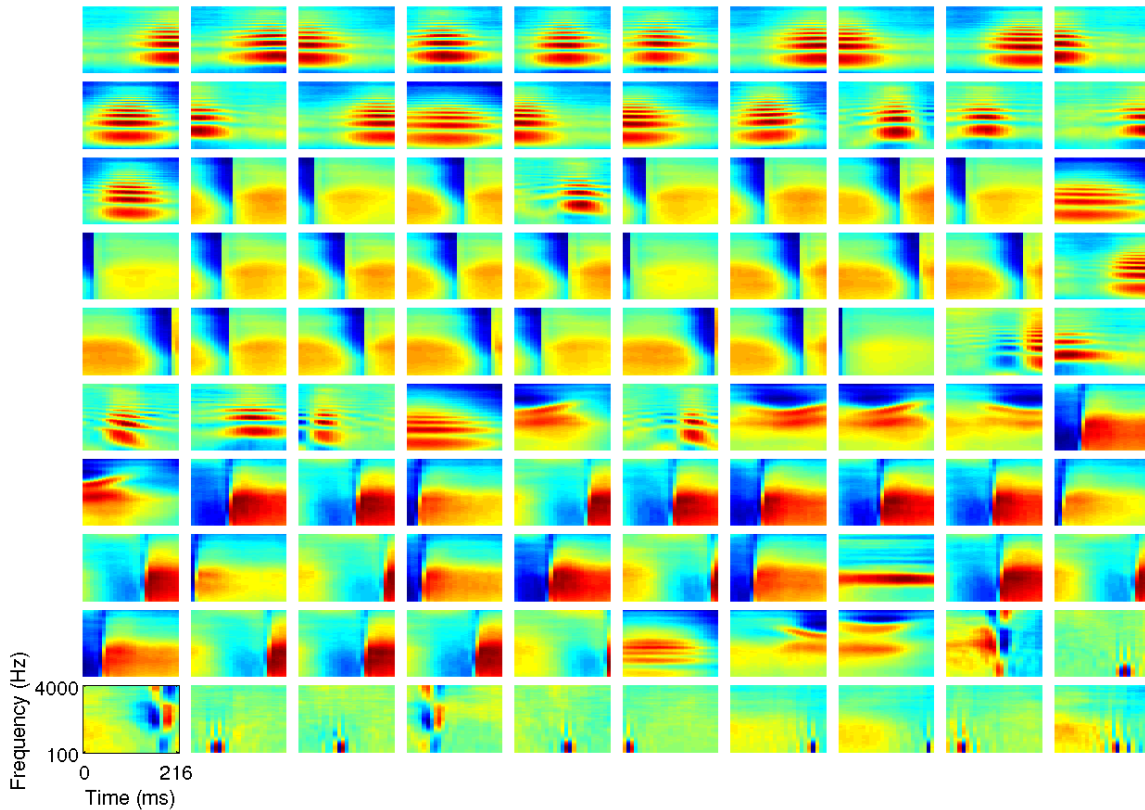


Figure 4.19: The full set of elements from a half-complete, L1-sparse dictionary trained with LCA [45] on spectrograms of speech. Same conventions as **Fig. 4.14**.

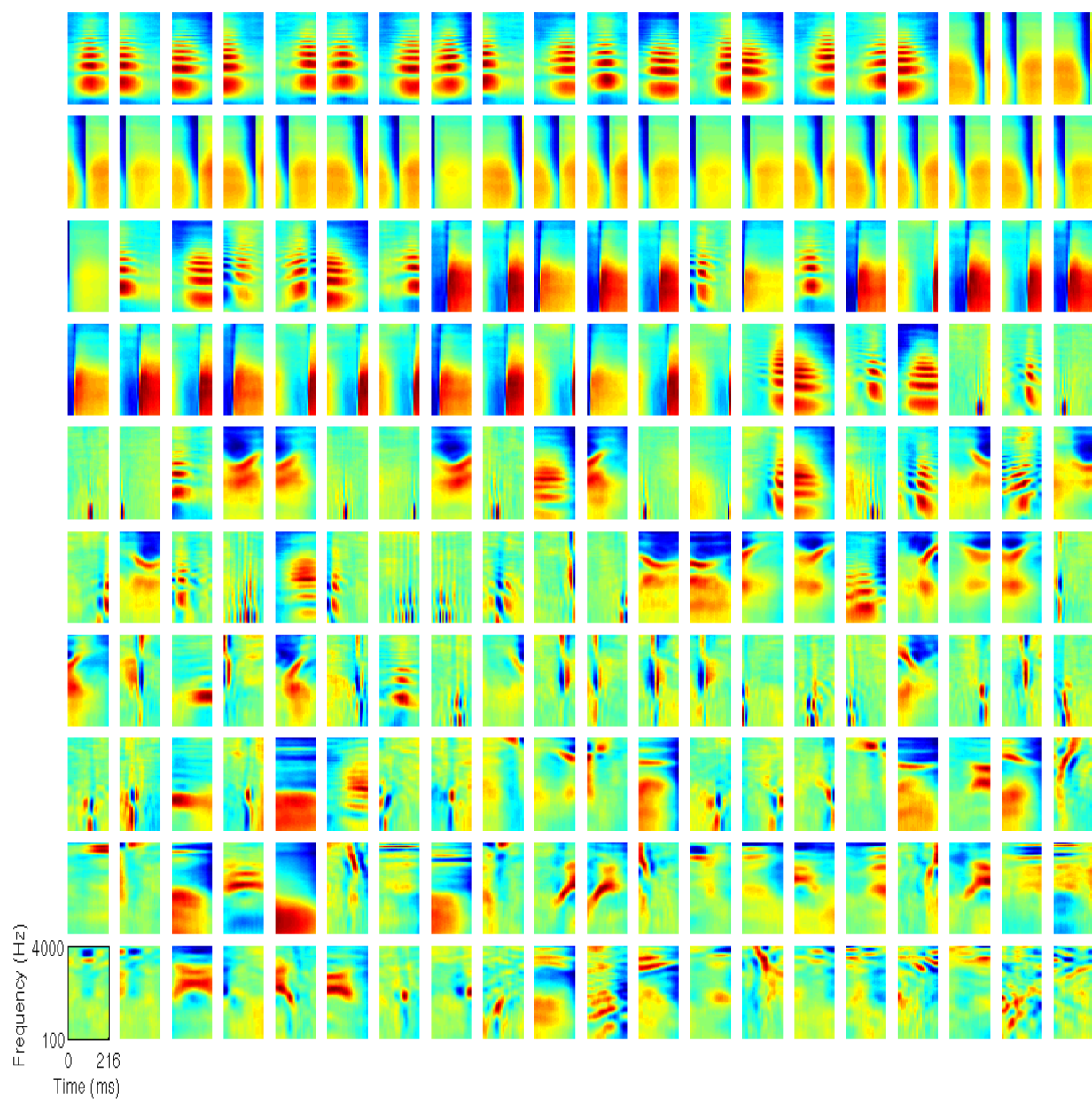


Figure 4.20: The full set of elements from a complete, L1-sparse dictionary trained with LCA [45] on spectrograms of speech. Same conventions as **Fig. 4.14**.

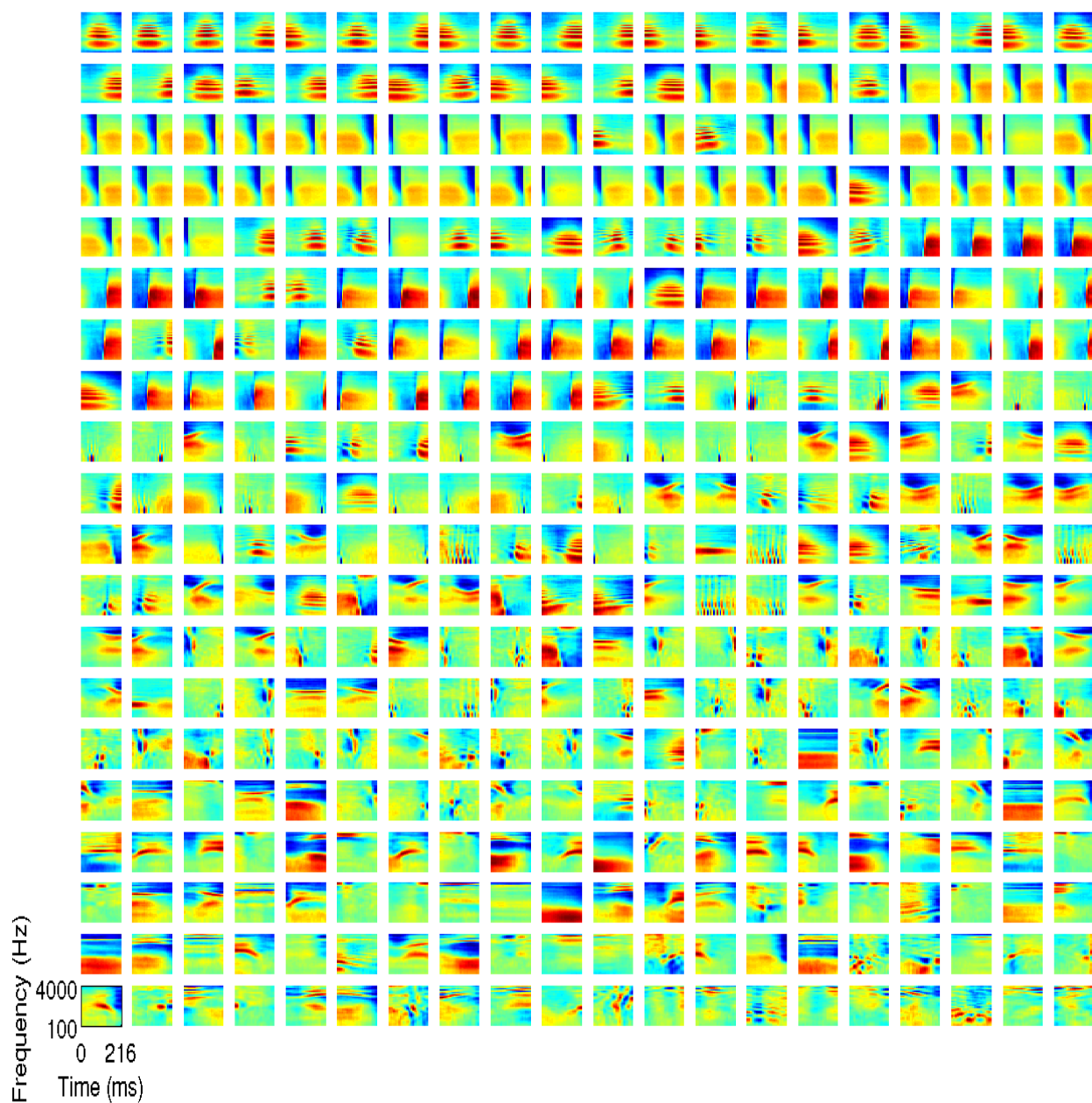


Figure 4.21: The full set of elements from a two-times overcomplete, L1-sparse dictionary trained with LCA [45] on spectrograms of speech. Same conventions as **Fig. 4.14**.



