Sparse Representations for the Cocktail Party Problem

Hiroki Asari* Barak A. Pearlmutter† Anthony M. Zador‡§

June 7, 2006
(CVS: hrtf_source.tex 1.326)

Abstract

A striking feature of many sensory processing problems is that there appear to be many more neurons engaged in the internal representations of the signal than in its transduction. For example, humans have about 30,000 cochlear neurons, but at least a thousand times as many neurons in the auditory cortex. Such apparently redundant internal representations have sometimes been proposed as necessary to overcome neuronal noise. We instead posit that they directly subserve computations of interest. Here we provide an example of how sparse overcomplete linear representations can directly solve difficult acoustic signal processing problems, using as an example monaural source separation using solely the cues provided by the differential filtering imposed on a source by its path from its origin to the cochlea (the head-related transfer function, or HRTF). In contrast to much previous work, the HRTF is used here to separate auditory streams rather than to localize them in space. The experimentally testable predictions that arise from this model—including a novel method for estimating a neuron's optimal stimulus using data from a multi-neuron recording experiment—are generic, and apply to a wide range of sensory computations.

*Cold Spring Harbor Laboratory, Watson School of Biological Sciences, One Bungtown Road, Cold Spring Harbor, NY 11724, USA, asari@cshl.edu.

†Hamilton Institute, National Univ. Ireland Maynooth, Co. Kildare, Ireland, barak@cs.nuim.ie.

‡Cold Spring Harbor Laboratory, One Bungtown Road, Cold Spring Harbor, NY 11724, USA, zador@cshl.edu.

§**Corresponding author.**

1 Introduction

Animals in nature confront an acoustic environment consisting of sounds from a rich, indeed often bewildering, combination of sources. Survival depends on responding appropriately to potential threats, food sources and mates (*e.g.* at a cocktail party), while at the same time ignoring the many irrelevant sound sources that may constitute the majority of the acoustic energy received. Source separation, or "stream segregation," is therefore one of the central problems in acoustic processing that organisms must solve. Animals must confront many of the same challenges in solving this problem as do artificial systems, and the insights gained from the one can be applied to the other. However, little is currently known about how animals solve this problem (but see Fishman et al., 2004; Micheyl et al., 2005), and no artificial system can solve it in a general setting.

Animals exploit a variety of binaural and monaural cues to separate acoustic sources (Bregman, 1990). For example, two tones occurring simultaneously are more likely to be grouped together perceptually—*i.e.* perceived as arising from the same source—than the same notes occurring sequentially. Such grouping makes sense under the assumption that the auditory system is trying to discover the statistically independent causes of the acoustic signals received at the ears (Bell and Sejnowski, 1995, 1997; Lewicki and Sejnowski, 2000; Simoncelli and Olshausen, 2001); simultaneous onset of two tones is unlikely to arise purely by chance, so it is more parsimonious to assume that the tones were caused by a single source (*e.g.* as harmonics of a single fundamental frequency.)

Many of the spectral, temporal and spatial cues used for stream segregation can be interpreted in this context.

A striking feature of this and many other sensory processing problems is that there appear to be many more neurons engaged in the internal representations of the signal than in its transduction. For example, humans have only about 30,000 cochlear neurons, but at least a thousand times as many neurons in the auditory cortex. Although such apparently redundant internal representations have sometimes been proposed as necessary to overcome neuronal noise, here we posit that they contribute to computation.

In order to extract the behaviorally relevant information embedded in natural acoustic environments, animals must be able to separate auditory streams originating from distinct acoustic sources ("cocktail party problem"). The auditory cortex has orders of magnitude more neurons than the cochlea, so many different patterns of cortical activity may faithfully represent any given pattern of cochlear activity. We propose that the cortex exploits this excess "representational bandwidth" (Deweese et al., 2005), or the excess degrees of freedom, by selecting the sparsest representation within an overcomplete set of features. This model suggests how this excess representational bandwidth can be used for computation, instead of merely to overcome neuronal noise as is usually assumed. We illustrate this model by showing how sparseness can be used to separate sources perceived monaurally. The model makes testable predictions about the dynamic nature of representations in the auditory cortex. Our results support the idea that sparse representations may underlie efficient computations in the auditory cortex.

Our approach is to adopt a practical computational framework for the cocktail party problem, and then explore the testable implications that follow. Here we describe a model of how the auditory system can exploit one particular sort of monaural segregation cue, namely the spectral cues introduced by the differential filtering imposed by the head-related transfer function (HRTF). Note that in contrast to much previous work, the HRTF is used here to separate auditory streams rather than to localize them in space; our model assumes that the locations of the sources has already been determined by other mechanisms.

The model posits that the neural representation of an acoustic stimulus is overcomplete in the sense that there are many more neurons available than are needed to represent the stimulus with high fidelity (Olshausen and Field, 1997; Lee et al., 1999; Lewicki and Sejnowski, 2000; Zibulevsky and Pearlmutter, 2001). Because the representation is overcomplete, there are many patterns of neural activity that all faithfully encode any given stimulus. We show that constraining neural activity to be sparse selects one of these representations, and that the resulting pattern of neural activity solves the source separation problem, even when multiple sources are audible to only a single ear. The framework is quite general, and can serve as a starting point for understanding how cortical circuits might exploit other sensory cues as well.

2 Methods

All programming was done in Matlab.

2.1 HRTF

The head related transfer function (HRTF) is the filter imposed by the head and the detailed shape of the ear on sounds received at the cochlea. The HRTF depends on the spatial position—both the relative azimuth and elevation—of the source (Yost et al., 1996). At some frequencies, the HRTF can attenuate sound from one location by as much as 40 dB more than from another (supplementary Figure 1A).

Although every individual has his or her own HRTF, the basic characteristics of HRTFs are similar across individuals. We used a representative left human pinna HRTF downloaded from http://www.itakura.nuee.nagoya-u.ac.jp/HRTF/ (Nishino et al., 2001).

2.2 Spectral basis for sources *via* NMF

We tested our algorithm on mixtures of musical sources. We used non-negative matrix factorization (NMF) to obtain basis elements for each source.

NMF is an algorithm for factorizing a data matrix under an elementwise non-negativity constraint (Lee and Seung, 1999). The original data matrix is given as an $n \times m$ matrix $\mathbf{V}$, each column of which here contains the n data values for one of m spectrogram segments. The data matrix $\mathbf{V}$ is approximated by NMF as $\mathbf{V} \approx \mathbf{QH}$ where the dimension of the factors $\mathbf{Q}$ and $\mathbf{H}$ are $n \times r$ and $r \times m$, respectively. The rank of factorization, r, is chosen so $nr + rm < nm$, so as to compress the original nm elements in the matrix $\mathbf{V}$ into a smaller number of elements, nr in $\mathbf{Q}$ plus rm in $\mathbf{H}$. Each column of $\mathbf{Q}$ contains one of the basis spectrograms, and the matrix $\mathbf{H}$ represents the coefficients for reconstructing the columns of the original data matrix $\mathbf{V}$ in this basis. The $\mathbf{Q}$ matrices obtained by NMF for each individual source were concatenated to form an overcomplete source-space basis matrix $\tilde{\mathbf{D}} = \left[\, \mathbf{Q}_1 \mid \mathbf{Q}_2 \mid \cdots \,\right]$. Each column of $\tilde{\mathbf{D}}$ was then filtered through each HRTF and the results concatenated to form the feature matrix, $\mathbf{D} = \left[\, h_1(t) * \tilde{\mathbf{D}} \mid \cdots \mid h_N(t) * \tilde{\mathbf{D}} \,\right]$.

In our experiments, the spectrograms in the data matrix $\mathbf{V}$ were obtained from music sounds, natural sounds, or speech sounds: commercial audio CDs (instrumental solos, classical and jazz, one each on cello, clarinet, trumpet, harp, and harpsichord, for a total of five), the audio CDs *The Diversity of Animal Sounds* and *Sounds of Neotropical Rainforest Mammals* (Cornell Laboratory of Ornithology, Ithaca, NY, USA), and spoken poetry (Dylan Thomas, T. S. Eliot, Frank O'Hara and William Butler Yeats on the commercial audio CD *Poetry speaks: Hear great poets read their work from Tennyson to Plath*, Sourcebooks Inc., 2001, ISBN 1570717206), respectively. Samples of 100–150 s were taken, stereo channels averaged, and the signal down-sampled from the original 44.1 kHz to 8 kHz. Log-scaled spectrograms were generated using a custom Matlab routine (available upon request) with a bin size of 5 ms and 75 frequency bands ranging from 55–3,951 Hz in steps of 1/12 octave. Each column of $\mathbf{V}$ held a strip of spectrogram, yielding a dimensionality of $n = 75$, and $m = 5{,}000$ samples were used for the training. Note that the training samples were distinct from those used for the testing. Specifically, we used 10,000 samples to assess the representational sparseness achieved by the NMF basis (Figure 2), and 20,000 random combinations of three sources (using the 10,000 samples in Figure 2) to assess separation performance (Figure 4).

Each NMF run consisted of 500 iterations with 10 restarts from random initial conditions, with the restart that yielded the minimum total error chosen. The factorization rank was $r = 15$. Concatenating the five $\mathbf{Q}$ matrices for the five instruments yielded a dictionary of 75 basis elements, each of which was filtered by each of three different HRTFs, resulting in a feature matrix $\mathbf{D}$ with 225 columns. The source locations were randomly chosen but 90° apart from each other in the simulations (*e.g.*in Figure 3 the three sources were located on your left, center and right, corresponding to the HRTFs for azimuth −90°, 0°and 90°respectively, with zero elevation; see also supplementary Figure 1B). The analyses on the natural sound and speech sound were performed in a similar manner, with 5,000 training samples for each data matrix $\mathbf{V}$.

2.3 Minimization

Pseudoinverses (L_2-norm minimization) were computed with Matlab's `pinv` routine, which uses an algorithm based on singular value decomposition (SVD). The linear programming problem (Eq. 10) was solved using the Matlab *Optimization Toolbox* `linprog` routine. We did not impose a non-negativity constraint on the coefficients. As a result, the dense solution consists of negative coefficients as well as positive ones, whereas all the substantially non-zero elements are positive for the sparse solution. Figure 6 shows the absolute values of the dense solution coefficients.

The linear programming problem given in Eq. 10 can be sensitive to noise. We therefore solved an augmented version of this that included a noise model. In particular, we assumed that the total amount of noise was bounded. Thus Eq. 10 was reformulated as:

$$\underset{\mathbf{c}}{\text{minimize}}\ \|\mathbf{c}\|_1 \text{ subject to } \|\mathbf{Dc} - \mathbf{y}\|_p \leq \beta \qquad (1)$$

where β is proportional to the noise level and with $p = 1$, 2, or ∞. Letting $\beta \to 0$ is equivalent to assuming that the noise is very small, and the solution converges to the zero-noise solution, Eq. 10.

The Gaussian noise case, $p = 2$, can be solved by semidefinite programming methods (Fletcher, 1985). Both $p = 1$ and $p = \infty$ can be solved using linear programming. All approaches yield qualitatively similar results.

