Efficient Coding of Natural Sounds

Michael S. Lewicki

Computer Science Department &
Center for the Neural Basis of Cognition
Carnegie Mellon University

Abstract

The auditory system encodes sound by decomposing the amplitude signal arriving at the ear into multiple frequency bands whose center frequencies and bandwidths are approximately logarithmic functions of the distance from the stapes. This particular organization is thought to result from the adaptation of cochlear mechanisms to the statistics of an animal's auditory environment. Here we report that several basic auditory nerve fiber tuning properties can be accounted for by adapting a population of filter shapes to optimally encode natural sounds. The form of the code is dependent on the class of sounds, resembling a Fourier transformation when optimized for animal vocalizations and a wavelet transformation when optimized for non-biological environmental sounds. Only for a combined set of vocalizations and environmental sounds does the optimal code follow scaling characteristics that are consistent with physiological data. These results suggest that the population of auditory nerve fibers encode a broad set of natural sounds in a manner that is consistent with information theoretic principles.

Correspondence:

Dr. Michael S. Lewicki, CMU/CNBC, 4400 Fifth Ave., Pittsburgh, PA 15213

Phone: 412-268-3921, fax: 412-268-5060, e-mail: lewicki@cnbc.cmu.edu, web: www.cs.cmu.edu/∼lewicki

Much is known about how the brain encodes sensory information, but why it has evolved to use the particular coding strategies it does have been the subject of long-standing interest and debate [1]. In the auditory system, cochlear nerve fibers are sharply tuned to specific frequencies and can be characterized as performing a short-term spectral analysis of acoustic signals [2]. To first approximation, the frequency and phase responses of auditory nerve fibers can be modeled as a bank linear filters that integrate auditory information over a timescale that varies with frequency [3–5]. Although these resemble filtering properties of Fourier and wavelet transforms, this observation alone does provide an adequate explanation for the auditory code, because it is not clear whether these transforms, which are derived largely from mathematical considerations, are optimal for processing the sensory environment experienced by the organism. A tonal decomposition might seem like a natural choice for harmonic sounds like vocalizations, but the natural environment is rich with sounds that are not harmonic. If these have equal behavioral significance, one would expect auditory systems to be adapted for processing a broad class of sounds, but the choice of the best code then is less obvious.

Can auditory sensory codes be explained by theoretical principles? One view based on information theory states that the goal of sensory coding is to encode the maximal amount of information about the stimulus by using a set of statistically independent features [1, 6–9]. In the auditory system, it has been shown that auditory nerves encode naturalistic stimuli with greater efficiency than white noise [10]. Testing the theoretical predictions of theory is important not only for gaining insight into the organization of the auditory neural code, but also how the codes of different organisms might be adapted for different auditory environments.

Efficient coding has had recent success in explaining the properties of receptive fields in primary visual cortex by deriving efficient visual codes from the statistics of natural images [11–14]. To test this theory in the auditory system, the technique of independent component analysis was used to derive efficient codes for different classes of natural sounds, including animal vocalizations, (non-biological) environmental sounds, and human speech. This yielded a theoretical prediction of the optimal code and provided an explanation for both the form of the filtering properties of cochlear nerves and their organization as a population. Previous explanations based on average power spectra, which does not take into account the temporal regularities, do not accurately predict the population characteristics of cochlear nerve fiber responses. The sound class that yielded a code most similar to that observed physiologically was a mixture of environmental sounds and animal vocalizations. Identical results were obtained with human speech, suggesting that its acoustic features make efficient use of the coding capacity of the auditory system.

Results

Auditory coding model

An auditory code based on the information theoretic principle of efficient coding can be derived by assuming the model

$$a_i(t) = \sum_{\tau=0}^{N-1} x(t-\tau)h_i(\tau) \tag{1}$$

in which the signal $x(t)$ in a time window of length N is encoded in the responses $a_1(t), ..., a_M(t)$. The goal of efficient coding is to derive a set of filters $h_1(t), ..., h_M(t)$ that minimize the statistical dependence of the outputs [6–9]. Methods of deriving efficient codes for models of the form in (1) fall under the rubric of sparse coding [11] or independent component analysis (ICA) [15, 16] and are equivalent to finding the features (or basis functions) that model the statistical distribution of the ensemble of waveforms within the analysis window [14].

Predicting Codes for Natural Sounds

The notion of an efficient code cannot be separated from the ensemble of signals that are being encoded [6, 17]. To make predictions for sensory codes, it is necessary to make conjectures about what class of stimuli the sensory system has evolved or is adapted to process. This could range from the general class of signals in the natural environment to only those crucial for survival. Here we analyze three classes of sounds chosen to be representative of a natural auditory environment: non-biological environmental sounds, animal vocalizations, and human speech, with each class containing a broad array of different sounds, animals, or speakers (see methods). Environmental sounds, such as rustling brush, crunching leaves and snapping twigs, are important for rapid and accurate auditory localization. These sounds are typically broadband, non-harmonic, and short in duration. Animal vocalizations by contrast, which plausibly have evolved to stand out from the background of environmental sounds, are typically harmonic, can have relatively narrow bandwidth, and are longer in duration. Speech represents a third distinct class of sounds, and is interesting not only because of its obvious relevance to the human auditory system, but also because it shares properties of both environmental sounds and vocalizations by virtue of having both consonants and vowels.