Figure 4.22: The full set of elements from a four-times overcomplete, L1-sparse dictionary trained with LCA [45] on spectrograms of speech. Same conventions as **Fig. 4.14**.

### 4.2.3 Sparsenet-trained Spectrogram Dictionaries

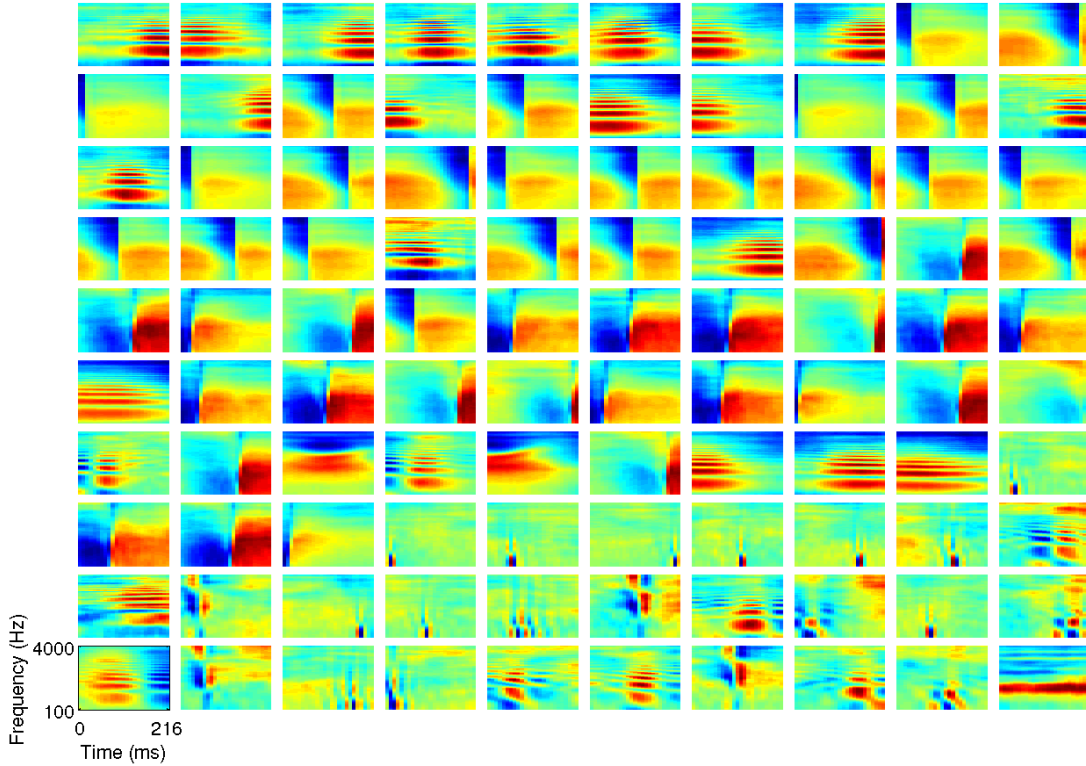


Figure 4.23: The full set of elements from a half-complete, L1-sparse dictionary trained with Sparsenet [11] on spectrograms of speech. Same conventions as **Fig. 4.14**.

I also trained some networks using Sparsenet [11] to produce soft sparse dictionaries, and again I obtained similar results as for our hard sparse dictionaries (**Figs. 4.23 and 4.24**). It has been proven mathematically [76] that signals that are actually L0-sparse can be uncovered effectively by L1-sparse coding algorithms, which might suggest that speech is indeed L0-sparse signal given that I find similar features using algorithms designed to maximize either L1 or L0 sparseness.

Thus, preprocessing with spectrograms, rather than a more nuanced cochlear model, and the degree of overcompleteness, greatly influenced the dictionaries produced by all of the sparse coding algorithms I explored, and both factors were stronger determinants of what shapes appear than the specific sparseness penalty I employed. Although I obtained similarly shaped STRFs for the various types of sparse penalties I employed, there were differences in the performance of the various dictionaries. In particular, the level of sparseness achieved across the population of model neurons exhibited different relationships with the fidelity of their representations, suggesting that some model choices resulted in population codes that were more efficient at using small numbers of neurons to represent stimuli efficiently, while others were more effective at increasing their representational power when incorporating more active neurons.

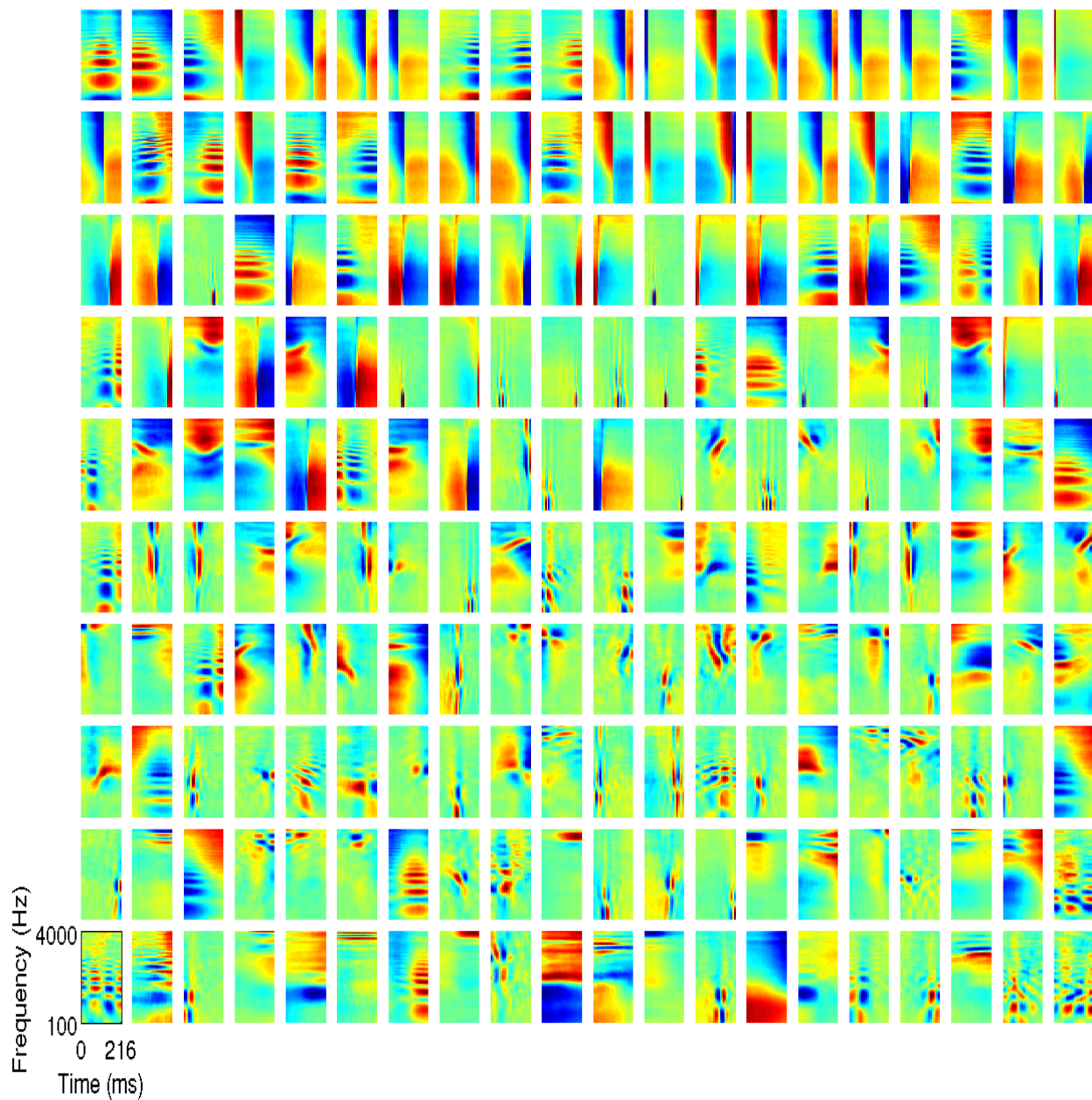


Figure 4.24: The full set of elements from a complete, L1-sparse dictionary trained with Sparsenet [11] on spectrograms of speech. Same conventions as **Fig. 4.14**.

### 4.3 Performance

The dependence of the signal-to-noise ratio (SNR) on the degree of sparseness for the dictionaries trained on spectrograms (Fig. 4.25) demonstrates that the sparse encoding algorithm exhibits a trade-off between reconstruction quality and sparseness, as one would expect. A few other general trends are evident as well. Most notably, the L0-sparse dictionaries have higher SNRs than the L1-sparse dictionaries for similar levels of sparseness. Also, the more overcomplete dictionaries have higher SNRs than half-complete ones, even with the same absolute number of active neurons. The half-complete and complete dictionaries do not show much improvement in performance even as the number of active neurons increases. Interestingly, we find that the performance of the L0-sparse dictionaries tend to saturate as the fraction of active neurons approaches unity whereas the corresponding curves for the L1-sparse dictionaries tend to curve upwards.

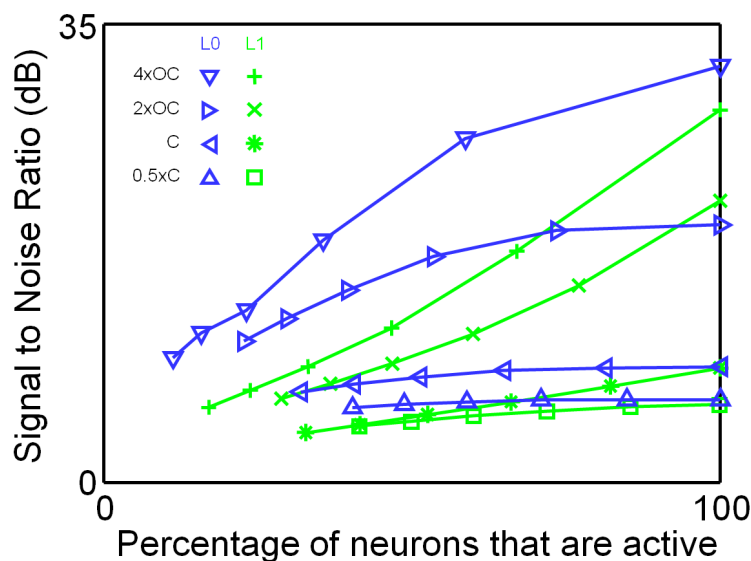


Figure 4.25: The signal to noise ratio (SNR) of sparse coding dictionaries trained on spectrograms increases with overcompleteness and with increasing numbers of active elements. Blue lines with triangles represent L0-sparse dictionaries, whereas green lines represent L1-sparse dictionaries. As expected, representations are more accurate with increasing numbers of active neurons and also when the level of overcompleteness is increased.