The solutions presented here all used $p = 1$. For this case, noise vectors $\mathbf{e}^+$ and $\mathbf{e}^-$ are introduced and included in the optimization, allowing Eq. 1 to be rewritten in standard form:

$$\begin{gathered}\mathbf{c}^+, \mathbf{c}^-, \mathbf{e}^+, \mathbf{e}^- \geq 0 \\ \mathbf{D}\mathbf{c}^+ - \mathbf{D}\mathbf{c}^- + \mathbf{e}^+ - \mathbf{e}^- = \mathbf{y} \\ \begin{bmatrix}1 & \cdots & 1\end{bmatrix}\mathbf{e}^+ + \begin{bmatrix}1 & \cdots & 1\end{bmatrix}\mathbf{e}^- \leq \beta\end{gathered} \tag{2}$$

We typically examined four different noise levels ($\log_{10}\|\mathbf{y}\|_1/\beta = 1, 2, 3, 4$), and selected the one with the best separation performance on average as the result (Figures 4 and 5).

SNRs were calculated as the reciprocal of the average across sources of $\langle (x_i(t) - \hat{x}_i(t))^2 \rangle / \langle x_i(t)^2 \rangle$ where $x_i(t)$ is the original spectrogram of the i^{th} source, $\hat{x}_i(t)$ is its estimate when recovered from the mixture, and the average $\langle \cdot \rangle$ is over time.

To measure the degree of sparse representations by NMF basis elements (Figure 2) and its relation to the separation performance (Figure 4), we introduced a *sparseness index*, defined as the number of non-zero elements in the presence of a single source divided by the dimension size. This index is unity for a dense representation, and approaches zero as the representation becomes sparser. The noise level was $\log_{10}\|\mathbf{y}\|_1/\beta = 1$ in Figure 2B, resulting in the reconstruction SNR of 18.3±3.8, 16.0±3.0, 18.0±3.6 (median ± interquartile range in dB) for music, natural sound, and speech ensembles, respectively.

2.4 Estimation of linear encoders and decoders

Given a set of stimuli $\mathbf{y}_k$ (for $k = 1, 2, \ldots$) and the corresponding responses $\mathbf{c}_k$ generated using Eq. 10, the optimal linear decoding filter $\hat{\mathbf{D}}$ was estimated by solving the following regression problem:

$$\underset{\hat{\mathbf{D}}}{\text{minimize}} \sum_k \left\| \mathbf{y}_k - \hat{\mathbf{D}}\mathbf{c}_k \right\|_2^2,$$

where $\|\cdot\|_2$ denotes the L_2 (Euclidean) norm. Similarly, the optimal linear encoder $\hat{\mathbf{E}}$ was obtained by solving the following equation:

$$\underset{\hat{\mathbf{E}}}{\text{minimize}} \sum_k \left\| \hat{\mathbf{E}}\mathbf{y}_k - \mathbf{c}_k \right\|_2^2.$$

Note that we used a fraction of the elements in $\mathbf{c}_k$ for the linear filter estimation, and showed the average results in Figures 7 and 8 over 200 random samplings of neurons. Also note that the i^{th} column of $\hat{\mathbf{D}}$ and the i^{th} row of $\hat{\mathbf{E}}$ correspond to the optimal linear decoder and encoder for the i^{th} neuron, respectively.

In Figure 7, we used a $1,168 \times 3,600$ feature matrix $\mathbf{D}$, each column of which held a feature spanning over 16 time bins (96 ms), with a bin size of 6 ms and 73 frequency bands ranging between 55–3,520 Hz in steps of 1/12 octave. As the original feature of a target neuron, we chose the one obtained from cello ensembles, and thus we used cello sounds as input stimuli in the simulation.

2.5 Asymmetry of sparse representations

To illustrate the asymmetry of linear encoding and decoding in the framework of our model, we ran simulations in 25 dimensions with 75 neurons. In the simulations, the 3-fold overcomplete features (a 25×75 feature matrix $\mathbf{D}$) were first generated randomly on the unit hypersphere. Neural activities for sample stimuli drawn from a Gaussian distribution were then determined by Eq. 10.

For simulated single unit data (Figure 8A), we computed the mutual information $I(c, s)$ between the simulated neural responses c and stimulus s using the $I(c, s) = H(c) - H(c|s) = H(c)$, where $H(c)$ is the response entropy and $H(c|s)$, the conditional of the response given the stimulus, is zero because the relation between stimuli and responses was deterministic. Thus the mutual information between the single neuron and the stimulus was just equal to the response entropy, which we estimated by direct binning from the histogram of neural responses. We compared this information to either: the mutual information between the optimal linear estimate of the response given the stimulus and the actual stimulus (encoding); or between the optimal linear estimate of the stimulus given the response and the actual response (decoding). For these information estimations we used the Gaussian approximation to bound the entropy of the reconstruction error (Bialek et al., 1991). We then normalized these linear information estimates to full mutual information to obtain the *reconstruction quality*,

$$1 - \left\langle \frac{\text{linear estimate of mutual information}}{\text{mutual information}} \right\rangle .$$

For multi-unit data (Figure 8B), the computation of the full mutual information (rather than the linear approximation) was computationally intractable. We therefore computed the following simpler measure of the reconstruction quality of the models:

$$1 - \left\langle \frac{\|\text{reconstruction error}\|_2}{\|\text{response or signal}\|_2} \right\rangle ,$$

where $\|\cdot\|_2$ denotes the L_2 norm and $\langle\cdot\rangle$ the mean over data. Note that the measure is based on the relative length of the model errors, and that it gives zero for pure noise and one for perfect reconstruction.

2.6 Context dependence of STRFs

In Figure 10, for demonstration purposes, we used a $1,168 \times 3,600$ feature matrix $\mathbf{D}$ (the same one as in Figure 7) and used two different sets of 300 active features (*i.e.* $1,168 \times 300$ packed matrices $\mathbf{D}_k$; see Appendix A.3) to estimate the STRFs for the two different contexts. Note that some of the features were active in both contexts (including the one shown in Figure 10A) whereas others only in either context.

3 Results

Our main goals are to explore a model of computation with sparse representations, and to generate new experimentally testable predictions from this model. To make our model concrete, we consider a specific computation—a special case of the monaural cocktail party problem in which the head-related transfer function (HRTF) provides the critical cue for disentangling sources. We focus on this special case *not* because it is of central importance from a psychophysical perspective—in a general setting, the HRTF is typically just one of many cues, and often not the most important—but rather because this problem provides a convenient way to illustrate the key predictions. The same sparse framework can be generalized to exploit other cues for source separation, and to other sensory processing problems (*e.g.* vision) as well.

The presentation is organized as follows. First we define the particular source separation problem we consider, in which there are several sources and a single ear. Next we show how a sparse overcomplete representation, like that seen at the cortical level in the auditory system, can be used to separate the sources. Finally, we identify experimentally testable predictions of the model.

3.1 Problem formulation

We have all experienced the basic cocktail party problem as a part of everyday life: we stand in a room full of people chatting, chairs scraping, fans humming and so forth, and strain to understand the words of a single interlocutor. This familiar but challenging scenario is interesting precisely because it tests the limits of what we humans can achieve. The cocktail party is, however, just an extreme example of a more general problem that the auditory system constantly confronts. It is rare that we can listen to an acoustic source without interference from other sources, yet our auditory system filters the interfering sources out of our conscious perception so effectively that we are often almost unaware of them. The apparent effortlessness with which we solve the cocktail party problem is deceptive, and is a testament to the effectiveness of our auditory system. Indeed, the problem of background noise represents one of the main factors limiting the widespread practical adoption of artificial speech recognition systems.

The auditory system uses a wide variety of psychophysical cues to segregate auditory streams (Bregman, 1990), including both binaural and monaural cues. Many monaural cues have been identified, such as common onset time or comodulation of stimulus power in different parts of the spectrum.

For simplicity we focus here on just one set of cues: those provided by the differential filtering imposed on a source by its path from its origin in space to the cochlea. This filtering is caused both by the head and the detailed shape of the ear (the head-related transfer function, or HRTF), and by the environment on sources at different positions in space (Yost et al., 1996). The HRTF is important for generating a three-dimensional experience of sound, so that acoustic sources that bypass the HRTF (*e.g.* those presented with headphones) are typically perceived unnaturally, as though arising inside the head (Wightman and Kistler, 1989; Kulkarni and Colburn, 1998). Whereas the importance of the HRTF in sound *localization* has been studied extensively (Knudsen and Konishi, 1979; Wightman and Kistler, 1989; Wenzel et al., 1993; Hofman and Opstal, 2002), its role in source *separation* as such has not. In contrast to much previous work, the HRTF is used here to separate auditory streams rather than to localize them in space.

It is often reasonable to assume that sound arriving from different locations should be treated as arising from distinct sources. For the purposes of the present paper, all sounds from a given position are *defined* to belong to the same source, and any sounds from a different position are defined to belong to different sources. We emphasize that although sound localization (the process by which an animal determines where in space a source is located) is related to source separation (the process by which an animal extracts different auditory streams from a single waveform), the two computations are distinct; neither is necessary nor sufficient for the other. Here we focus on the separation problem, and assume that source localization occurs by other mechanisms.

The particular source separation problem we consider is as follows. Suppose there are N acoustic sources located at known distinct positions in space, with $x_i(t)$ being the time course of the stimulus sound pressure of the i^{th} source at its point of origin. Associated with each position is a known filter given by $h_i(t)$. In what follows we will refer to $h_i(t)$ as the HRTF, but in general $h_i(t)$ will include not just the filtering of the head and external ear, but also the filter function of the acoustic environment (reverberation, *etc.*)

The signal $y(t)$ at the ear is the sum of the filtered signals,

$$y(t) = \sum_{i=1}^{N} h_i(t) * x_i(t) = \sum_{i=1}^{N} \tilde{x}_i(t) \tag{3}$$

where $*$ indicates convolution and $\tilde{x}_i(t) = h_i(t) * x_i(t)$ is the i^{th} source in isolation following filtering. (We can say that $x_i(t)$ is the i^{th} source measured in source space, while $\tilde{x}_i(t)$ is the same source measured in sensor space.) The organism's goal in source separation is to recover the underlying sources $x_i(t)$ from the signal $y(t)$, using knowledge of the directional filters $h_i(t)$. For example, if $x_{\text{alice}}(t)$ and $x_{\text{bob}}(t)$ are speech streams generated by two speakers (sources) Alice and Bob at a cocktail party, then the goal is to disentangle these two streams using the only signal available, the sum $y(t) = h_{\text{alice}}(t) * x_{\text{alice}}(t) + h_{\text{bob}}(t) * x_{\text{bob}}(t)$. Note that the actual spatial locations of the sources are not computed during the separation; we do not address the localization problem in this paper.

The particular monaural version of this problem that we consider here is a special—more difficult—case of the binaural (or, in artificial systems, the multiple microphone) problem.