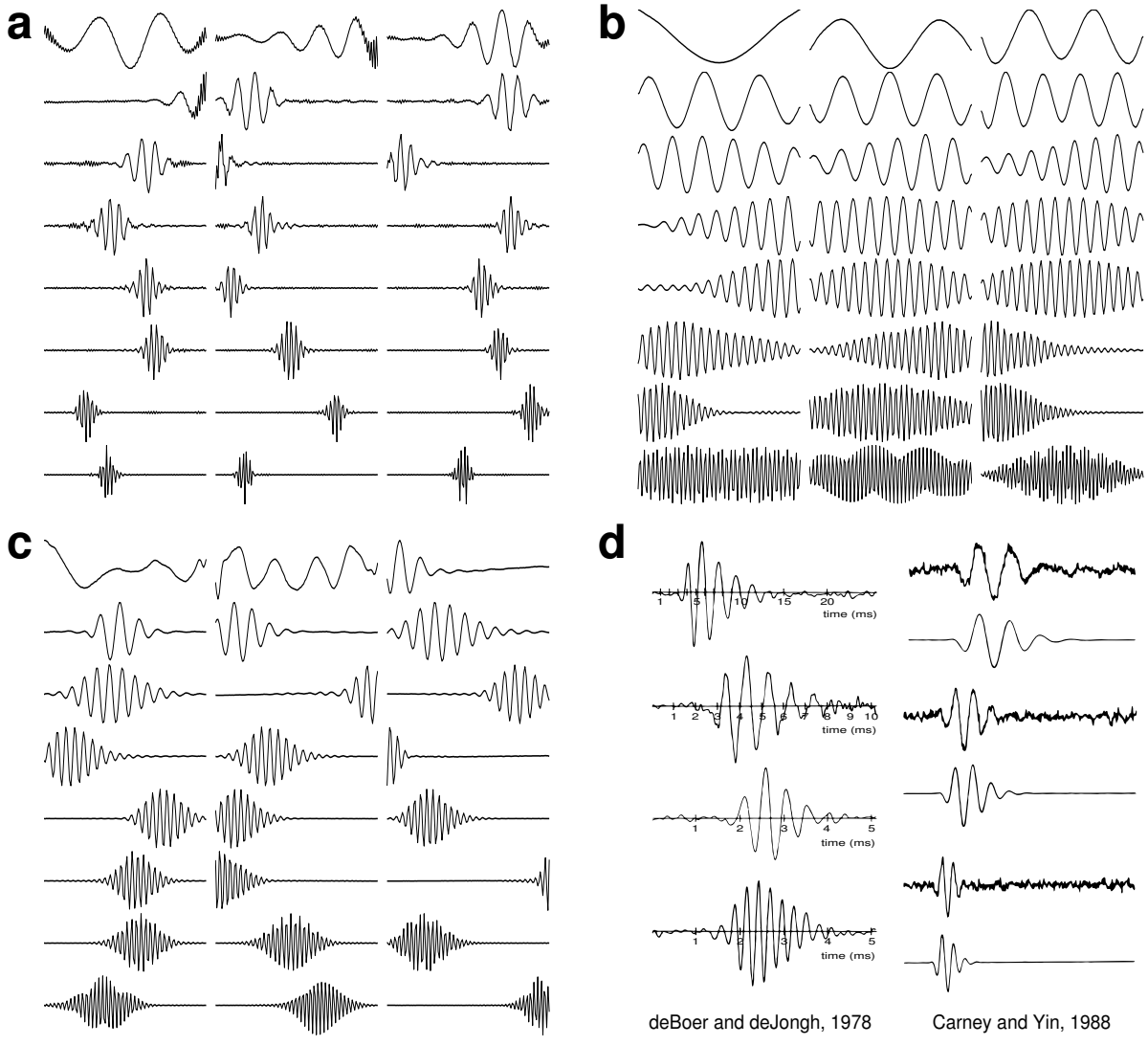
Efficient codes for each sound ensemble were derived using a generalized independent components algorithm (see methods). For reasons of computational efficiency, the analysis window was limited to 128-samples, which corresponds to a window width of approximately 8 msecs. Each filter is defined by 128 points over the analysis window and the algorithm places no constraints filter shape, i.e. each of the 128 points that define the filter is a free parameter, so the filters could take on any spectral or temporal

pattern. Those shapes that do emerge are determined by the statistical structure of the ensemble. A subset of the filter shapes for environmental sounds is shown in fig.1a. Even though the ensemble consisted of non-harmonic sounds like rustling brush and cracking twigs, the majority of the filters have a single peak resonance frequency and an amplitude envelope that is localized in time. Note that similar filters shapes can appear at multiple temporal positions, because the set of filters is optimized to encode the waveform over the whole the analysis window.