## 4.4 Modulation Power Spectra

My four-times overcomplete, spectrogram-trained dictionary exhibits a clear tradeoff in spectrotemporal resolution (red points, **Fig. 4.26**), similar to what has been found experimentally in IC [31]. This trend is not present in the half-complete cochleogram-trained dictionary (blue open circles, **Fig. 4.26**). Rather, these elements display a limited range of temporal modulations, but they span nearly the full range of possible spectral modulations. Thus, by this measure the spectrogram-trained dictionary is a better model of IC than the cochleogram-trained model. In a later section, I compare the shapes of the various classes of model STRFs with individual neuronal STRFs from IC, and again find good agreement between my overcomplete spectrogram-trained model and the neural data.

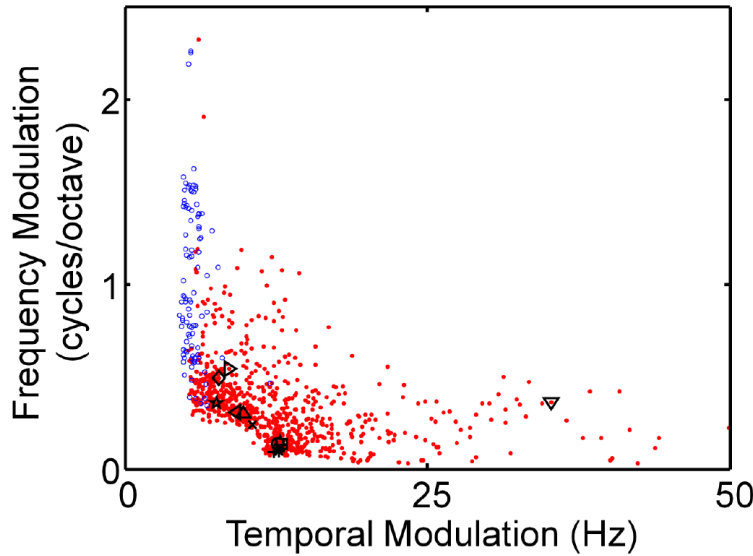


Figure 4.26: The four-times overcomplete spectrogram-trained dictionary elements (red dots; same dictionary as in **Fig. 4.17**) display a clear tradeoff between spectral and temporal modulations, similar to what has been reported for Inferior Colliculus (IC) [31]. By contrast, the half-complete cochleogram-trained dictionary (blue circles; same dictionary as in **Fig. 4.1**) exhibits a much more limited range of temporal modulations, with no such tradeoff in spectrotemporal resolution. Each data point represents the centroid of the modulation spectrum of the corresponding element. The elements shown in **Fig. 4.17** are indicated on the graph with the same symbols as before.

One might ask if the spectral and temporal tradeoff is simply due to the uncertainty principle as is the case in the auditory nerve. An issue that arose was how the uncertainty principle applied to my data. The uncertainty principle for Fourier Transformed signals is that the product of the frequency bandwidth  $\sigma_f$  and the duration of the windowed signal  $\sigma_t$  is constrained [74].

$$\sigma_f \sigma_t \geq \frac{1}{4\pi} \quad (4.1)$$

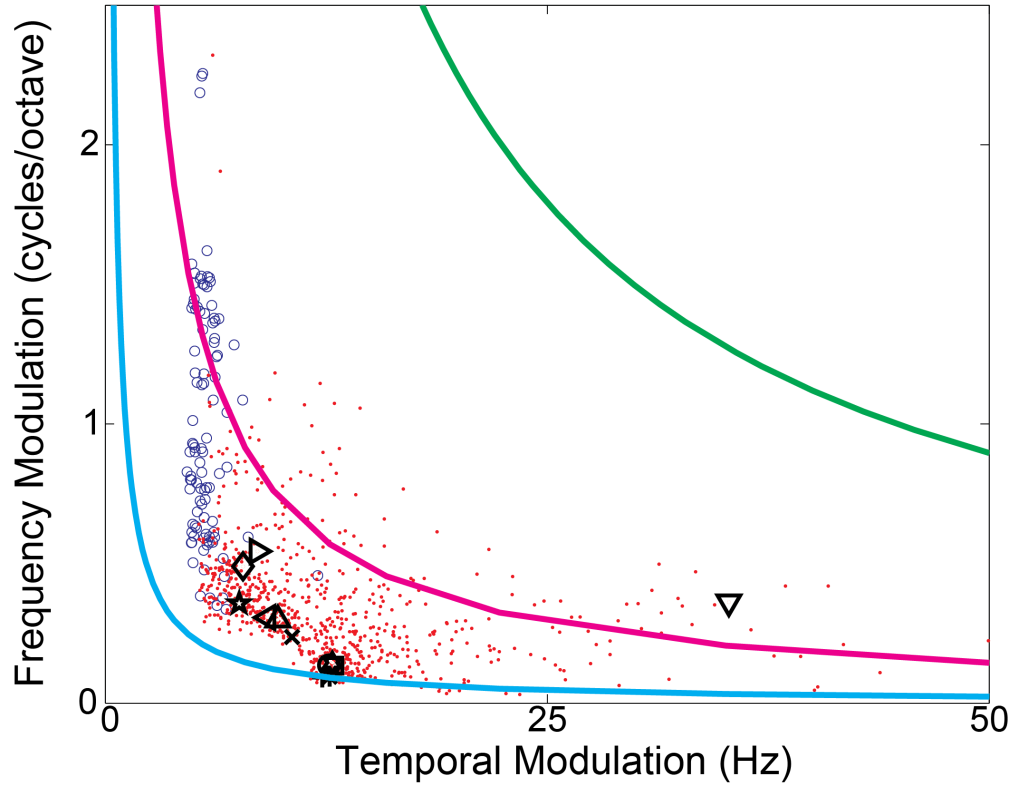


Figure 4.27: The four-times overcomplete spectrogram-trained dictionary elements (red dots; same dictionary as in **Fig. 4.17**) and the half-complete cochleogram-trained dictionary (blue circles; same dictionary as in **Fig. 4.1**). Each data point represents the centroid of the modulation spectrum of the corresponding element. The elements shown in **Fig. 4.17** are indicated on the graph with the same symbols as in that figure. As described in the text, the three colored lines represent different manifestations of the uncertainty principle. The green line is the principle applied at the lowest frequencies (100 Hz), magenta at the middle frequencies (619 Hz), and the cyan line for the highest frequencies (4000 Hz).

I can use this to limit the modulation frequencies of my dictionary units by looking at the maximum values for the modulation frequencies which are:  $\max(\omega_t) = \frac{1}{2\sigma_t}$  and  $\max(\omega_f) = \frac{1}{2\sigma_f}$ . If I invert this, the range of modulation frequencies possible is  $|\omega_t \cdot \omega_f| \leq \pi$ .

My processing of the waveforms used log-spaced frequencies which is not compatible with that form of the uncertainty principle since I measure my spectral cycles in octaves instead of Hz. If I multiple my spectral and temporal modulation frequencies, the product would not be unitless. There is not a clear analogous formulation for log-spaced frequencies. I found the limit at the top, middle, and bottom frequencies of my spectrograms.

To calculate these three lines (**Fig. 4.27**), I looked at three areas of the frequency range and treated them as if they were linearly spaced.

At the lowest frequencies, the fastest possible linear frequency modulation would be 1 cycle going from 100 to 102.9355 Hz, which equals 340.7 cyc/kHz. In our log-spacing, these three frequencies cover 0.0126 decades. For all frequencies, the fastest possible spectral frequency mod-

ulation would be 1 cyc/0.0126 decades. Combined with the unit factor (0.301 decades= 1 octave), the fastest modulation is 23.8889 cyc/octave. Altogether, the uncertainty principle with my units:

$$\omega_f \cdot \omega_t \cdot \frac{23.8889}{340.7} \leq \pi \rightarrow \omega_f \cdot \omega_t \leq 44.8 \frac{Hz \cdot cyc}{oct} \quad (4.2)$$

This equation is the green line represented in the figure. For the middle frequencies (618.8795 to 637.0647 Hz, magenta line), the resulting equation was  $\omega_t \cdot \omega_f \leq 7.2330 \text{ Hz} \cdot \text{cyc/oct}$ . For the top frequencies (3885.9 to 4000 Hz, cyan line) the resulting equation was  $\omega_t \cdot \omega_f \leq 1.1573 \text{ Hz} \cdot \text{cyc/oct}$ . Similar results were obtained for the cochleogram frequencies where I looked at the bottom frequencies for comparison,  $\omega_t \cdot \omega_f \leq 44.6318 \text{ Hz} \cdot \text{cyc/oct}$  (green line as well). If the bottom frequency line is used, then clearly all data points are almost an order of magnitude away from the uncertainty principle. This would imply that the tradeoff is due to more than just the uncertainty principle. At the lower frequency line, the majority of the points are above the line. This is clearly a violation of the uncertainty principle so it cannot possibly be the right measure for the majority of the points. Unfortunately, there is no simple way to characterize where a modulation is taking place, especially since each STRF can have modulations at multiple frequencies.

## 4.5 Analysis of Best Frequencies

In order to match the properties of experimental STRFs, I performed another population analysis. I examined the relationship between best frequency and temporal modulation frequency. The previous section demonstrated the trade-off between spectral and temporal modulation frequency. There is also an anti-correlation between best frequency and temporal modulation frequency: lower best frequency is linked to higher temporal modulation frequency. I first looked into this based on the analysis of [31]. Rodriguez et al. found that best frequency increased orthogonal to the isofrequency lamina of IC. These same neurons also showed an increase in temporal extent as the best frequency increased (see Fig. 1 of [31]). Qualitatively, I observed this in **Fig. 4.28** which shows forty-five hand-chosen elements that match the experimental pattern of localized excitation preceded or followed by inhibition. The STRFs with high best frequencies are broad in time whereas those at the lowest frequencies are very narrow in time