3.2 Neural representation for source separation

How might a neural system solve the source separation problem described above? We begin by assuming that each short segment (*e.g.* 5 ms) of each acoustic source $\tilde{x}_i(t)$ (as it sounds at the cochlea) is represented in the activities c_{ij} of a population of neurons indexed by components j and source positions i,

$$\tilde{x}_i(t) = \sum_j c_{ij}\, d_{ij}(t), \tag{4}$$

where $d_{ij}(t)$ are stimulus *features*, *i.e.* elements of a (not necessarily orthogonal, and possibly overcomplete) linear basis. We will interpret the neural activities c_{ij} as the spike rate of the corresponding neurons during each segment. The signal $y(t)$ is then given by

$$y(t) = \sum_{ij} c_{ij}\, d_{ij}(t). \tag{5}$$

We have introduced Eq. 5 as an analytic model: given a stimulus $y(t)$, find a set of features $d_{ij}(t)$ and neural activities c_{ij} that represent that stimulus; if the features span the stimulus space, then such a representation will always exist. Below we will focus on the case where the feature set permits a sparse representation, *i.e.* where only a few of the neural activities c_{ij} are significantly nonzero. (Although *sparse* might colloquially refer to the case where most of the activities c_{ij} are exactly zero, here we use a generalized notion of sparseness, common in the literature, that requires only that most activities be close to zero.)

Eq. 5 assumes a linear relationship between an auditory stimulus $y(t)$ and its neural representation in terms of features $d_{ij}(t)$. The assumption of linearity is common in both visual and auditory physiology. For example, it is often assumed that a population of neurons in cortical area V1 represents a visual scene in terms of a collection of oriented edges; in this case the scene and the features in Eq. 5 would be rewritten as functions of spatial rather than temporal coordinates, but the formulation would be otherwise identical. Similarly, in auditory physiology, stimuli are sometimes represented as a weighted sum of basis elements such as moving ripples (Kowalski et al., 1996; Klein et al., 2000); in the context of Eq. 5, this implies assuming a one-to-one correspondence between a basis element (derived from the ripple basis) $d_{ij}(t)$ and the firing rate c_{ij} of a corresponding neuron.

In order to relate the neural representation of the signal $y(t)$ in Eq. 5 to the sources $x_i(t)$, we further assume that each source can be expressed as a linear combination of (not necessarily orthogonal) basis elements $q_j(t)$,

$$x_i(t) = \sum_j c_{ij}\, q_j(t), \tag{6}$$

where the basis elements $q_j(t)$ are related to the features d_{ij} by convolution with each filter $h_i(t)$,

$$d_{ij}(t) = h_i(t) * q_j(t). \tag{7}$$

Combining these expressions, the signal $y(t)$ received at the ear is related to the sum of the filtered sources by

$$\begin{aligned} y(t) &= \sum_i h_i(t) * x_i(t) \\ &= \sum_i h_i(t) * \Big(\sum_j c_{ij}\, q_j(t)\Big) \\ &= \sum_{ij} c_{ij} \big(h_i(t) * q_j(t)\big) \\ &= \sum_{ij} c_{ij}\, d_{ij}(t). \end{aligned} \tag{8}$$

There are thus more features $d_{ij}(t)$ in the neural representation than there are basis elements $q_j(t)$. In particular, if there are known to be N sources, then there are N-fold more features $d_{ij}(t)$ than basis elements $q_j(t)$. As before, Eqs. 5–7 represent an analytic model: given a set of features $d_{ij}(t)$ (or equivalently a set of basis elements $q_j(t)$ and position-dependent filters $h_i(t)$), and an input $y(t)$, find an appropriate set of neural activities c_{ij}.

The basis elements $q_j(t)$ reflect statistical correlations within sources; each source typically consists of several such elements. These basis elements can be thought of as an internal model of the components of acoustic sources, in the same way that edges might be thought of as components of visual sources (objects). Because the neural representation involves pre-filtering with the HRTF (Eq. 7), the coefficient c_{ij} associated with feature $d_{ij}(t)$ is then better thought of as representing the hypothesis that an element $q_j(t)$ is present at position i. In the same way, neurons in the primary visual cortex can be thought of as representing the hypothesis (d_{ij}) that an oriented edge (q_j) is present at a particular position (i) in the visual field. In other words, the elements $q_j(t)$ reflect only the properties of the stimulus, whereas the features $d_{ij}(t)$ arise from interaction of these elements with the sense organs.

A population of neural activities satisfying Eqs. 5–7 has effectively solved the source separation problem, since a given source i can be reconstructed merely by summing over all neurons associated with position i. This formulation therefore recasts the source separation into a new problem: finding the appropriate neural activities c_{ij}. Such a representation, if it could be found, is especially appealing because it permits the sources to be reconstructed (by Eq. 6) directly in terms of the stimulus elements $q_j(t)$ as they sound at the source (*i.e.* prior to filtering); the reconstruction is therefore invariant to changes in stimulus position. In the next section we will show that sparseness provides the key to specifying the appropriate representation.

For notational and computational convenience we discretize time and rewrite Eq. 5 in matrix form (using **bold** to indicate vectors and matrices):

$$\mathbf{y} = \mathbf{Dc}, \tag{9}$$

where $\mathbf{y}$ is a column vector whose N_{row} elements correspond to the discrete-time sampled elements $y(t_k)$, $\mathbf{c}$ is a column vector of length N_{col} representing the complete neural activity pattern c_{ij}, and $\mathbf{D}$ is an $N_{\text{row}} \times N_{\text{col}}$ matrix whose columns $\mathbf{d}_{ij}$ hold the features with elements $d_{ij}(t_k)$.

3.2.1 Sparse neural representation of sources

Source separation thus requires finding the neural activities $\mathbf{c}$ such that the neural representation represents the sources $\mathbf{x}_i$ as closely as possible. We assume that the neural representation is overcomplete (Riesenhuber and Poggio, 2000; Olshausen and Field, 1997), *i.e.* that the number of neurons (features) is large ($N_{\text{col}} > N_{\text{row}}$). In this case, many different neural activity patterns $\mathbf{c}$ could represent the stimulus $\mathbf{y}$ equally well (Figure 1A). However, the goal is not merely to represent the stimulus $\mathbf{y}$, but to find a representation in which the underlying sources $\mathbf{x}_i$ are apparent and from which they can be readily recovered.

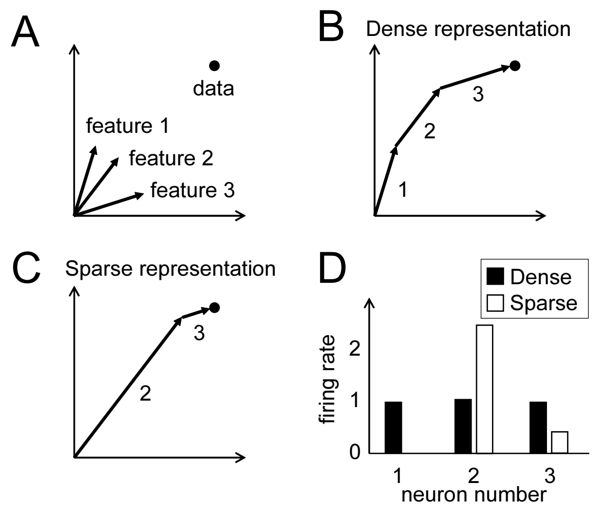

Figure 1: **Overcomplete representation in two dimensions.** **(A)** Three non-orthogonal feature vectors $\mathbf{d}_{ij}$ in $N = 2$ dimensions constitute an overcomplete representation, offering many possible ways to represent a data point $\mathbf{y}$ with no error. **(B)** The conventional solution is given by the pseudoinverse, which yields a dense representation because it minimizes the squared sum of the neural activity, $\sum_{ij} c_{ij}^2$. This representation invokes all features about evenly. **(C)** The sparse solution invokes at most $N = 2$ features because it minimizes $\sum_{ij} |c_{ij}|$. **(D)** Comparison of neural activity for the two cases. For the dense representation, all three neurons participate about equally, whereas for the sparse representation activity is concentrated in neuron 2.

Since the neuronal population does not have access to the sources themselves, but only to their sum $\mathbf{y}$, not enough information is available to recover the sources uniquely. The source separation problem is thus ill-posed. (In the same way, knowing that the sum of two scalars a and b is 12 is not sufficient to recover a and b, and any choice for a and b that satisfies $a + b = 12$ is a possible solution.) The problem can be made well-posed by adding additional constraints (regularizers) on the responses, as is often done in computational vision (Poggio et al., 1985). Here we consider a sparseness regularizer on the neural representation (Chen et al., 1998; Lee et al., 1999; Lewicki and Sejnowski, 2000; Zibulevsky and Pearlmutter, 2001; Vinje and Gallant, 2000; Simoncelli and Olshausen, 2001; Hahnloser et al., 2002; Bell and Sejnowski, 1997; Olshausen and Field, 1996, 1997; Olshausen and O'Connor, 2002). In neural terms, this sparseness assumption corresponds to representing the acoustic stimulus $\mathbf{y}$ in terms of the *minimum number of spikes* (Figure 1C), a biologically appealing constraint which leads to an energy-efficient representation (Laughlin and Sejnowski, 2003; Levy and Baxter, 1996). Thus we assume that the neural representation $\mathbf{c}$ satisfies (see Appendix A.1):

$$\text{minimize } \sum_{ij} |c_{ij}| \text{ subject to } \mathbf{y} = \mathbf{D}\mathbf{c}. \quad (10)$$

Eq. 10 specifies a linear programming problem with a single global optimum. Formally, the solution minimizes the L_1 norm $\|\mathbf{c}\|_1 = \sum_{ij} |c_{ij}|$ of the solution vector. In practice, the problem we consider allows for reconstruction noise (see Eq. 1 in Method 2.3).

3.2.2 Dense neural representation of sources

An alternative regularizer is that implicit in the pseudoinverse (Strang, 1988), corresponding to the usual least-squares solution (see Appendix A.1),

$$\text{minimize } \sum_{ij} c_{ij}^2 \text{ subject to } \mathbf{y} = \mathbf{D}\mathbf{c}. \quad (11)$$

The pseudoinverse finds the solution $\mathbf{c}$ that minimizes the L_2 norm, *i.e.* the squared neural activity $\sum_{ij} c_{ij}^2$ (Figure 1B). However, it is not obvious why it would be useful for the brain to minimize this quantity, which has units of spikes-squared, rather than some other quantity (such as spikes; see below). Moreover, we show in the next section that it fails in practice to separate the sources successfully.

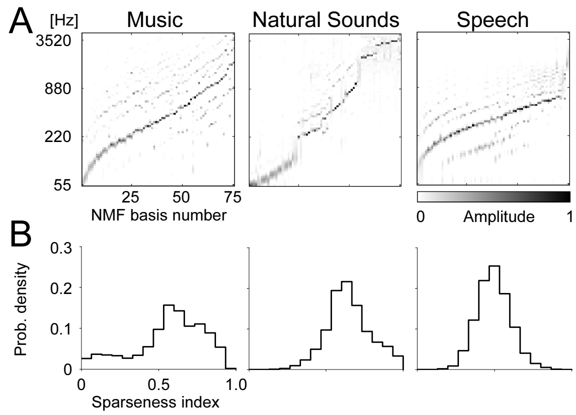

Figure 2: **Non-negative matrix factorization (NMF) can be used to find the parts of sound ensembles. (A)** NMF basis elements for three sound classes (music, natural sounds, and speech) were aligned in columns by the peak frequency. Note that power is concentrated in the fundamental frequency, but higher harmonics are clearly visible. Also note that each column, which reflects statistical correlations present in the sources, is an example of $q_j(t)$ defined in Eq. 6; it is the filtered versions $d_{ij}(t)$ that form the neural representation in Eq. 7. **(B)** The ability of the NMF bases in (A) to represent sounds in a sparse model is quantified in terms of the "sparseness index," defined as the number of non-zero elements in the presence of a single source divided by the dimension size. This index is unity for a dense representation, and approaches zero as the representation becomes sparser. The distribution of the "sparseness index" was 0.61 ± 0.27, 0.64 ± 0.17, and 0.49 ± 0.13 (median $\pm$ interquartile range) for music, natural sounds, and speech, respectively, over 10,000 test samples; *see Methods for details*.