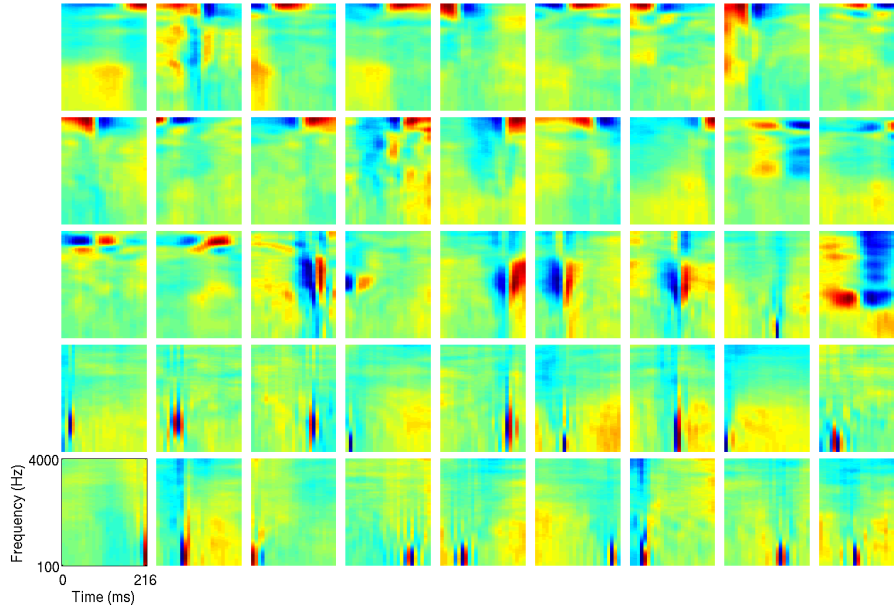
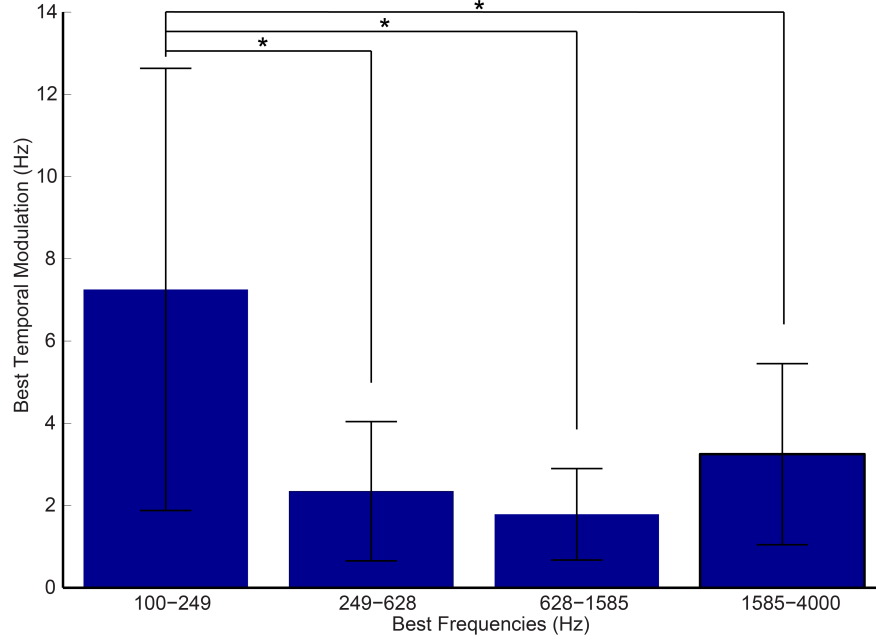


Figure 4.28: The selected forty-five elements demonstrate a relationship between temporal extent and best frequency. Units with lower best frequencies have short temporal extents and vice versa. All elements are from the four-times overcomplete L0-sparse spectrogram-trained dictionary ordered from highest to lowest best frequency. Best Frequency is defined by the overall maximum value in the STRF.

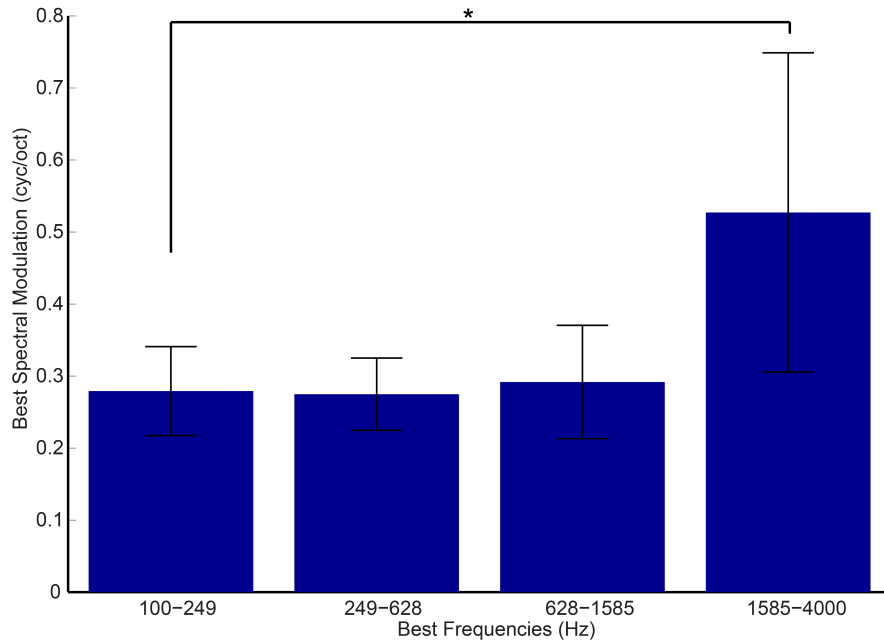
To quantify my results, I examined all eight hundred model STRFs and organized them by their best frequencies. I put all elements into one of four equally spaced frequency groups (each spanned 1.3 octaves): 100-249, 249-628, 628-1585, and 1585-4000 Hz with 132, 429, 130, and 109 units respectively. I calculated the median values of all four groups for the best temporal modulation frequency and spectral modulation frequency and plotted them in bar charts (**Fig. 4.29(a)** and **4.29(b)**). There is a decrease in the median values of the best temporal modulation frequencies as the best frequencies increase. Additionally, there is a slight correlation between best spectral

modulation frequency and best frequency. This mimics the results seen in [31].

The opposite trend is found in the auditory nerve [77]. In AN, the fibers are efficient for encoding the lower-order statistics of sound [78]; the temporal modulations are enhanced at the cost of fine spectral modulations so temporal modulations increase with best frequency. Since the opposite trend is found in IC, IC is not just inheriting the properties of the AN. Perhaps, IC is optimized for the higher-order statistics of natural sounds [78]. It is interesting to note that sparse coding predicts both of these trends depending on the input to the encoding network. Sparse coding of waveforms [17] predicts the shapes of AN fibers so the model elements show the same trend viewed in experimental AN fibers. In my case, sparse coding of spectrograms predicts the STRF shapes and trends seen in IC. This suggests that the auditory system is in some way taking advantage of the structure of natural sounds since sparse coding can predict these seemingly opposing trends.



(a)



(b)

Figure 4.29: Median values of dictionary subsets binned into equal-spaced best frequency bands (each represents 1.3 octaves). The elements are from the four-times overcomplete L0-sparse spectrogram-trained dictionary. (a) Median values of best temporal modulation frequency. (b) Median values of best spectral modulation frequency (cyc/oct). Asterisks signify significance based on the Mann-Whitney-Wilcoxon rank-sum test ( $p < 0.05$ ), and error bars represent median absolute deviation. Analysis inspired by [31].

## 4.6 Comparison to Experimental Data

Sparse coding models of natural scenes have produced visual features that closely match experimentally measured neural receptive fields found in primary visual cortex [11, 12, 75]. In the present study, my model’s STRFs resemble those of neurons recorded in multiple areas in the auditory pathway. Specifically, I find features that correspond to receptive fields found in IC [31, 30, 55, 60], MGBv [32], and A1 [53, 62, 50]. I am unaware of any previous theoretical work that has provided accurate predictions for receptive fields in these areas.

**Figs. 4.30, 4.31, and 4.32** present several examples of previously reported experimental receptive fields that qualitatively match some of my model’s dictionary elements.

IC neurons often exhibit highly localized inhibition and suppression patterns (**Fig. 4.30**), sometimes referred to as “ON” or “OFF” responses, depending on the temporal order of excitation and suppression. I show multiple examples drawn from the complete, two-times overcomplete, and four-times overcomplete dictionaries, trained on spectrograms, that exhibit these patterns. The receptive fields of three neurons recorded in gerbil IC exhibit suppression at a particular frequency followed by excitation at the same frequency (**Fig. 4.30a**). Such neurons are found in my model dictionaries (**Fig. 4.30b**). The reverse pattern is also found in which inhibition follows excitation as shown in two cat IC STRFs (**Fig. 4.30c**) with matching theoretical examples from our model dictionaries (**Fig. 4.30d**). Note that the experimental receptive fields extend to higher frequencies because the studies were done in cats and gerbils, which are sensitive to higher frequencies than we were probing with my human speech training set. In addition, note the difference in time-scales between my spectrogram representation and the experimental STRFs. One possible explanation for this is the different timescales of speech and sounds which are behaviorally relevant to cats and rodents.

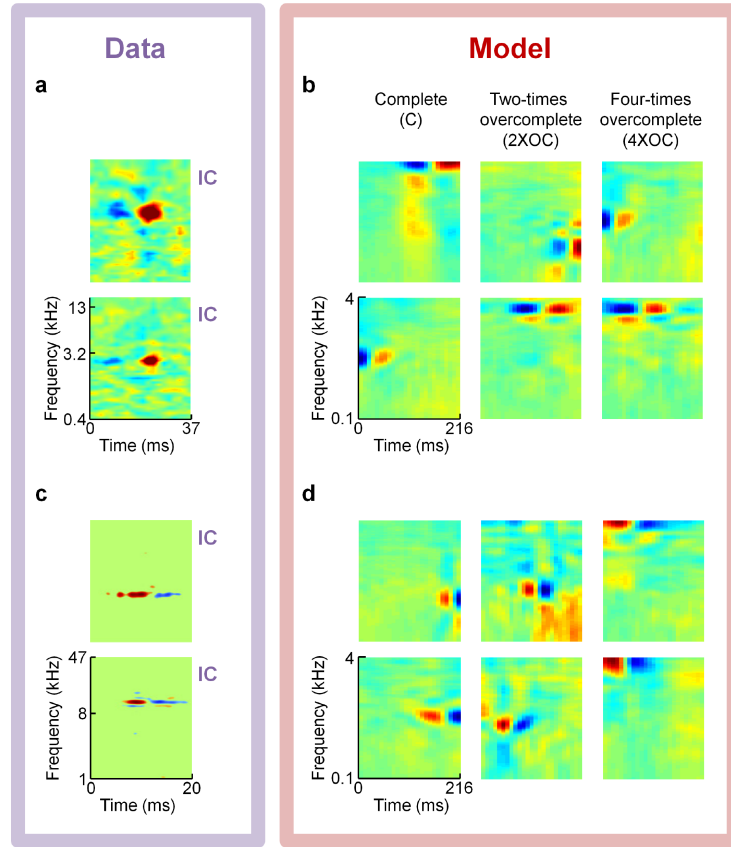


Figure 4.30: Complete and overcomplete sparse coding models trained on spectrograms of speech predict Inferior Colliculus (IC) spectro-temporal receptive field (STRF) shapes with excitatory and inhibitory subfields that are localized in frequency but separated in time. **(a)** Two examples of Gerbil IC neural STRFs [30] exhibiting ON-type response patterns with excitation following suppression; data courtesy of N.A. Lesica. **(b)** Representative model dictionary elements from each of three dictionaries that match this pattern of excitation and suppression. The three dictionaries were all trained on spectrogram representations of speech, using a hard sparseness (L0) penalty; the representations were complete (left column; **Fig. 4.15**), two-times overcomplete (middle column; **Fig. 4.16**), and four-times overcomplete (right column; **Fig. 4.18**). **(c)** Two example neuronal STRFs from cat IC [31] exhibiting OFF-type patterns with excitation preceding suppression; data courtesy of M.A. Escabí. **(d)** Other model neurons from the same set of three dictionaries as in panel **b** also exhibit this OFF-type pattern.

A common feature of thalamic and midbrain neural receptive fields is a localized checkerboard pattern of excitation and inhibition (**Fig. 4.31**), typically containing between four to 12 prominent subfields. I present experimental gerbil IC, cat IC and cat MGBv STRF's of this type in **Fig. 4.31a** beside similar examples from my model (**Fig. 4.31b**). This pattern is displayed by many elements in my sparse coding dictionaries, but to my knowledge it has not been predicted by previous theories.

Finally, I also find some less localized receptive fields that strongly resemble experimental data. Some model neurons (**Fig. 4.32b**) consist of an inhibition/excitation pattern that extends

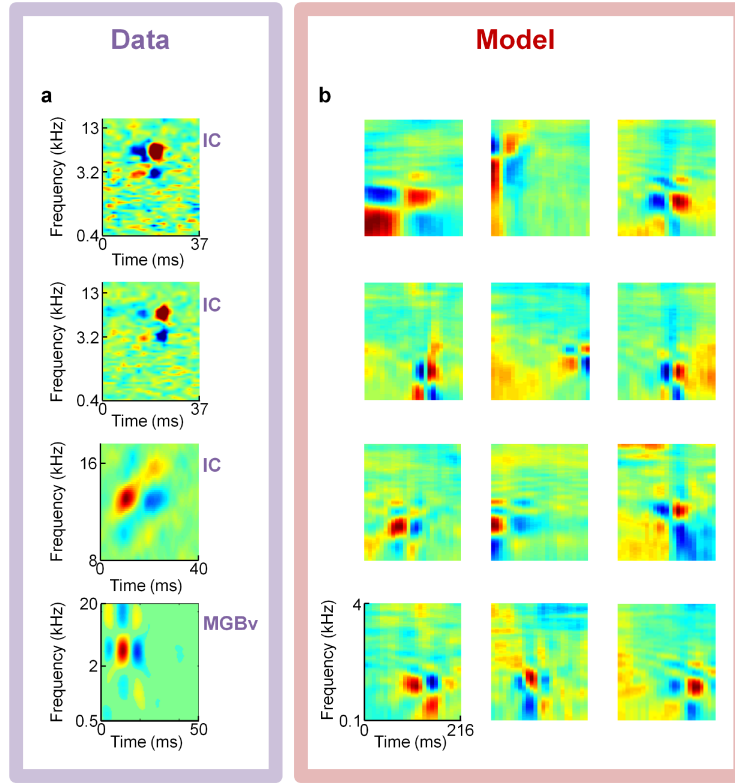


Figure 4.31: Neuronal midbrain and thalamus receptive fields and model comparisons. An overcomplete sparse coding model trained on spectrograms of speech predicts Inferior Colliculus (IC) and auditory thalamus (ventral division of the medial geniculate body; MGBv) spectro-temporal receptive fields (STRFs) consisting of localized checkerboard patterns containing roughly four to nine distinct subfields. **(a)** Example STRFs of localized checkerboard patterns from two Gerbil IC neurons [30], one cat IC neuron [60], and one cat MGBv neuron [53] (top to bottom). Data courtesy of N.A. Lesica (top two cells) and M.A. Escabí (bottom two cells). **(b)** Elements from the four-times overcomplete, L0-sparse, spectrogram-trained dictionary with similar checkerboard patterns as the neurons in panel **a**.

across most frequencies, reminiscent of broadband OFF and ON responses as reported in cat IC and rat A1 (**Fig. 4.32a**).

Another shape seen in experimental STRF's of bat IC (top), and cat A1 (bottom; **Fig. 4.32c**) is a diagonal pattern of excitation flanked by inhibition at the higher frequencies. This pattern of excitation flanked by inhibition is present in my dictionaries (**Fig. 4.32d**), also at the highest frequencies probed. These resemble the FM sweeps that are over-represented in the bat auditory system and are sometimes found in speech.

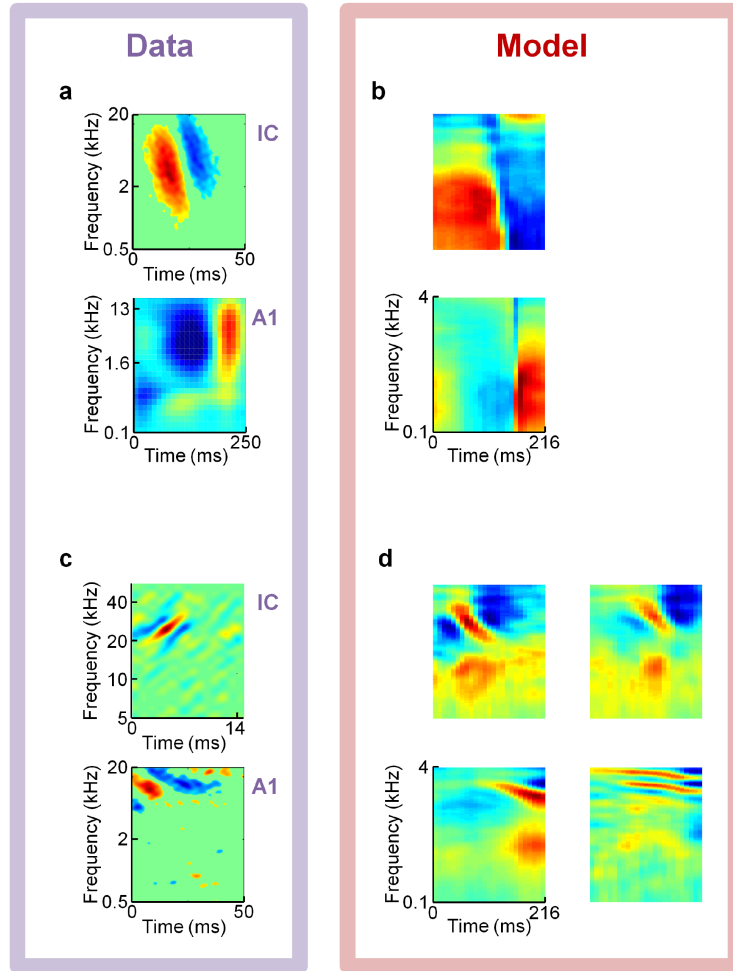


Figure 4.32: Neuronal midbrain and thalamus receptive fields and model comparisons on spectrograms of speech predicts several classes of broadband spectro-temporal receptive field (STRF) shapes found in Inferior Colliculus (IC) and primary auditory cortex (A1). (**a,b**) An example broadband OFF-type STRF from cat IC [53] (top; data courtesy of M.A. Escabí) and an example broadband ON-type subthreshold STRF from rat A1 [62] (bottom; data courtesy of M. Wehr) shown in panel **a** resemble example elements from a four-times overcomplete, L0-sparse, spectrogram-trained dictionary shown in panel **b**. (**c**) STRFs from a bat IC neuron [55] (top; data courtesy of S. Andoni) and a cat A1 neuron [53] (bottom; data courtesy of M.A. Escabí) each consist of a primary excitatory subfield that is modulated in frequency over time, flanked by similarly angled suppressive subfields. (**d**) Example STRFs from four elements taken from the same dictionary as in panels **b** exhibit similar patterns as the neuronal STRFs in panel **c**.

## 4.7 Linear Filters

My final analysis was linear filter estimation (STA). I worked exclusively with the four-times overcomplete L0-sparse spectrogram-trained dictionary.

### 4.7.1 The Model

I treated my model elements as if they are the receptive fields of real auditory neurons, and probed them with nine different stimulus sets (see Sec. 3.5 for full description of each category) by using the original dictionary to encode these new sounds using the LCA algorithm. The inferred coefficients are the correlates to spikes from neurons or the subthreshold membrane potential. Using the encodings, I averaged over all the probe stimuli in a stimulus set to get linear filter estimates for each model neuron; either standard STAs or whitened STAs.

### 4.7.2 Standard Spike-Triggered Average

I first studied standard STAs in which the stimulus correlations were not removed. As expected, not accounting for stimulus correlations typically produces worse estimates (measured by mean absolute error). I show one example neuron (**Fig. 4.33**) of a checkerboard pattern of excitation and inhibition (a hallmark of IC neurons [53, 30, 31]). The original is seen in the top left of the figure (red box). Some of the probe stimulus sets such as the White Noise (WWN), Classical Music (CM), Dynamic Moving Ripples (DMR), Natural Scenes (NS), and Image White Noise (IWN) qualitatively capture the main features of the original receptive field. The other types poorly represent the original receptive field by not exhibiting the featured checkerboard pattern. As a population, the eight hundred model neurons were poorly estimated by the probe stimuli using standard STAs. Main features of receptive fields were distorted or completely absent.

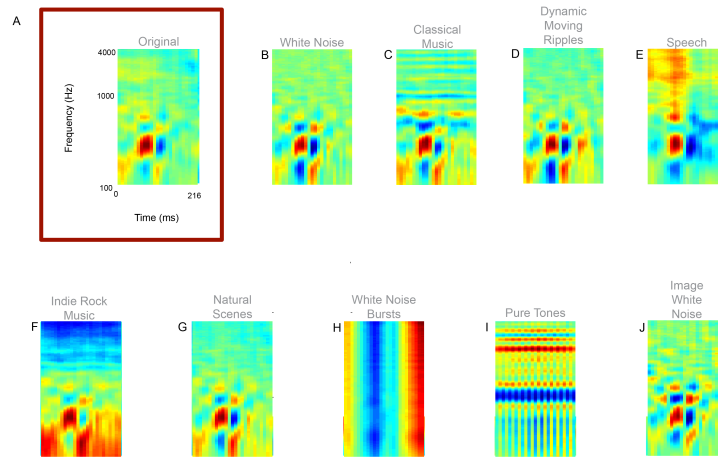


Figure 4.33: Different probe stimulus sets produce different standard spike-triggered averages for a checkerboard pattern with correlations intact. **(a)** The original receptive field from the four-times overcomplete L0-sparse spectrogram-trained dictionary. **(b)** Estimate from white noise. **(c)** Estimate from classical music. **(d)** Estimate from dynamic moving ripples. **(e)** Estimate from a holdout speech of set. **(f)** Estimate from indie rock music. **(g)** Estimate from natural scenes. **(h)** Estimate from white noise bursts. **(i)** Estimate from pure tones. **(j)** Estimate from image white noise. Each rectangle represents the spectro-temporal receptive field (STRF) of a single element in the dictionary; time is plotted along the horizontal axis (from 0 to 216 msec) and log frequency is plotted along the vertical axis, with frequencies ranging from 100 Hz to 4000 Hz.