3.3 Separation of harmonic sources

Successful source separation based on Eq. 10 requires that two conditions be satisfied. First, the sources must be sparsely representable, as is the case with natural auditory stimuli (Attias and Schreiner, 1997; Lewicki, 2002; Klein et al., 2003; Smith and Lewicki, 2006). Second, the sources must have spectral correlations matched to the HRTF. We found that the model was able to separate acoustic sources consisting of mixtures of music, natural sounds and speech.

3.3.1 Finding a feature set

We used nonnegative matrix factorization (NMF) to generate a set of basis features from spectrograms obtained from samples of solo instrumental music, natural sounds and speech (Figure 2). NMF is an algorithm for factorizing a data matrix—a matrix whose columns contain the snippets of solos—under non-negativity constraints (Lee and Seung, 1999). In contrast to some other decomposition approaches, such as principal component analysis (PCA), NMF often yields representations in which the elements are fairly local, and which can be interpreted as "parts."

When applied to music, NMF typically yielded elements suggestive of musical notes, each with a strong fundamental frequency and weaker harmonics at higher frequencies. In many cases, listeners could easily use timbre to identify the instrument from which a particular element was derived. When applied to sounds from other ensembles (natural sounds and speech), NMF yielded elements that had rich harmonic structure, but it was not in general easy to "interpret" the elements (*e.g.* as vowels). Nonetheless these elements captured aspects of the statistical structure of the underlying ensemble of sounds, and led to sparse representations of the ensembles (Figure 2B).

The choice of NMF in this context was merely a matter of convenience; we could have used any basis that captured the spectral correlations in the sources and permitted a sparse representation. Finding good overcomplete dictionaries from samples of a stimulus ensemble is a subject of ongoing research (Kreutz-Delgado et al., 2003). We do not imagine that NMF is the "algorithm" by which features are established in real neural circuits—such features must surely arise through a complex interaction of genetic and environmental cues. We need not, therefore, expect to find a precise correspondence between the features obtained by NMF and those observed in the auditory cortex. In this respect our results complement previous work on finding the features underlying auditory or visual scenes (Olshausen and Field, 1997, 1996; Lewicki, 2002; Bell and Sejnowski, 1997; Schwartz and Simoncelli, 2001); the emphasis here is not on the elements themselves, but rather on how they work together to form a representation that separates sources.

3.3.2 Separation

To test the model's ability to separate sources, we generated digital mixtures of three sources positioned at three distinct positions in space (Figure 3). On the *left column* are the spectrograms of the sources at their origin. Two of the sources (a harp playing the note "D", *center* and *bottom*) were chosen to be identical; this example is thus particularly challenging, since the only cue for separating the sources is the filtering imposed by the HRTF.

Separation was nevertheless quite successful (compare *left* and *right* columns). These results were typical: whenever the underlying assumptions about the sparseness of the stimulus were satisfied, sources consisting of mixtures of music, natural sounds or speech were all separated well (Figure 4). Separation worked particularly well for mixtures of sparsely representable sources (*i.e.* smaller sparseness index values), whereas it did not work for sources that were not sparsely represented (*i.e.* larger sparseness index values.) Figure 5 shows that separation without differential pre-filtering by the HRTF was unsuccessful, as was separation using the Gaussian prior instead of the sparseness prior (dense representation.)

The neural representations underlying separation provide insight into these results. Figure 6A shows the representations of each of the three sources (the same as in Figure 3) presented in isolation. In each panel, the activity in a population of 225 neurons (corresponding to the 225 features $\mathbf{d}_{ij} = \mathbf{h}_i * \mathbf{q}_j$) is indicated by the intensity of points on a 15×15 grid. Since the sources occupy three positions i, there are three copies of the basis $\mathbf{q}_j$ in each panel (corresponding to the three filters $\mathbf{h}_i$.) The activity patterns are sparse; only a relatively small number of units are active in each representation. Note that because the middle and the right sources (*source 2* and *3*, respectively) in this example were chosen to be identical, the middle and right neural representations differ only by a shift.

The procedure for recovering a source from such a representation is straightforward: the estimate of the left source (*source 1*) is simply the summed activity of the left third of the neurons—those representing features prefiltered by the HTRF corresponding to the leftmost position in space; and likewise for the middle and right thirds. The HRTF can thus be seen as a kind of "tag" for grouping together elements from a single source. This suggests dividing source separation into two conceptually distinct steps (although in practice the steps occur simultaneously.) In the first step, the stimuli are decomposed into the appropriate features. In the second step, the features are tagged and bundled together with other features from the same source. It is for this bundling step that the HRTF along with the prior knowledge of souce locations is essential.

The failure of the dense representation to separate sources (Figure 4) results from a failure of the first step. Instead of decomposing the sources into a small number of features, the dense representation (Figure 6C) assumes that each instrument contributed about equally to the received signal, and so finds a representation in which a large fraction of neurons are active. That is, instead of "explaining" the sources in terms two harps and a trumpet, the dense representations also finds some clarinet, some cello, *etc.*, at all positions. This is intrinsic to the dense solution, since it finds the "minimum power" solution in which neural activity is spread among the population (Figure 1B).

The failure of even the sparse approach when the spectral cues induced by the HRTF are absent (Figure 5, *leftmost point showing* 0*-degree separation*) results from a failure at the second step. That is, the sparse approach finds a useful decomposition at the first step even without the HRTF, but in the absence of HRTF cues the active features are not tagged, and so the features cannot be assigned appropriately to distinct sources. Other psychophysical cues relevant for source separation, such as common

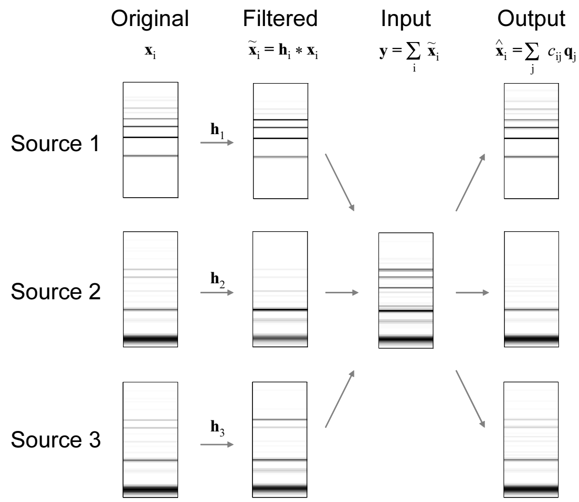

Figure 3: **Separation of three musical sources.** Three musical instruments at three distinct spatial locations were filtered (by $\mathbf{h}_1, \ldots, \mathbf{h}_3$, respectively) and summed to produce the *input* $\mathbf{y}$, and then separated using a sparse overcomplete representation to produce the *output*. Note that two of the sources (a harp playing the note "D", *center* and *bottom*) were chosen to be identical; this example is thus particularly challenging, since the only cue for separating the sources is the filtering imposed by the HRTF. Nevertheless, separation was good (compare *left* and *right* columns.)

onset time, might provide alternative or additional tags in this same framework. A more general formulation of source separation might allow tagging on longer time scales, so that a feature active at one moment might be more (or less) likely to be active the next, reflecting the fact that sources tend to persist, but we do not pursue that approach further here.

3.4 Experimental predictions

Our model of sparse representations makes at least three experimentally testable predictions.

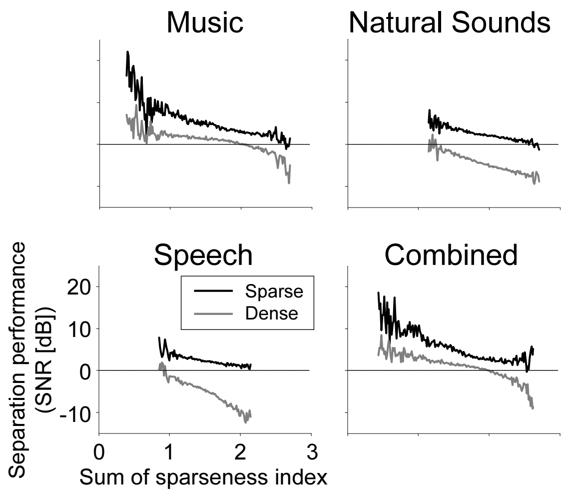

Figure 4: **Performance of different separation approaches with three sources.** The separation performance (SNR across sources) is shown as a function of the sum of the "sparseness index" of the three sources (average over 20,000 sample sets). Note that sparse prior (black) always outperforms dense prior (gray), and that excellent separation was achieved especially when the sources are sparsely representable. Also note that the model does not depend strongly on choosing the basis carefully, as demonstrated by the good performance of the "combined" example in which a concatenated basis was taken from all the ensembles.

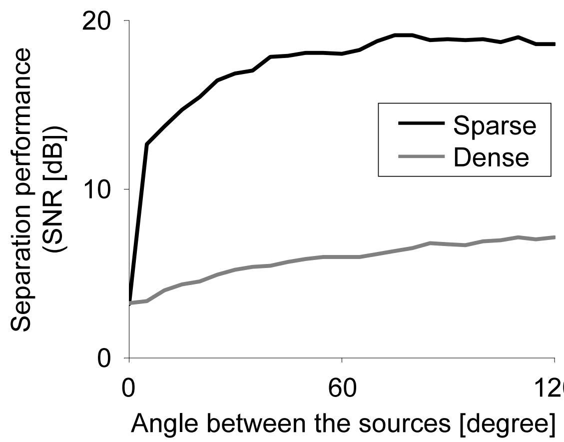

Figure 5: **Separation performance for different source locations.** Using a typical example of three novel stimuli (trumpet and two same harp), separation performance (y-axis) was examined with all the possible combinations of the three sources (from 0 to 120 degrees apart; x-axis). The average performance is shown here under either sparse (black) or dense (gray) prior. Note that separation was unsuccessful at angle zero since we cannot exploit *differential* filtering, whereas the performance gets better as the sources get further apart.