### 4.7.3 Whitened Spike-Triggered Average

I then examined whitened STAs of my model neurons. **Fig. 4.34** shows the same original receptive field as **Fig. 4.33** along with linear filter estimates where stimulus correlations have been removed. Unlike the previous figure, a different group of probe stimuli provide the best estimates: CM, HS, and NS. For certain receptive fields taking stimulus correlations into effect can have a negative effect for some probe stimulus sets. It's unclear why this would be the case. On average, there was great improvement of STRF estimation when the correlations were taken into account, both visually and in terms of mean absolute error.

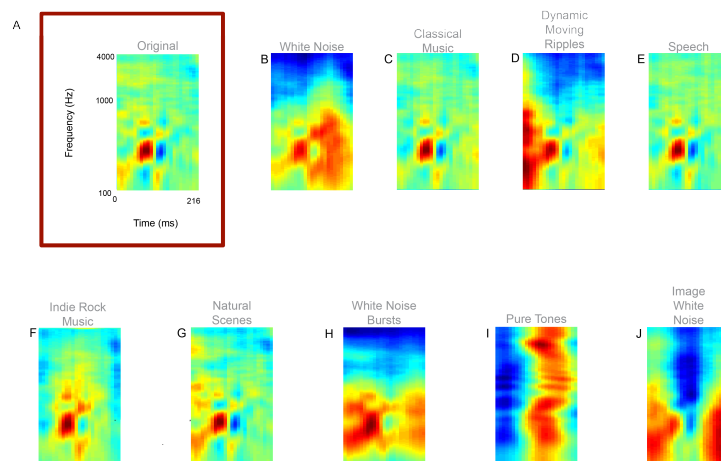


Figure 4.34: Different probe stimulus sets produce different whitened spike-triggered averages for the same checkerboard pattern as **Fig. 4.33** with correlations removed. (a) The original receptive field from the four-times overcomplete L0-sparse spectrogram-trained dictionary. (b) Estimate from white noise. (c) Estimate from classical music. (d) Estimate from dynamic moving ripples. (e) Estimate from a holdout speech of set. (f) Estimate from indie rock music. (g) Estimate from natural scenes. (h) Estimate from white noise bursts. (i) Estimate from pure tones. (j) Estimate from image white noise.

**Fig. 4.35** shows a harmonic stack, a shape that has not been reported as a receptive field shape in the mammalian auditory midbrain, thalamus, and cortex. My results suggest that this auditory feature cannot be recovered with traditional stimulus sets such as PT, WN, and DMR. Another reason that these broadband shapes might not have been reported is because experimentalists probe around the frequency area to which the neuron most strongly responds (the so-called ‘Frequency Response Area’). This could make finding broadband sounds impossible.

**Fig. 4.36** shows a tightly bound on/off set response localized to a few frequencies. These have been seen at low best frequencies as found in an experiment with DMR [31]. Only CM, HS, IRM, and NS were able to pick up on this feature using my simple whitened STA.

The final shape is a tight on/off pattern that lasts for many cycles in time (**Fig. 4.37**). A feature with such extended temporal modulations has not been reported before. Again only a few probe stimulus sets were able to capture this type of shape. Though I show only five out of eight hundred

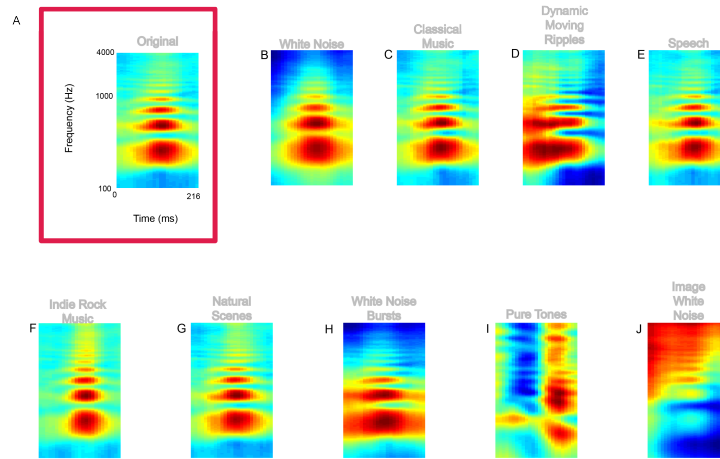


Figure 4.35: Different probe stimulus sets produce different receptive field estimates for a harmonic stack with correlations removed. **(a)** The original receptive field from the four-times overcomplete L0-sparse spectrogram-trained dictionary. **(b)** Estimate from white noise. **(c)** Estimate from classical music. **(d)** Estimate from dynamic moving ripples. **(e)** Estimate from a holdout speech of set. **(f)** Estimate from indie rock music. **(g)** Estimate from natural scenes. **(h)** Estimate from white noise bursts. **(i)** Estimate from pure tones. **(j)** Estimate from image white noise.

units, these trends existed for the majority of the dictionary.

Examining the dictionary as a whole, some cell-types were easily uncovered. One surprise was that DMR could only uncover a subset of the model auditory features even though they are often used as experimental stimuli. Part of the reason for these issues might be because I did not do any regularization. DMR often uncovered a noisy version of the main feature. Since experimentalists often do some kind-of thresholding or smoothing of their STRFs, the features might be more apparent.

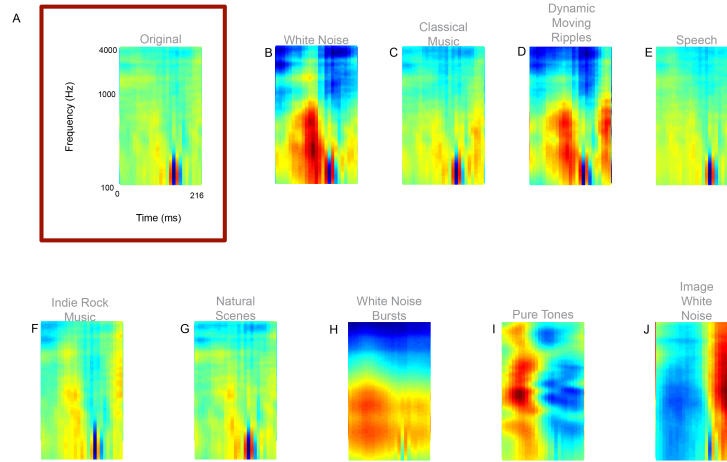


Figure 4.36: Different probe stimulus sets produce different receptive field estimates for a tightly localized on/off pattern with correlations removed. (a) The original receptive field from the four-times overcomplete L0-sparse spectrogram-trained dictionary. (b) Estimate from white noise. (c) Estimate from classical music. (d) Estimate from dynamic moving ripples. (e) Estimate from a holdout speech of set. (f) Estimate from indie rock music. (g) Estimate from natural scenes. (h) Estimate from white noise bursts. (i) Estimate from pure tones. (j) Estimate from image white noise.

#### 4.7.4 Mean Absolute Error

To quantify the global performance of the different probe stimulus sets over all model neurons, I calculated the mean absolute error (MAE) by treating the original basis function as ground truth and the linear filter estimates as predictions. For each probe stimulus set, I then averaged over all eight hundred predictions.

Almost all of the probe stimulus sets showed decreases in MAE once stimuli correlations were removed (**Fig. 4.38**). The stimulus sets that performed best over all cell types were CM, HS, and NS. This is reflected in the specific examples shown above. Of these, HS is guaranteed to do a good job since probing the dictionary with the same type of stimulus used for training is mathematically proven to uncover the original dictionary. The other two types probably do a good job since they contain the full range of complexity of natural sounds.

Pure tones are the least accurate out of all of the stimuli which makes sense as they have no way of accurately capturing the temporal structure of the data. Additionally, the pure tones were only active at a single frequency so concurring frequencies would be impossible to fully capture.

The different types of white noise also failed despite the fact that mathematically uncorrelated stimuli such as Gaussian white noise should be able to capture the full system. However, the previous statement is only true in the case of infinite stimuli. Both in this model work and real experimental work, there is no way to fully explore the stimulus space sufficiently to be able to estimate receptive fields.

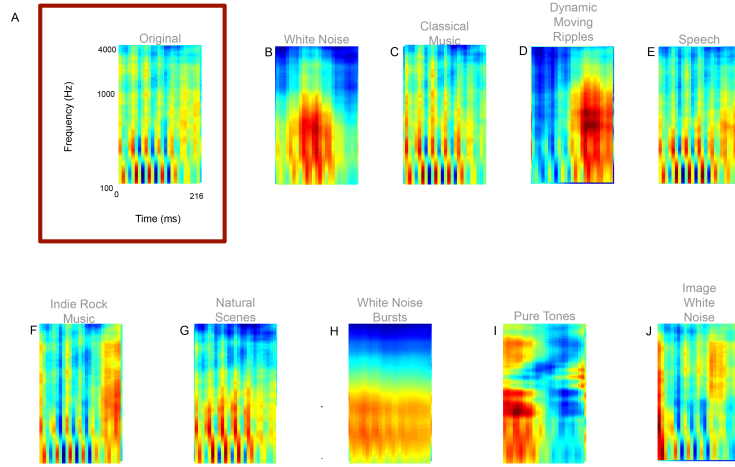


Figure 4.37: Different probe stimulus sets produce different receptive field estimates for a tightly localized on/off pattern that lasts for many cycles in time with correlations removed. (a) The original receptive field from the four times overcomplete L0-sparse spectrogram-trained dictionary. (b) Estimate from white noise. (c) Estimate from classical music. (d) Estimate from dynamic moving ripples. (e) Estimate from a holdout speech of set. (f) Estimate from indie rock music. (g) Estimate from natural scenes. (h) Estimate from white noise bursts. (i) Estimate from pure tones. (j) Estimate from image white noise.

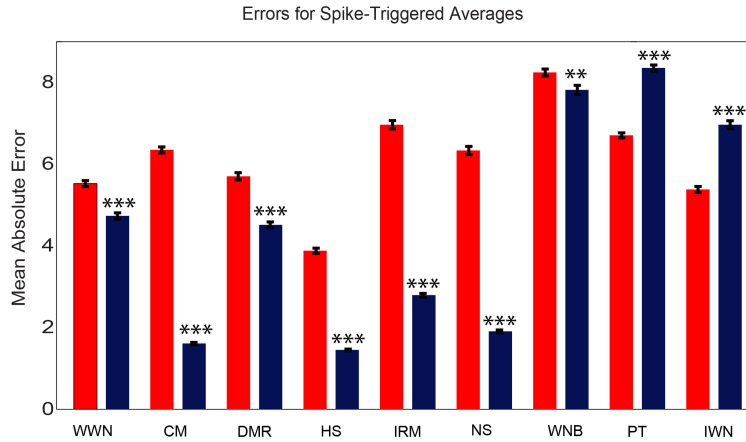


Figure 4.38: Removing stimulus correlations lowers the mean absolute error (MAE) of most probe stimulus sets. (Red) MAE for standard spike-triggered average for all nine probe stimuli averaged over all eight hundred estimates. (Blue) MAE for whitened spike-triggered average for all nine probe stimuli averaged over all eight hundred estimates. WWN: Waveform White Noise; CM: Classical Music; DMR: Dynamic Moving Ripples; HS: Holdout set of speech; IRM: Indie Rock Music; NS: Natural Scenes; WNB: White Noise Bursts; PT: Pure Tones; IWN: Image White Noise. Bars indicate standard error and asterisks indicate significance.

# Chapter 5

## Conclusions

I have applied the principle of sparse coding to spectrogram and cochleogram representations of human speech recordings in order to uncover some important features of natural sounds. Of the various models I considered, I have found that the specific form of preprocessing (*i.e.*, cochleograms vs. spectrograms) and the degree of overcompleteness are the most significant factors in determining the complexity and diversity of receptive field shapes. Importantly, I have also found that features learned by my sparse coding model resemble a diverse set of receptive field shapes in IC, as well as MGBv and A1. Even though a spectrogram may not provide as accurate a representation of the output of the cochlea as a more explicit cochleogram model, such as the one I explored here, I have found that sparse coding of spectrograms yields closer agreement to experimentally measured receptive fields, demonstrating that we can infer important aspects of sensory processing in the brain by identifying the statistically important features of natural sounds without having to impose many constraints from biology into my models from the outset.

Indeed, it is worth emphasizing that the agreement I have found did not result from fitting the data, *per se*; it emerged naturally from the statistics of the speech data we used to train my model. Specifically, the model parameters I explored — undercomplete vs. overcomplete representation, L0 vs. L1 sparseness penalty, and cochleogram vs. spectrogram preprocessing — represent a low-dimensional space of essentially eight different choices compared with the rich, high-dimensional space of potential STRF shapes I could have obtained.

Intriguingly, while I have emphasized the agreement between my model and IC, the receptive fields I have found resemble experimental data from multiple levels of the mammalian ascending auditory pathway. This may reflect the possibility that the auditory pathway is not strictly hierarchical, so that neurons in different anatomical locations may perform similar roles, and thus are represented by neurons from the same sparse coding dictionary.

This view is consistent with the well-known observation that there is a great deal of feedback from higher to lower stages of processing in the sub-cortical auditory pathway [79], as compared with the visual pathway, for example. The tradeoff in spectrotemporal resolution we have found in our model resembles that of IC, which is the lowest stage of the ascending auditory pathway

to exhibit a tradeoff that cannot be accounted for by the uncertainty principle, as is the case for auditory nerve fibers [31], but it remains to be seen if such a tradeoff also exists in MGBv or A1.

A related issue is that an individual neuron might play different roles depending on the stimulus ensemble being presented to the nervous system. In fact, changing the contrast level of the acoustic stimuli used to probe individual IC neurons can affect the number of prominent subfields in the measured STRF of the neuron [30]. My model does not specify which neuron should represent any given feature, it just predicts the STRFs that should be represented in the neural population in order to achieve a sparse encoding of the stimulus.

Moreover, for even moderate levels of overcompleteness, my sparse coding dictionaries include categories of features that have not been reported in the experimental literature. For example, the STRF shown in **Fig. 4.17k** represents a well-defined class of elements in my sparse dictionaries, but I am unaware of reports of this type of STRF in the auditory pathway. One thing to note is that harmonic stacks are the least active elements in my model. If these are represented in the auditory system, there would be a selection bias against them since these neurons would not fire that much. A typical experiment might only uncover one or two examples of such STRFs so they might not make it into publication. Thus, my theoretical receptive fields could be used to develop acoustic stimuli that might drive auditory neurons that do not respond to traditional probe stimuli. In particular, my dictionaries contain many broadband STRFs with complex structures. These broadband neurons may not have been found experimentally since by necessity researchers often probe neurons extensively with stimuli that are concentrated around the neuron's best frequency.

I have presented several classes of STRFs from my model that qualitatively match the shapes of neural receptive fields, but in many cases the neurons are sensitive to higher frequencies than the model neurons. This is likely due to the fact that I trained my network on human speech, which has its greatest power in the low kHz range, whereas the example neural data available in the literature come from animals with hearing that extends to much higher acoustic frequencies, and with much higher-pitched vocalizations, than humans.

Even if sparse coding is, indeed, a central organizing principle throughout the nervous system, it could still be that the sparse representations I predict with my model correspond best to the subthreshold, postsynaptic responses of the membrane potentials of neurons, rather than their spiking outputs. In fact, I show an example of a subthreshold STRF (**Fig. 4.32** bottom) that agrees well with one class of broadband model STRFs (**Fig. 4.32**). The tuning properties of postsynaptic responses are typically broader than spiking responses, as one would expect, which could offer a clue as to which is more naturally associated with model dictionary elements. If my model elements are to be interpreted as subthreshold responses, then the profoundly unresponsive regions surrounding the active subfields of the neuronal STRFs could be more accurately fit by my model STRFs after they are post-processed by being passed through a model of a spiking neuron with a finite spike threshold.

It is encouraging that sparse encoding of speech can identify acoustic features that resemble neuronal STRFs from auditory midbrain, as well as those in thalamus and cortex, and it is notable that the majority of these features bear little resemblance to the Gabor-like shapes and elongated

edge detectors that have been predicted by sparse coding representations of natural images. Clearly, my results are not an unavoidable consequence of the sparse coding procedure itself, but instead reflect the structure of the speech spectrograms and cochleograms I have used to train my model. Previous work to categorize receptive fields in A1 has often focused on oriented features that are localized in time and frequency [28, 80], and some authors have suggested that such Gabor-like features are the primary cell types in A1 [33], but the emerging picture of the panoply of STRF shapes in IC, MGBv, and A1 is much more complex, with several distinct classes of features, just as I have found with my model. An important next step will be to develop parameterized functional forms for the various classes of STRFs I have found, which can assume the role that Gabor wavelets have played in visual studies. I hope that this approach will continue to yield insights into sensory processing in the ascending auditory pathway.

My linear filter estimation study has yielded some interesting results on the effects of probe stimulus sets on STRF estimation. Here I have shown that the probe stimuli do make a huge difference in receptive field estimation.

As expected, removing stimulus correlations vastly improves receptive field estimation, but does not fully correct the STRF estimation. Even with correlations removed, some probe stimulus sets still outperform others. However, Klein et. al [63] suggests that too many dynamic moving ripples are needed to give an accurate estimate for a typical experiment. This could explain why the ripples could not uncover some cell shapes.

Additionally, it is surprising that filter estimates occasionally appear worse when they are whitened. Qualitatively, the receptive fields can appear worse even though quantitatively the error is smaller. This is because the algorithm picks the best global solution so the overall error can decrease although auditory features that are salient to human eyes may look worse.

Many probe stimulus sets could give different estimators because the whitened STA is still a biased estimator even in the limit of infinite data if certain conditions are not met [81]. For example, the estimator is only unbiased if the distribution is radially symmetric. If this were true, then the standard STA would be the same as the whitened STA. Since this is not the case for most stimuli, the linear filter estimate is not a full description of the neuron.

Another issue I did not explore was regularization. This is more of an issue with experiments where there is a limit to the amount of data taken. In that case, the autocovariance matrix may have eigenvalues that are close to zero resulting in a noisy pseudoinverse [62]. Regularization constrains parameter values to alleviate this issue [82]. This is a direction for future work.

My results suggest that some stimuli do give better STAs than others particularly classical music, human speech, and natural scenes. I hope that my work will be useful to experimentalists when they are choosing their stimuli.

Overall, I have shown that sparse coding of speech data can predict some properties of the auditory system, which supports the efficient coding hypothesis. It is also telling that my work on speech was able to uncover features in many different animals, suggesting that speech spans the full space of natural sounds and that brains are adapted to have some general features in common. There are many possible future directions for this work. The first would be to find some automated

way to cluster the different cell-types found in my dictionary. I also hope that someday there will be parameterizations for all of the different shapes, in the same way that the canonical parameterization for V1 RFs is a Gabor function. It will also be interesting to someday invert these STRFs and listen to the sounds that they represent. The linear filter estimation work can also be continued by looking at issues of regularization as well as more complicated measures like spike-triggered covariance or maximally informative dimensions.

# Bibliography

- [1] Laughlin S (2001) Energy as a constraint on the coding and processing of sensory information. *Curr Opin Neurobiol* 11: 475-480.
- [2] Attneave F (1954) Some informational aspects of visual perception. *Psychol Rev* 61: 183-193.
- [3] Barlow H (1961) Possible principles underlying the transformations of sensory messages. In: Rosenblith W, editor, *Sensory Communication*, Cambridge: MIT Press. pp. 217-234.
- [4] Atick J, Redlich A (1992) What does the retina know about natural scenes? *Neural Comput* 4: 196-210.
- [5] Laughlin S (1981) A simple coding procedure enhances a neuron's information capacity. *Z Naturforsch* 36c: 910-912.
- [6] Rieke F, Warland D, de Ruyter van Steveninck R, Bialek W (1999) *Spikes: Exploring the neural code*. MIT Press.
- [7] DeWeese MR (1996) Optimization principles for the neural code. *Network* 7(2): 325–331.
- [8] Zhao L, Zhaoping L (2011) Understanding auditory spectro-temporal receptive fields and their changes with input statistics by efficient coding principles. *PLoS Comp Bio* 7(8): e1002123.
- [9] Földiák (1990) Forming sparse representations by local anti-hebbian learning. *Biol Cybern* 64: 165–170.
- [10] Levy W, Baxter R (1996) Energy efficient neural codes. *Neural Computation* 8: 531–1111.
- [11] Olshausen B, Field D (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381: 607–609.
- [12] Rehn M, Sommer F (2007) A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. *J Comput Neurosci* 22: 135–146.
- [13] DeWeese MR, Hromdka T, Zador A (2005) Reliability and representational bandwidth in the auditory cortex. *Neuron* 48(3): 479–488.