3.4.1 Optimal feature estimation requires multi-neuron recording

In this model, the firing rate of a given neuron $\{i\,j\}$ is maximized when there is a perfect match between the stimulus and that neuron's feature, *i.e.* when $\mathbf{y} = \mathbf{d}_{ij}$. Since the feature $\mathbf{d}_{ij}$ is used in the linear reconstruction of the stimulus from the neural activities (Eq. 5), one might imagine that the optimal stimulus (*i.e.* the stimulus that maximizes the firing rate) can be obtained by estimating the optimal linear decoder of the target neuron considered alone. Experiments based on this idea have shown that the optimal linear decoder can sometimes drive neurons in the auditory cortex to fire vigorously (deCharms et al., 1998).

Surprisingly, this model predicts that the linear estimate of the decoder obtained in this way is *not* the optimal stimulus, even though the optimal decoder is linear. Instead, finding the optimal stimulus requires recording from *all* the neurons involved in the representation. This follows from the fact that we have assumed that the features are not orthogonal (see also Appendix A.2). Note that in this model, optimal decoding (Eq. 5) need not take neural correlations into account, even when they are present.

This first prediction is illustrated by a simulation (Figure 7). The y-axis shows the firing rate of a target neuron (normalized to its maximum firing rate) in response to the presentation of the stimulus that matches the optimal linear decoder constructed by recording the activity of a target neuron and a variable number of other neurons. When the optimal linear decoder is estimated from only the target neuron, the firing rate is sub-maximal. As the number of neurons used to estimate the optimal linear

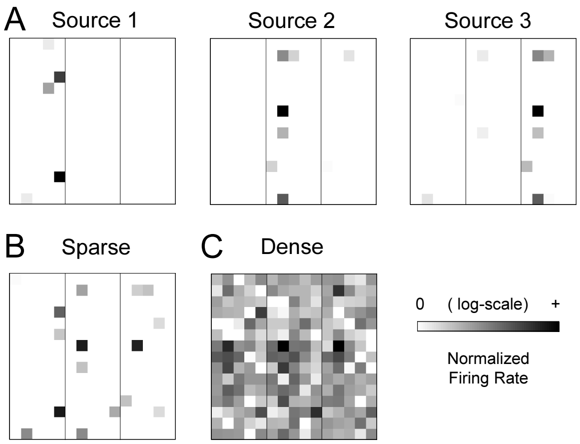

Figure 6: **Neural representations underlying source separation.** Each panel shows the activity of a population of 225 neurons, corresponding to the 225 features $\mathbf{d}_{ij} = \mathbf{h}_i * \mathbf{q}_j$. The intensity of each dot in the 15×15 grid is proportional to the log of the firing rate of each neuron. Since the sources occupy three positions i, there three copies of the basis $\mathbf{q}_j$ in each panel (corresponding to the three filters $\mathbf{h}_i$.) The copies are arranged from left to right for convenience, and separated by vertical lines. However, the arrangement is for purposes of illustration only; we do not mean to imply any spatial organization of sources within the cortex. The sources are the same as in the previous figure. **(A)** Sparse representations of the three sources (corresponding to the *original* spectrograms in Figure 3) presented in isolation. Only a relatively small number of units are active in each panel. **(B)** Sparse representation of the mixed sources (*input* spectrogram in Figure 3.) Note that activity is approximately the sum of the activities of the isolated sources in (A). **(C)** Dense representation of the mixed sources. Note that most units are active.

decoder is increased (x-axis), the response of the target neuron converges to unity, indicating that the optimal decoder has converged to the target neuron's feature.

Figure 7 represents a novel and testable prediction of the model: jointly estimating the optimal linear decoder from a population of neurons should yield a stimulus that is closer to op-

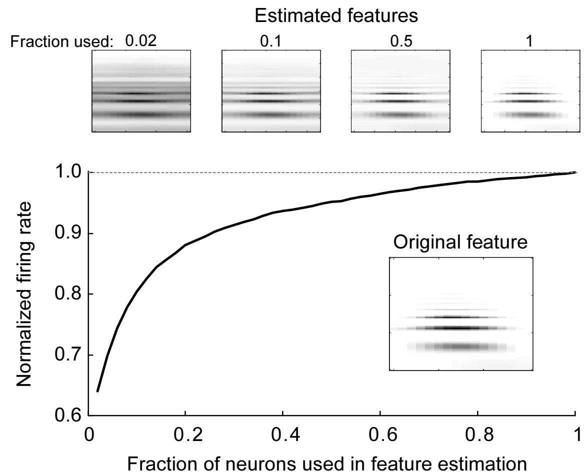

Figure 7: **Prediction 1: Stimulus optimization requires multi-neuron recording.** The y-axis shows the simulated firing rate of a target neuron (normalized to its maximum firing rate) in response to the presentation of the optimal linear decoder constructed by recording the activity of a target neuron and a variable number of other neurons. When the optimal linear decoder is estimated from only the target neuron, the firing rate is sub-maximal. As the number of neurons used in this simulation to estimate the optimal linear decoder is increased (x-axis), the response of the target neuron converges to unity, indicating that the optimal decoder has converged to the target neuron's feature.

timal. Moreover, it also leads to a novel experimental approach for finding the optimal stimulus. Note that although in principle the activity of all neurons involved in the representation must be recorded, in practice the activity of even a few can be useful. With modern techniques (*e.g.* tetrodes) for isolating the activity of several nearby neurons, this approach might be practical.

3.4.2 Linear decoding and nonlinear encoding

A second testable prediction of the model is that there should be an asymmetry between encoding and decoding: the optimal encoding function is nonlinear but the optimal decoding function is linear. Here *decoding* refers to the process of "reading out" a neural representation (*e.g.* by forming an estimate or reconstruction of the stimulus), whereas *encoding* refers to the process by which the nervous system constructs a pattern of neural activities from a stimulus. Surprisingly, however, this asymmetry emerges only for populations of neurons; the optimal linear encoder and decoder of an isolated neuron perform about equally, and both underperform the optimal nonlinear decoder (Figure 8).

The fact that optimal decoding of a neuronal population is linear—*i.e.* that the optimal linear decoder of the neuronal population response provides perfect reconstruction of the stimulus under the model, so no nonlinear model can do better—is a direct consequence of our fundamental assumption (Eq. 5) that the neural representation is a linear combination of features. The linearity of neural *de*coding does not imply that the neural *en*coding function—the inverse transformation from the stimulus to the response—need be linear; and in general it is not.

Sparseness induces a nonlinear encoding function; more precisely, it induces a *piecewise linear* encoding function (Figure 9). Sparseness implies that only at most N_{row} out of the possible N_{col} features $\mathbf{d}_{ij}$ are active in the representation of a particular stimulus; the precise subset of active neurons changes for different stimuli. Piecewise linearity arises because the encoding function is linear for all stimuli that activate the same subset of features, but changes for different subsets (see also Appendix A.2). Note that not just any nonlinear function can be implemented. For example, any saturating nonlinearities must be introduced by a preprocessor, since doubling the stimulus $\mathbf{y}$ necessarily doubles the neural representation $\mathbf{c}$, *i.e.* $\mathbf{y} = \mathbf{D}\mathbf{c}$ implies $2\mathbf{y} = \mathbf{D}(2\mathbf{c})$.

The prediction that there is an asymmetry between the linearity of the decoding function and the nonlinearity of the encoding function can be tested experimentally (Figure 8). Given an ensemble of stimulus-response pairs (*i.e.* the neural responses to an ensemble of sounds) obtained from a population of neurons, the model predicts that a linear stimulus reconstruction approach (*i.e.* a decoding model) will outperform a linear "forward" (*i.e.* encoding) model, but only if the optimal linear reconstructors are estimated from a population of neurons.

The idea that a linear approximation is better suited for the neural decoding than encoding function was first exploited to estimate the information rate of fly visual neurons (Bialek et al., 1991). By contrast, our model predicts that, if the neural representation is sparse and overcomplete, then the asymmetry should emerge only in multi-neuron recordings. To our knowledge, this asymmetry has not been tested for high-level auditory representations. Our model thus makes a strong prediction: that linear decoding does not provide an advantage over linear encoding for single neuron experiments, whereas the former outperforms the latter for multi-neuron experiments.

3.4.3 Context-dependence of STRFs

A third prediction that follows from the piecewise linearity of the encoding function is that the linear component of receptive fields should depend on the acoustic context. Following conventional usage in auditory physiology, we will use the term spectrotemporal receptive field, or STRF, to refer only to the *linear* component of the encoding function, even though the encoding function itself may be highly nonlinear (Theunissen et al., 2000, 2001; Kowalski et al., 1996). (In visual physiology, "STRF" is used to refer to the "*spatial* temporal receptive field," but the quantities are analogous.) The STRF is the analog (in a high-dimensional input space) of the slope of a neuron's tuning curve in one dimension.

In an experimental setting, piecewise linearity predicts that the STRF should depend on the acoustic context. We define the acoustic context of a feature $\mathbf{d}_{ij}$ with respect to a stimulus $\mathbf{y}$ as the collection of other features activated simultaneously by that stimulus. In music, for example, the features tend to resemble musical notes, and the acoustic context can be thought of as the set of notes (*e.g.* in a chord) that accompany a given note. Figure 10 shows the STRF of the same neuron (a trumpet feature) in two different contexts (either clarinet or flute.) The gross features of the STRF (*e.g.* the excita-

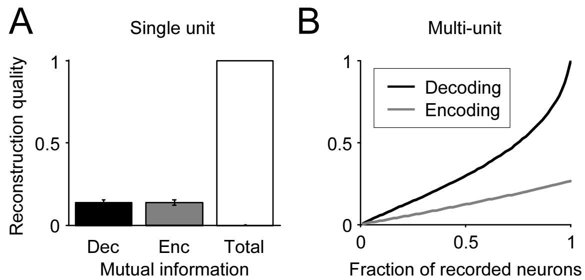

Figure 8: **Prediction 2: Linear decoding outperforms linear encoding for multi-neuron but not single neuron experiments. (A)** The mutual information between a simulated neuron's response and the stimulus (*Total*) was compared with the mutual information between the stimulus and the optimal linear estimate of the stimulus obtained from the activity of a single neuron (*Dec*), and between the actual response and an optimal linear estimate of the response obtained from the stimulus (*Enc*). The two linear estimates were comparable, and both captured only a fraction of the total information, indicating that encoding and decoding are comparable for single neurons. **(B)** Decoding outperforms encoding in a simulated multineuron experiment. The reconstruction quality is plotted as a function of the optimal linear decoder (*dark curve*) or the optimal linear encoder (*light curve*). The reconstruction quality is a normalized measure of the accuracy of reconstruction, defined as $1 - \langle \|\mathrm{error}\|_2 / \|\mathrm{signal}\|_2 \rangle$; *see methods for details*. Encoding and decoding perform comparably when only a few neurons are recorded, but as the number of neurons recorded increases, the reconstruction quality of decoding grows faster. When the activity of all neurons involved in the representation is recorded (75 in this simulation), decoding is perfect.

tory band around 880 Hz) are preserved in both contexts, but the secondary features (*e.g.* the addition of an inhibitory sideband) is context-sensitive. Changes in the STRF for different features and different contexts can be larger or smaller than in this example. Stimulus context thus changes the neural encoding function, suggestive of the non-classical receptive field modulation observed in visual and auditory cortexes

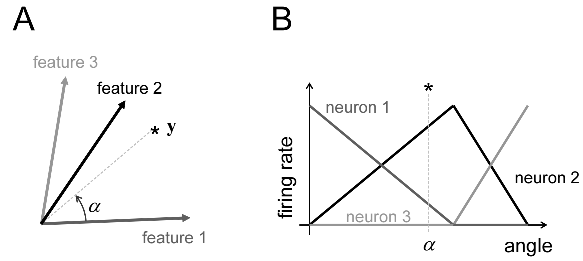

Figure 9: **Encoding is nonlinear (piecewise linear). (A)** Three features in two dimensions constitute an overcomplete basis. A sample signal $\mathbf{y}$ is indicated with an '$*$'. **(B)** Tuning curves for the three features are piecewise linear. The firing rate of each of the three units in (A) is given as a function of angle for stimuli of unit length; the point $\mathbf{y}$ in (A) is at about 45°. Because the sample space is two-dimensional, any given point is represented by at most two active neurons. Decoding is linear: the point $\mathbf{y}$ is recovered by a weighted sum of the features, with the corresponding neural activities constituting the weights. Encoding, however, is nonlinear: the slope of all active neurons' activation functions can change at the boundaries, whenever any neuron becomes active or inactive. The basic intuition shown here generalizes to the other examples in this paper, in which the dimensionality of the space (given by the number of elements in the spectrogram) is much higher.

(David et al., 2004; Valentine and Eggermont, 2004).

Context-dependence as defined here is stronger than simple nonlinearity. Specifically, the prediction is that there should exist extended subregions of stimulus space where the encoding function of a given target neuron is one linear function, and across some boundary in stimulus space switch to a second linear function. These boundaries are demarcated by the activation of another (non-target) neuron in the population and the de-activation of a second (non-target) neuron (Figure 9). This prediction could be tested using a multi-neuron recording technique.

The locally linear encoding induced by sparseness may help reconcile some of the apparent contradictions in the auditory literature. STRFs obtained using a "moving ripple" basis can predict responses to linear combinations of basis elements (Kowalski et al., 1996). However, linear encoding (STRF) models fail to predict neural responses when the stimulus domain is extended to include a wide selection of complex sounds (Machens et al., 2004; Linden et al., 2003), consistent with the idea that ripples represent a subspace within which encoding is linear. Context sensitivity may also provide an explanation for a proposed neural correlate of comodulation masking release in which the addition of a pure tone can suppress the response to temporally-modulated noise (Nelken et al., 1999); this form of contextual modulation cannot be explained by any purely linear encoding model.

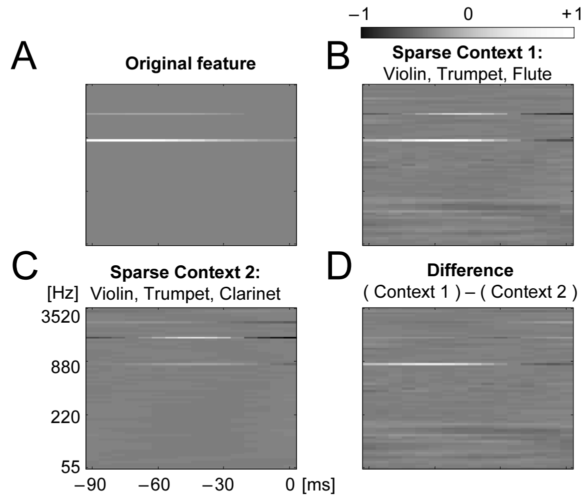

Figure 10: **Prediction 3: Dependence of STRF on context. (A)** Spectrogram of trumpet feature, showing a strong fundamental around 880 Hz and some higher harmonics. **(B,C)** The STRFs corresponding to the feature in (A) when that feature is activated in two different contexts (clarinet or flute played simultaneously), derived under the assumption of a sparse neural representation. The STRF provides the *encoding* from the stimulus to neural activity. The color at any point of the STRF indicates the value (in spikes/second) of the kernel which is convolved with the spectrogram of the stimulus to generate a neural response. Under the sparse assumption, the encoding is piecewise linear, and the STRFs shown are two out of the many possible pieces. The STRF is obtained from the appropriate row of the matrix $\mathbf{D}_k^\diamond$ (see Appendix A.3). **(D)** The difference between the two spectrograms. Note that they show the same basic harmonic structure, but differ in details such as the relative contributions of the excitatory and inhibitory sidebands. The differences can be as large as the STRFs themselves.

4 Discussion

Our main result is that the appropriate 'sparse' neural representation implicitly separates a mixture of sound sources into its constituent auditory streams. In this model, sources at different positions in space were separated with only monaural information by exploiting the differential filtering imposed by the HRTF, under the assumption that the source locations have already been identified by other mechanisms. This model provides a possible explanation for an important question about cortical organization: Why are there so many more neurons in the auditory (or visual) cortex than in the cochlea (or retina)? The answer we provide, motivated by the ability of an overcomplete sparse representation to separate sources, is potentially quite general, and may be applicable to other brain regions as well.

This model was motivated foremost by the computational demands of source separation. Source separation is a complex computation, and we could no more expect to solve the whole problem in its entirety here than we could expect to solve completely its visual analog—scene segmentation—or any of the many other challenging problems in computational vision. We have instead concentrated on a restricted form of the problem involving only the spatial cues introduced by the HRTF, with the expectation that the same framework can be generalized to understand how some other cues might be used.

Sparseness provides a powerful and useful constraint on neural activities. Our results complement and extend previous work on finding features that permit an efficient representation of auditory or visual scenes (Olshausen and Field, 1997, 1996; Lewicki, 2002; Bell and Sejnowski, 1997; Schwartz and Simoncelli, 2001; Klein et al., 2003; Smith and Lewicki, 2006). In our framework the HRTF "tags" these features so they can be assigned to the appropriate source. Other psychophysical cues important for acoustic stream segregation, such as common onset time, could be used in a similar way.

4.1 Efficient coding and sparseness

Our approach is compatible with the "efficient coding hypothesis" (Barlow, 1961), according to which the goal of sensory processing is to construct an efficient representation of the sensory environment. Building on these results, we used NMF to derive a set of basis elements with which auditory stimuli could be represented sparsely. Although we did not test bases obtained using other approaches, we do not expect that our results would be sensitive to the particular method used to find the basis; any sparse basis would likely have worked.

The principle of efficient, or sparse, coding has been used to predict receptive field properties of both auditory and visual neurons (Simoncelli and Olshausen, 2001; Lewicki, 2002; Smith and Lewicki, 2006; Bell and Sejnowski, 1997; Olshausen and Field, 1996). However, the focus of the present work is not on the receptive field properties themselves, but rather on how the resulting sparse representation can subserve a computation. Our predictions are therefore not about the detailed structure of receptive fields, but rather about how receptive fields interact.

The motivation for sparseness here is not coding (or metabolic (Laughlin and Sejnowski, 2003; Levy and Baxter, 1996)) efficiency *per se*, but rather performance on a particular computational problem: source separation. There are contexts in which coding efficiency imposes important constraints on representation; for example, the retina compresses visual information collected at 10^8 photoreceptors into a signal that is carried by only about 10^6 fibers in the optic nerve. For source separation, however, sparseness provides a mathematical instantiation of Occam's Razor: it allows a search for the most likely interpretation to be conducted by searching for the sparsest interpretation.

Sparse encoding implies that most stimuli should elicit only modest firing in most neurons, as has been observed experimentally for both simple and complex auditory and visual stimuli (DeWeese et al., 2003; Machens et al., 2004; Vinje and Gallant, 2000), but it does *not* imply that responses must be weak for all stimuli. Sparseness implies merely that stimuli elicit only a small number of spikes across the neuronal population; indeed, a neuron encoding some particular feature $\mathbf{d}$ will fire maximally when the stimulus $\mathbf{d}$ is presented. Sparseness in this model is therefore a constraint on the activity of the population of neurons involved in a representation, rather than on the activity of any single neuron. Our results are thus fully consistent with experiments indicating that it is sometimes possible to optimize stimuli online to obtain high firing rates (Barbour and Wang, 2003; deCharms et al., 1998), since in our framework such a stimulus is the feature associated with the neuron.

Directly assessing the sparseness of a neuronal representation experimentally is difficult. The key issue is how many neurons (or spikes) participate in the representation of a typical stimulus. Ideally this would be measured by recording all spikes from all neurons simultaneously, but this is not possible using the experimental techniques currently available. Nevertheless, there is growing evidence that natural auditory (DeWeese et al., 2003; Machens et al., 2004) and visual (Vinje and Gallant, 2000; Baddeley et al., 1997) stimuli activate only a relatively small number of neurons (Olshausen and Field, 2004). Thus the representational sparseness assumed by this model should be viewed as at least provisionally consistent with the current experimental evidence about cortical representations.

4.2 Overcomplete representations as a model of cortex

Although there is nothing in the model that explicitly ties it to one or another brain area, we think it most likely that, at least in mammals, the operations we describe occur in the cortex, rather than at subcortical stations. First, receptive fields in auditory cortex are heterogeneous, and often have the broad and complex spectrotemporal structure (Sutter, 2000) required to exploit the HRTF.

Secondly, and more significantly, auditory cortex has the characteristics expected from an overcomplete representation. There are about 30,000 auditory nerve fibers, but more than a thousand times as many auditory cortex neurons. Assuming that the "representational fidelity"—the amount of information that can be represented by a single spike—of neurons in cortex is comparable to that at the periphery, the "representational capacity" of cortex is far in excess of what is needed to form merely a complete representation. This model suggests a way in which this excess representational bandwidth can be used for computation, instead of merely to overcome neuronal noise as has sometimes been proposed.

Although we do not know how sparse cortical representations are achieved, it seems likely that the underlying circuitry involves lateral interactions. Indeed, "sparsification" is similar to divisive normalization approaches, which are motivated by both circuit and computational considerations (Schwartz and Simoncelli, 2001). Explicit circuit dynamics and connection weights can be obtained using gradient descent to minimize the total neural activity in Eq. 10.

4.3 Model predictions

We have identified three clear experimental predictions of the model. First, the optimal linear decoder estimated from an experiment in which the activity of multiple neurons are recorded should maximize a target neuron's firing rate. Second, there should be an asymmetry between the performance of the optimal linear encoder and decoder, but this asymmetry should become evident only in interpreting multi-neuron recording experiments; the model predicts that the optimal linear encoder and decoder in a single neuron experiment both underperform the "true" optimal (*i.e.* nonlinear) decoder. Finally, the STRF should be dynamically influenced by acoustic context. These predictions can be used to test—and falsify—the model.

Perhaps the most surprising prediction of this model is how stimulus optimization for a target neuron can improve by recording from other neurons involved in the representation. To our knowledge, online stimulus optimization using data from more than one neuron has not been previously proposed, but could be practical using modern multi-neuron recording techniques.

We have assumed that the neural decoding function—the transformation from the neural response to the stimulus—is linear. However, we have shown that sparseness implies that the neural encoding function—the inverse transformation from the stimulus to the response—is in general nonlinear (piecewise linear). This asymmetry, which emerges only in multi-neuron recording experiments, is a strong and testable prediction of the model.