- [14] DeWeese MR, Wehr M, Zador A (2003) Binary spiking in auditory cortex. *J Neurosci* 23(21): 7940–7949.
- [15] Hromdka T, DeWeese M, Zador A (2008) Sparse representation of sounds in the unanesthetized auditory cortex. *PLoS Biol* 6(1): e16.
- [16] Lewicki M (2002) Efficient coding of natural sounds. *Nature Neuroscience* 5: 356–1111.
- [17] Smith E, Lewicki M (2006) Efficient auditory coding. *Nature* 439: 978-982.
- [18] Klein D, König P, Körding K (2003) Sparse spectrotemporal coding of sounds. *EURASIP Journal on Applied Signal Processing* EURASIP Journal on Applied Signal Processing: 659-667.
- [19] Aertsen A, Johannesma P (1981) A comparison of the spectro-temporal sensitivity of auditory neurons to tonal and natural stimuli. *Biol Cybern* 42: 142-156.
- [20] Körding K, König P, Klein D (2001) Learning of sparse auditory receptive fields. *Proc Int Joint Conf Neural Networks (IJCNN)* .
- [21] Henaff M, Jarrett K, Kavukcuoglu K, LeCun Y (2011) Unsupervised learning of sparse features for scalable audio classification. *Proceedings of International Symposium on Music Information Retrieval (ISMIR)* .
- [22] deBoer E, Kuyper P (1968) Triggered correlation. *IEEE Transaction on Biomedical Engineering* 15: 169–179.
- [23] Young E (1998) What’s the best sound? *Science* 280: 1402-1403.
- [24] Dayan P, Abbott L (2001) *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, Massachusetts: MIT Press.
- [25] Eggermont J, Johannesma P, Aertsen A (1983) Reverse-correlation methods in auditory research. *Quarterly Reviews of Biophysics* 16: 341–414.
- [26] DeAngelis G, Ohzawa I, Freeman R (1993) Spatiotemporal organization of simple cell receptive fields in the cat’s striate cortex: II. linearity of temporal and spatial summation. *Journal of Neurophysiology* 69: 1118–1135.
- [27] Reid R, Shapley R (1992) Spatial structure of cone inputs to receptive fields in primate lateral geniculate nucleus. *Nature* 356: 716–718.
- [28] Depireux D, Simon J, Klein D, Shamma S (2001) Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *J Neurophysiol* 85: 1220–1111.
- [29] Theunissen F, Sen K, Doupe A (2000) Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *J Neurosci* 20: 2315–1111.

- [30] Lesica N, Grohe B (2008) Dynamic spectrotemporal feature selectivity in the auditory mid-brain. *J Neurosci* 28: 5412-5421.
- [31] Rodríguez F, Read H, Escabí M (2010) Spectral and temporal modulation tradeoff in the inferior colliculus. *J Neurophysiol* 103: 887-903.
- [32] Miller L, Escabí M, HL R, Schreiner C (2002) Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *J Neurophysiol* 89: 516-527.
- [33] deCharms R, Blake D, Merzenich M (1998) Optimizing sound features for cortical neurons. *Science* 280: 1439–1111.
- [34] Schwartz O, Pillow J, Rust N, Simoncelli E (2006) Spike-triggered neural characterization. *Journal of Vision* 6: 484–507.
- [35] Theunissen F, David S, Singh N, Hsu A, Vinje W, et al. (2001) Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Comput Neural Syst* 12: 289-316.
- [36] Klein D, Simon DDJ, , Shamma S (2000) Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design. *J Comput Neurosci* 9: 85-111.
- [37] Linden J, Liu R, Sahani M, Schreiner C, Merzenich M (2003) Spectrotemporal structure of receptive fields in areas a1 and aaf of mouse auditory cortex. *J Neurophysiol* 90: 2660-2675.
- [38] Barlow H (1972) Single units and sensation: A neuron doctrine for perceptual psychology. *Perception* 1: 371–394.
- [39] Hebb D (1949) *The organization of behavior*. New York: Wiley and Sons.
- [40] Field D (1987) Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America* 4: 2379-2394.
- [41] Ruderman D (1994) Origins of scaling in natural images. *Vision Res* 37: 3385-3398.
- [42] Bell A, Sejnowski T (1997) The “independent components” of natural scenes are edge filters. *Vision research* 37: 3327–3338.
- [43] Lewicki M, Sejnowski T (2000) Learning overcomplete representations. *Neural Computation* 12: 337–111.
- [44] Ringach D (2002) Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *J Neurophys* 88: 455–1111.
- [45] Rozell C, Johnson D, Baraniuk R, Olshausen B (2008) Sparse coding via thresholding and local competition in neural circuits. *Neural Computation* 20: 2526–2563.

- [46] Wang J, Olshausen B, Ming V (2008) A sparse subspace model of higher-level sound structure. In: COSYNE.
- [47] Purves D, Augustine G, Fitzpatrick D, Hall W, LaMantia AS, et al., editors (2008) Neuroscience (4th edition). Sunderland, MA: Sinauer Associates, Inc.
- [48] Rieke F, Bodnar D, Bialek W (1995) Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proc R Soc Lond B* 262: 259–265.
- [49] Nelken I, Rotman Y, Yosef O (1999) Responses of auditory-cortex neurons to structural features of natural sounds. *Nature* 397: 154–1111.
- [50] Fritz J, Shamma S, Elhilali M, Klein D (2003) Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat Neurosci* 6: 1216-1223.
- [51] Barbour D, Wang X (2003) Contrast tuning in auditory cortex. *Science* 299: 1073–1111.
- [52] Nelken I (2004) Processing of complex stimuli and natural scenes in the auditory cortex. *Curr Opin in Neurosci* 14: 474–1111.
- [53] Escabí M, Read H (2005) Neural mechanisms for spectral analysis in the auditory midbrain, thalamus, and cortex. *International Review of Neurobiology* 70: 207-252.
- [54] Wang X, Lu T, Snider R, Liang L (2005) Sustained firing in auditory cortex evoked by preferred stimuli. *Nature* 435: 341–1111.
- [55] Andoni S, Li N, Pollak G (2007) Spectrotemporal receptive fields in the inferior colliculus revealing selectivity for spectral motion in conspecific vocalizations. *J Neurosci* 27: 4882-4893.
- [56] Nagel K, Doupe A (2008) Organizing principles of spectro-temporal encoding in the avian primary auditory area field l. *Neuron* 58: 938–1111.
- [57] Shechter B, Marvit P, Depireux D (2010) Lagged cells in the inferior colliculus of the awake ferret. *Eur J Neurosci* 31(1): 42-8.
- [58] Amin N, Gill P, Theunissen F (2010) Role of the zebra finch auditory thalamus in generating complex representations for natural sounds. *J Neurophys* 104: 784-79.
- [59] Backoff P, Clopton B (1991) A spectrotemporal analysis of dc of single unit responses to wideband noise in guinea pig. *Hearing Research* 53: 28-40.
- [60] Qiu A, Schreiner C, Escabí M (2003) Gabor analysis of auditory midbrain receptive fields: spectro-temporal and binaural composition. *J Neurophysiol* 90: 456-476.
- [61] Singh N, Theunissen F (2003) Modulation spectra of natural sounds and ethological theories of auditory processing. *J Acoust Soc Am* 114: 3394–1111.

- [62] Machens C, Wehr M, Zador A (2004) Linearity of cortical receptive fields measured with natural sounds. *J Neurosci* 24: 1089-1100.
- [63] Klein D, Simon J, Depireux D, Shamma S (2006) Stimulus-invariant processing and spectrotemporal reverse correlation in primary auditory cortex. *J Comput Neurosci* 20: 111-136.
- [64] Gill P, Zhang J, Woolley S, Fremouw T, Theunissen F (2006) Sound representation methods for spectro-temporal receptive field estimation. *J Comp Neuro* 21(1): 5-20.
- [65] Atencio C, Sharpee T, Schreiner C (2008) Cooperative nonlinearities in auditory cortical neurons. *Neuron* 58: 956-966.
- [66] Woolley S, Theunissen PGF (2006) Stimulus-dependent auditory tuning results in synchronous population coding of vocalizations in the songbird midbrain. *J Neurosci* 26(9): 2499-512.
- [67] Elliott T, Theunissen F (2009) The modulation transfer function for speech intelligibility. *PLoS Comput Bio* 5(3): e1000302. doi:10.1371/journal.pcbi.1000302.
- [68] Woolley S, Fremouw T, Hsu A, Theunissen F (2005) Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nat Neuroscience* 8: 1371-9.
- [69] Lyon R (1982) A computational model of filtering, detection, and compression in the cochlea. In: *IEEE*. pp. 1282-1285.
- [70] Slaney M (1998) Auditory toolbox version 2. Technical Report 010, Interval Research Corporation.
- [71] Lee H, Largman Y, Pham P, Ng A (2009) Unsupervised feature learning for audio classification using convolutional deep belief networks. In: *NIPS*.
- [72] Garofolo J, et al (1993) Timit acoustic-phonetic continuous speech corpus. In: *Linguistic Data Consortium*. Philadelphia.
- [73] Stevens S, Volkman J, Newman EB (1937) A scale for the measurement of the psychological magnitude pitch. *J Acoust Soc Am* 8: 185-190.
- [74] Cohen L (1995) *Time-Frequency Analysis*. Englewood Cliffs, New Jersey: Prentice Hall PTR.
- [75] Olshausen B, Cadieu C, Warland D (2009) Learning real and complex overcomplete representations from the statistics of natural images. In: *Proc. SPIE*. volume 7446.
- [76] Donoho D (2004) Compressed sensing. *IEEE Tran Inform Theory* 52: 1289-1396.
- [77] Jorin P, Yin T (1992) Responses to amplitude-modulated tones in the auditory nerve of the cat. *J Acoust Soc Am* 91: 215-232.

- [78] Rodríguez F, Chen C, Read H, Escabí M (2010) Neural modulation tuning characteristics scale to efficiently encode natural sound statistics. *J Neurosci* 30 (47): 15969-15980.
- [79] Read H, Winer J, Schreiner C (2002) Functional architecture of auditory cortex. *Curr Opin Neurobiol* 12: 433-440.
- [80] Shamma S (2001) On the role of space and time in auditory processing. *TRENDS Cog Sci* 5: 340–348.
- [81] Paninski L (2003) Convergence properties of three spike-triggered analysis techniques. *Network: Comput Neural Syst* 14: 437-464.
- [82] Hastie T, Tibshirani R, Friedman J (2001) *The elements of statistical learning theory*. New York: Springer.

# Appendix A

## List of Abbreviations

Abbreviation	Term
A1	Primary Auditory Cortex
AGC	Adaptive Gain Control
AN	Auditory Nerve
BF	Best Frequency
CM	Classical Music
DMR	Dynamic Moving Ripples
FM	Frequency-Modulated
IC	Inferior Colliculus
ICA	Independent Components Analysis
IRM	Indie Rock Music
IWN	Image White Noise
LCA	Locally Competitive Algorithm
LTC	Local Thresholding Circuit
MAE	Mean Absolute Error
MGBv	ventral division of the Medial Geniculate Body
MID	Maximally Informative Dimension
ML	Modulated-limited
MPS	Modulation Power Spectrum
MTF	Modulated Transfer Function
PCA	Principal Components Analysis
PT	Pure Tones
RF	Receptive Field
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level
STA	Spike-Triggered Average
STRF	Spectro-Temporal Receptive Field
TM	Temporally-Modulated

TORC	Temporally-Orthogonal Ripple Combination
V1	Primary Visual Cortex
WNB	White Noise Bursts
WWN	Waveform White Noise