This asymmetry further implies the context-dependence of the STRF. We speculate that this may explain why linear (STRF) encoding models work well in restricted domains (Kowalski et al., 1996) but fail for richer stimulus ensembles (Machens et al., 2004; Linden et al., 2003). However, context dependence has not yet been tested directly.

4.4 Relation to independent component analysis (ICA)

Our formulation of the source separation problem (Eq. 3) differs in two respects from the one usually considered in the ICA literature. First, we have assumed pre-filtering of each source by a known filter $\mathbf{h}$, whereas in the usual formulation the weighting of each source is given by an unknown scalar. Second, in most formulations the receiver is assumed to have access to the sources via several sensors, each of which is exposed to a different (linear) combination of the sources, whereas here we assume only a single sensor with input $\mathbf{y}$.

Most approaches to solving such multi-sensor formulations focus on recovering an "unmixing matrix" which inverts the mixing matrix governing the weighting of each source at each sensor. In such cases, it is generally sufficient to assume simply that the sources are statistically independent (Comon et al., 1991; Comon, 1994; Bell and Sejnowski, 1995; Belouchrani et al., 1997; Amari and Cichocki, 1998). If, however, there is only a single sensor, the problem is degenerate and such approaches fail: separating multiple sources from a single sensor requires assumptions stronger than simple independence (Cauwenberghs, 1999; Roweis, 2001; Hochreiter and Mozer, 2001; Jang and Lee, 2003; Smaragdis, 2004). The intermediate case—at least two sensors but more sources than sensors—simplifies the problem considerably (Bofill and Zibulevsky, 2001; Rickard and Dietrich, 2000; Linsker, 2001), because binaural cues as well as monaural ones can be utilized.

Recent advances in ICA have emphasized the utility of sparse overcomplete representations for source separation problems in acoustic, visual and other domains (Farid and Adelson, 1999; Lee et al., 1999; Lewicki and Sejnowski, 2000; Zibulevsky and Pearlmutter, 2001; Li et al., 2004; Levin and Weiss, 2004). Here we have built on these ideas, and developed a novel approach to separating multiple "augmented" (pre-filtered) signals combined at a single sensor.

Our framework can be generalized to exploit other cues used in single sensor separation, such as common onset time. It can also readily be extended to make use of binaural information. Each HRTF function is made single-input two-output, and the lengths of the column vectors corresponding to the post-HRTF dictionary elements and the observation vector are doubled. Interaural time and level disparity can then be used to separate sources. Information from two (or more) sensors can thus be naturally incorporated.

4.5 HRTF and source separation

The model assumes that sources have a statistical structure consisting of spectral correlations that can be exploited by filtering by the HRTF. One novel contribution of this work is its specific proposal for how the HRTF can be used for source separation, a process related to but distinct from sound localization. Spectral cues are not strictly required for sound localization: binaural cues can provide robust cues even in the absence of spectral cues. Conversely, source separation can proceed when spectral cues are weak—or indeed, even when spatial cues are completely absent, as for example when picking out a violin from within a concerto played over a single speaker. This illustrates a general principle: no single cue is essential to source separation, and the auditory system will promiscuously exploit any cues that are available.

Nevertheless, it is clear that HRTF cues, when present, help in source separation (Yost et al., 1996). We have shown how neural systems can exploit these cues using sparse representations. We speculate that sparseness may represent a general adaptation used by the nervous system to separate acoustic sources, and that similar principles may be also involved in source separation in other modalities, such as olfaction.

Acknowledgments

We thank Tomas Hromadka, Mike DeWeese, Zach Mainen, Carlos Brody, Bruno Averbeck, and Barbara Shinn-Cunningham for helpful comments. Supported by Higher Education Authority of Ireland (An tÚdarás Um Ard-Oideachas) and Science Foundation Ireland grant 00/PI.1/C067 (BAP), a Farish-Gerry Fellowship (HA) and the Sloan Foundation, Mathers Foundation, NIH, Packard Foundation and the Redwood Neuroscience Institute (AMZ).

A Appendix

A.1 Notes on regularizers

L_0 minimization: Minimizing sparseness in the L_1 sense is not the only possible choice. One natural alternative is the L_0 norm, which minimizes the total number of active neurons—the total number of nonzero activities c_{ij}—rather than the total number of spikes. Although this constraint also seems biologically sensible, it leads to a computationally intractable (NP-complete) combinatorial problem (Donoho and Elad, 2003); moreover, in many cases it leads to the same solution as the minimum L_1 solution

(Li et al., 2004), particularly in the presence of a noise model. We therefore consider only the L_1 solution here.

Conditioning of dense representation: From Eq. 11 it is clear that the uniqueness of the solution depends on the invertibility and condition number (ratio of largest to smallest singular values) of $\mathbf{D}$, whose columns $\mathbf{d}_{ij} = \mathbf{h}_i * \mathbf{q}_j$ depend in turn on both the filters $\mathbf{h}_i$ and the source elements $\mathbf{q}_j$. In particular, if there is no filter $\mathbf{h}_i$, as in the usual formulation of source separation, then the columns of $\mathbf{D}$ are identical, and the solution is therefore degenerate. The greater the difference between the filtered copies of the sources—the columns of $\mathbf{D}$—the smaller the condition number of $\mathbf{D}$ and the more numerically stable the problem.

Probabilistic interpretation of regularizers: Interpreted probabilistically, the regularizers on neural representations (Eqs. 10 and 11) correspond to maximum likelihood estimates using different *a priori* assumptions about the processes generating the stimuli, whose estimates are represented as the neural activities participating in a representation (Figure 1D). The pseudoinverse assumes that the underlying causes represented by the activities c_{ij} were drawn from a Gaussian distribution, while the sparseness regularizer assumes instead a Laplacian distribution, $p(c_{ij}) \propto e^{-|c_{ij}|}$. Because a Laplacian distribution has more elements very close to (and very far from) zero than does a Gaussian with the same variance, it corresponds to a sparser description in terms of c_{ij}. Note that without a noise term, the maximum likelihood estimates using any prior yields perfect (zero reconstruction error) representations of the stimulus; the prior here is on the distribution of the underlying causes represented by the coefficients c_{ij}, rather than on the distribution of reconstruction errors (as for example in robust fitting methods.) Only when a noise term is added (Eq. 1) do the neuronal activities c_{ij} cease to represent the stimulus perfectly.

A.2 Asymmetry of sparse representations

Why does an overcomplete sparse representation predict an asymmetry between encoding and decoding (Section 3.4.2)? To understand this, let us begin by considering the more familiar case of a complete (but not overcomplete) orthonormal representation, such as a Fourier or ripple basis. In this case, there is perfect symmetry between the encoding and decoding transformations. (These transformations are referred to as *analysis* and *synthesis* in the wavelet literature.) That is, if the columns of $\mathbf{D}$ in the decoding equation $\mathbf{y} = \mathbf{Dc}$ (Eq. 9) are orthonormal, then the inverse (encoding) transformation is given by $\mathbf{c} = \mathbf{D}^T\mathbf{y}$, where we have used the fact that the inverse of an orthonormal matrix is its transpose, $\mathbf{D}^T = \mathbf{D}^{-1}$. In this case, the encoding and decoding filters—the rows and columns of $\mathbf{D}$—are identical. This explains the familiar symmetry between the forward and inverse Fourier transform, in which the rows and columns of $\mathbf{D}$ are sinusoids.

If the columns of a square matrix $\mathbf{D}$ are not orthonormal, then its inverse (if it exists) is not equal to its transpose. However, the encoding transformation $\mathbf{c} = \mathbf{D}^{-1}\mathbf{y}$ is still linear, since it can be expressed as a linear combination of the columns of $\mathbf{D}^{-1}$. Even in the overcomplete case (Eq. 11), linearity of encoding is preserved if the dense representation is used, since in that case the encoding filters are the columns of the pseudoinverse $\mathbf{D}^{\diamond}$.

In the sparse overcomplete case, asymmetry arises because the inverse transformation is in general nonlinear. Here, the representation is found by solving the optimization problem of Eq. 10. Note that this in turn gives a reason that stimulus optimization requires multi-neuron recordings (Section 3.4.1). Specifically, because the neural responses cannot be explained by a single (or a global) linear encoding function, the feature (or the stimulus that elicits the maximum response) cannot be estimated correctly by a linear regression using single unit data.

A.3 Context dependence of STRFs

In order to understand the context dependence of STRFs (Section 3.4.3), we consider the encoding function associated with the sparse representation of a particular stimulus $\mathbf{y}_k$. Consider the "packed matrix" $\mathbf{D}_k$ whose columns are the subset of features involved in the sparse representation of the stimulus $\mathbf{y}_k$, *i.e.* only the columns corresponding to the nonzero elements of $\mathbf{c_k}$. This matrix satisfies the decoding relation

$$\mathbf{y}_k = \mathbf{D}_k \bar{\mathbf{c}}_k, \tag{12}$$

where $\bar{\mathbf{c}}_k$ consists of $\mathbf{c}_k$ without zero elements. However, because of the sparseness (L_1) prior, the matrix $\mathbf{D}_k$ is at most full-rank, and is constructed from only at most N_{row} features. It is thus not overcomplete, and so the encoding can be specified by a matrix using the pseudoinverse,

$$\bar{\mathbf{c}}_k = \mathbf{D}_k^{\diamond} \mathbf{y}_k. \tag{13}$$

Thus the encoding function is continuous and piecewise linear, with the linear segments defined by pseudoinverses $\mathbf{D}_k^{\diamond}$ and the discontinuities in the first derivative occurring whenever the stimulus activates (or deactivates) a new feature (*i.e.* an element of $\mathbf{c}$ becomes non-zero or zero) and thereby changes $\mathbf{D}_k$.

The STRF for the i^{th} neuron is then obtained from the i^{th} row of the pseudoinverse of the packed matrix $\mathbf{D}_k$. (Recall that, following convention, we use the term "STRF" to refer only to the linear component of the encoding function from stimulus to response.)

Three comments on the STRF estimation are in order (Section 2.6). First, we note that constructing the matrix $\mathbf{D}_k$ requires knowledge of the solution $\mathbf{c}_k$, so that this does not actually constitute an algorithm for finding $\mathbf{c}_k$. Second, in the special case of the dense representation, both encoding and decoding are linear. In this case, the encoding function for any stimulus (Eq. 11) is simply the pseudoinverse $\mathbf{D}^{\diamond}$ of the full matrix $\mathbf{D}$.

Finally, we note that the fact that even if two STRFs obtained from a single neuron's response to two different stimulus ensembles differ, and each accounts perfectly for the data to which it was fit, this does not rule out the possibility that there might exist some third STRF that perfectly fits the amalgamation of the two data sets. Consider two stimuli $\mathbf{y}_k$ and $\mathbf{y}_{k'}$ which activate only features in $\mathbf{c}_k$ and $\mathbf{c}_{k'}$, respectively. The pseudoinverse $\mathbf{D}_k^{\diamond}$ corresponding to the first stimulus will be valid (Eq. 13) for any stimulus composed of any subset of features that are nonzero in $\mathbf{c}_k$, but will not be valid for any features that were not active. Therefore, any stimulus consisting of sums of features in the union of the feature sets $\mathbf{c}_k$ and $\mathbf{c}_{k'}$ will yield incorrect results if either of the corresponding packed pseudoinverse matrices are used. However, a new pseudoinverse matrix constructed from features in the union of the feature sets $\mathbf{c}_k$ and $\mathbf{c}_{k'}$ could yield correct values for a superset of the stimulus ensembles for which either of the original two STRFs were valid.

Such supersets can only be constructed from a number of features that is limited by the rank of $\mathbf{D}$. That is, there does not exist a matrix $\mathbf{D}_k^{\diamond}$ that can be used in (Eq. 13) to generate the correct $\mathbf{c}$ for *all* stimuli. For the example in Figure 10, we confirmed that there does not exist an STRF that includes both as special cases.

References

S. Amari and A. Cichocki. Adaptive blind signal processing: Neural network approaches. *Proceedings of the IEEE*, 86(10):2026–2048, Oct. 1998.

H. Attias and C. Schreiner. Temporal low-order statistics of natural sounds. In *Advances in Neural Information Processing Systems*, 1997.

R. Baddeley, L. F. Abbott, M. C. Booth, F. Sengpiel, T. Freeman, E. A. Wakeman, and E. T. Rolls. Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proc R Soc Lond B Biol Sci*, 264:1775–1783, 1997.

D. L. Barbour and X. Wang. Auditory cortical responses elicited in awake primates by random spectrum stimuli. *J. Neurosci.*, 23(18):7194–7206, 2003.

H. B. Barlow. Possible principles underlying the transformations of sensory messages. In W. A. Rosenblith, editor, *Sensory Communication*, pages 217–234. MIT Press, 1961.

A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neu. Comp.*, 7(6): 1129–1159, 1995.

A. J. Bell and T. J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

A. Belouchrani, K. Abed Meraim, J.-F. Cardoso, and É. Moulines. A blind source separation technique based on second order statistics. *IEEE Trans on Signal Proc*, 45(2):434–444, Feb. 1997.

W. Bialek, F. Rieke, R. R. de Ruyter van Stevenick, and D. Warland. Reading a neural code. *Science*, 252:1854–1857, 1991.

P. Bofill and M. Zibulevsky. Underdetermined blind source separation using sparse representations. *Signal Processing*, 81(11):2353–2362, 2001.

A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, Massachusetts, 1990. ISBN 0-262-02297-4.

G. Cauwenberghs. Monaural separation of independent acoustical components. In *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS'99)*, volume 5, pages 62–65, Orlando FL, 1999.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

P. Comon. Independent component analysis: A new concept. *Signal Processing*, 36:287–314, 1994.

P. Comon, C. Jutten, and J. Hérault. Blind separation of sources, part II: Problem statement. *Signal Processing*, 24:11–20, 1991.

S. V. David, W. E. Vinje, and J. L. Gallant. Natural stimulus statistics alter the receptive field structure of v1 neurons. *J Neurosci*, 24(31): 6991–7006, 2004.

R. C. deCharms, D. T. Blake, and M. M. Merzenich. Optimizing sound features for cortical neurons. *Science*, 280(5368):1439–1443, 1998.

M. R. DeWeese, M. Wehr, and A. M. Zador. Binary spiking in auditory cortex. *J. Neurosci.*, 23(21):7940–7949, 2003.

M. R. Deweese, T. Hromadka, and A. M. Zador. Reliability and representational bandwidth in the auditory cortex. *Neuron*, 48:479–488, 2005.

D. L. Donoho and M. Elad. Maximal sparsity representation via $l1$ minimization. *Proc. Natl. Acad. Sci. U. S. A.*, 100:2197–2202, Mar. 2003.

H. Farid and E. H. Adelson. Separating reflections from images by use of independent components analysis. *J. of the Optical Society of America*, 16(9):2136–2145, 1999.

Y. I. Fishman, J. C. Arezzo, and M. Steinschneider. Auditory stream segregation in monkey auditory cortex: Effects of frequency separation, presentation rate, and tone duration. *J Acoust Soc Am*, 116(3):656–70, 2004.

R. Fletcher. Semidefinite matrix constraints in optimization. *SIAM J. Control and Opt.*, 23: 493–513, 1985.

R. H. Hahnloser, A. A. Kozhevnikov, and M. S. Fee. An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature*, 419(6902):65—70, 2002.

S. Hochreiter and M. C. Mozer. Monaural separation and classification of mixed signals: A support-vector regression perspective. In *3rd International Conference on Independent Component Analysis and Blind Signal Separation*, San Diego, CA, Dec. 9-12 2001.

P. M. Hofman and A. J. V. Opstal. Bayesian reconstruction of sound localization cues from responses to random spectra. *Biol Cybern*, 86 (4):305–316, 2002.

G.-J. Jang and T.-W. Lee. A maximum likelihood approach to single-channel source separation. *J. of Mach. Learn. Research*, 4:1365–1392, Dec. 2003. URL http://www.jmlr.org/papers/v4/jang03a.html.

D. Klein, P. Konig, and K. P. Kording. Sparse spectrotemporal coding of sounds. *Journal on Applied Signal Processing*, 7:659–667, 2003.

D. J. Klein, D. A. Depireux, J. Z. Simon, and S. A. Shamma. Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design. *Journal of Computational Neuroscience*, 9(11):85–111, 2000.

E. I. Knudsen and M. Konishi. Mechanisms of sound localization in the barn owl. *Journal of Comparative Physiology*, 133:13–21, 1979.

N. Kowalski, D. A. Depireux, and S. A. Shamma. Analysis of dynamic spectra in ferret primary auditory cortex II: Prediction of unit responses to arbitrary dynamic spectra. *J. Neurophysiol.*, 76(5):3524–3534, 1996.

K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neu. Comp.*, 15(2):349–396, 2003.

A. Kulkarni and H. S. Colburn. Role of spectral detail in sound-source localization. *Nature*, 396(6713):747–749, 1998.

S. B. Laughlin and T. J. Sejnowski. Communication in neuronal networks. *Science*, 301(5641): 1870–1874, 2003.

D. D. Lee and H. S. Seung. Learning the parts of objects with nonnegative matric factorization. *Nature*, 401:788–791, 1999.

T.-W. Lee, M. S. Lewicki, M. Girolami, and T. J. Sejnowski. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters*, 4(5):87–90, 1999.

A. Levin and Y. Weiss. User assisted separation of reflections from a single image using a sparsity prior. In *Proc. of the European Conference on Computer Vision (ECCV)*, Prague, May 2004.

W. B. Levy and R. A. Baxter. Energy efficient neural codes. *Neu. Comp.*, 8(3):531–543, 1996.

M. S. Lewicki. Efficient coding of natural sounds. *Nat. Neurosci.*, 5(4):356–363, 2002.

M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neu. Comp.*, 12 (2):337–65, 2000.

Y. Li, A. Cichocki, and S. Amari. Analysis of sparse representation and blind source separation. *Neu. Comp.*, 16(6):1193–1234, 2004.

J. F. Linden, R. C. Liu, M. Sahani, C. E. Schreiner, and M. M. Merzenich. Spectrotemporal structure of receptive fields in areas AI and AAF of mouse auditory cortex. *J. Neurophysiol.*, 90(4):2660–2675, 2003.

R. Linsker. Separation of a mixture of acoustic sources into its components. US Patent 6,317,703, Nov. 13 2001.

C. K. Machens, M. S. Wehr, and A. M. Zador. Linearity of cortical receptive fields measured with natural sounds. *J. Neurosci.*, 24(5):1089–1100, 2004.

C. Micheyl, B. Tian, R. P. Carlyon, and J. P. Rauschecker. Perceptual organization of tone sequences in the auditory cortex of awake macaques. *Neuron*, 48(1):139–48, 2005.

I. Nelken, Y. Rotman, and O. B. Yosef. Responses of auditory-cortex neurons to structural features of natural sounds. *Nature*, 397(6715): 154–157, 1999.

T. Nishino, Y. Nakai, K. Takeda, and F. Itakura. Estimating head related transfer function using multiple regression analysis. *IEICE Trans. A*, J84-A(3):260–268, 2001. In Japanese.

B. A. Olshausen and D. J. Field. Sparse coding of sensory inputs. *Curr Opin Neurobiol*, 14(4): 481–487, 2004.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23): 3311–3325, 1997.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

B. A. Olshausen and K. N. O'Connor. A new window on sound. *Nat. Neurosci.*, 5:292–293, 2002.

T. Poggio, V. Torre, and C. Koch. Computational vision and regularization theory. *Nature*, 317 (6035):314–319, 1985.

S. T. Rickard and F. Dietrich. DOA estimation of many W-disjoint orthogonal sources from two mixtures using DUET. In *Proceedings of the 10th IEEE Workshop on Statistical Signal and Array Processing (SSAP2000)*, pages 311–314, Pocono Manor, PA, Aug. 2000.

M. Riesenhuber and T. Poggio. Models of object recognition. *Nat. Neurosci.*, 3 Suppl:1199–1204, 2000.

S. T. Roweis. One microphone source separation. In *Adv. in Neu. Info. Proc. Sys. 13*, pages 793–799. MIT Press, 2001.

O. Schwartz and E. P. Simoncelli. Natural signal statistics and sensory gain control. *Nat. Neurosci.*, 4(8):819–825, 2001.

E. P. Simoncelli and B. A. Olshausen. Natural image statistics and neural representation. *Annu. Rev. Neurosci.*, 24(1):1193–1216, 2001.

P. Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *Fifth International Conference on Independent Component Analysis*, LNCS 3195, pages 494–499, Granada, Spain, Sept. 22–24 2004. Springer-Verlag.

E. C. Smith and M. S. Lewicki. Efficient auditory coding. *Nature*, 439:978–982, 2006.

G. Strang. *Linear Algebra and Its Applications*. Thomson, 3rd edition, 1988. ISBN 0155510053.

M. L. Sutter. Shapes and level tolerances of frequency tuning curves in primary auditory cortex: quantitative measures and population codes. *J. Neurophysiol.*, 84(2):1012–1025, 2000.

F. E. Theunissen, K. Sen, and A. J. Doupe. Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *J. Neurosci.*, 20(6):2315–2331, 2000.

F. E. Theunissen, S. V. David, N. C. Singh, A. Hsu, W. E. Vinje, and J. L. Gallant. Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network*, 12(3): 289–316, 2001.

P. A. Valentine and J. J. Eggermont. Stimulus dependence of spectro-temporal receptive fields in cat primary auditory cortex. *Hear Res*, 196(1–2):119–133, 2004.

W. E. Vinje and J. L. Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276, 2000.

E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman. Localization using nonindividualized head-related transfer functions. *J Acoust Soc Am*, 94(1):111–123, 1993.

F. L. Wightman and D. J. Kistler. Headphone simulation of free-field listening. II: Psychophysical validation. *J Acoust Soc Am*, 85(2):868–878, 1989.

W. A. Yost, R. H. Dye, Jr., and S. Sheft. A simulated "cocktail party" with up to three sound sources. *Percept Psychophys*, 58(7): 1026–1036, 1996.

M. Zibulevsky and B. A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neu. Comp.*, 13(4):863–882, Apr. 2001.