The properties of the filter population change with the statistical structure of the sound ensemble. The ICA filter shapes for animal vocalizations (fig.1b) is essentially a Fourier representation, with most filters having sinusoidal oscillations and little amplitude modulation. This type of representation would be expected as the sounds in ensemble of animal vocalizations where largely harmonic. That the filters are not localized in time reflects the fact that the statistical regularities of animal vocalizations occur on

---

Figure 1 *(following page)*: Auditory filters derived from the efficient coding of different classes of natural sounds: environmental sounds, animal vocalizations, and human speech (a-c). Individual waveforms show the entire filter in the time domain (approx. 8 msecs). The set of filters shapes is optimized to form an efficient code by maximizing the statistical independence of their response to the sound ensemble. Each plot shows a representative subset of the total population of 128 filters, displayed in increasing order of peak resonance frequency. (**a**) Efficient coding of non-harmonic environmental sounds yields a set of filters that resemble a wavelet representation. The majority have a dominant resonance frequency and an amplitude envelope that is localized in time. (**b**) Efficient coding of animal vocalizations results in filters that resemble a Fourier representation. All filters are sinusoidal and the majority extend over the entire length of the analysis window. The Moire-like patterns visible at the highest frequencies arise from the cyclic alignment of the underlying filter resonance frequency and the sampling frequency. (**c**) Efficient coding of an speech, which contains both harmonic and non-harmonic sounds (i.e. vowels and consonants), yields a representation intermediate between those in (a) and (b). (**d**) Comparison to cochlear filter shapes measured experimentally at the auditory nerve. These were obtained using reverse correlation which measures both the frequency and phase properties of the response of auditory nerve fibers. The *left column* shows filters redrawn from [4]. The peak resonance frequencies are 0.53, 1.0, 2.1, and 4.7 kHz, from top to bottom (note different time scales). The *right column* shows the measured filters (upper curves) and modeled functions (lower curves) redrawn from [5]. Each waveform is 20 msecs in duration, and the peak resonance frequencies are 364, 642, 999 Hz. Like the filter shapes predicted by efficient coding, the auditory nerve filters integrate the auditory signal over a time period that depends on frequency.


a
b
c
d
time (ms)
time (ms)
time (ms)
time (ms)
deBoer and deJongh, 1978
Carney and Yin, 1988

a much larger timescale than that for environmental sounds. A much larger analysis window would be required to show the temporal extent of this regularity. Efficient coding of speech also yields sinusoidal filters, but with amplitude envelopes that lie intermediate between those for environmental sounds and animal vocalizations (fig.1c). That the independent components of natural sound could produce filters that are localized in time and frequency was first observed by Tony Bell who applied ICA to melodic tooth tapping [18]. The power spectra of selected filters is shown in figure 2. The trade off between time and frequency is clearly visible in the progression from environmental sounds to vocalizations, which reflects the correlation time of the different sound ensembles.

The filters derived from efficient coding are qualitatively similar to filters that model the response properties of auditory nerve fibers. Examples of these are shown in fig.1d (redrawn from refs. [4] and [5]) were obtained using reverse correlation which provides an estimate of a linear filters that determine both the temporal and spectral properties of the auditory nerve response [3–5]. The so-called revcor filter shapes bear some resemblance to those derived theoretically for environmental sounds and human speech, and also show the dependence of the width of the amplitude envelope with frequency. Similar filter shapes also account the spectral analysis properties of the human auditory system [19].

These results are consistent with the notion that the sensory code used by the auditory system forms an efficient code for a mixture of non-harmonic broadband sounds and harmonic vocalizations. But it is possible that the intermediate temporal envelopes observed for speech arose because of a particular acoustic property of speech and not because it contains both harmonic and non-harmonic sounds. To test this explicitly, the same analysis was performed on a combined sound ensemble consisting of environmental sounds and animal vocalizations. The relative proportion of each ensemble could be chosen, and a proportion of two to one yielded filters similar to those for speech (fig.3).

Principal component analysis (PCA or the Karhunen-Loéve transform) has long been used in efficient coding of speech signals (e.g. [20]). To contrast the predictions of this approach with ICA, a set of filters for on the same ensemble of environmental sounds was derived using PCA. Fig.4 shows that the first few principal components are sinusoidal, but most are not localized in either time or frequency and bear little resemblance to auditory filters. As is well known, PCA achieves a weaker form of statistical independence by choosing filter shapes that decorrelate the outputs, and implicitly assumes the outputs follow a Gaussian distribution. The datasets used here, however, the outputs are highly non-Gaussian, and decorrelation, which results in a less efficient code, is not sufficient to explain the auditory filter properties. Furthermore, the PCA filters are restricted to be orthogonal, which highly restricts the class of filters that can be used to model structure in the sound ensemble. This restriction is not imposed by ICA.

To check whether there might be a bias in the efficient coding algorithm to give wavelet-like filters,

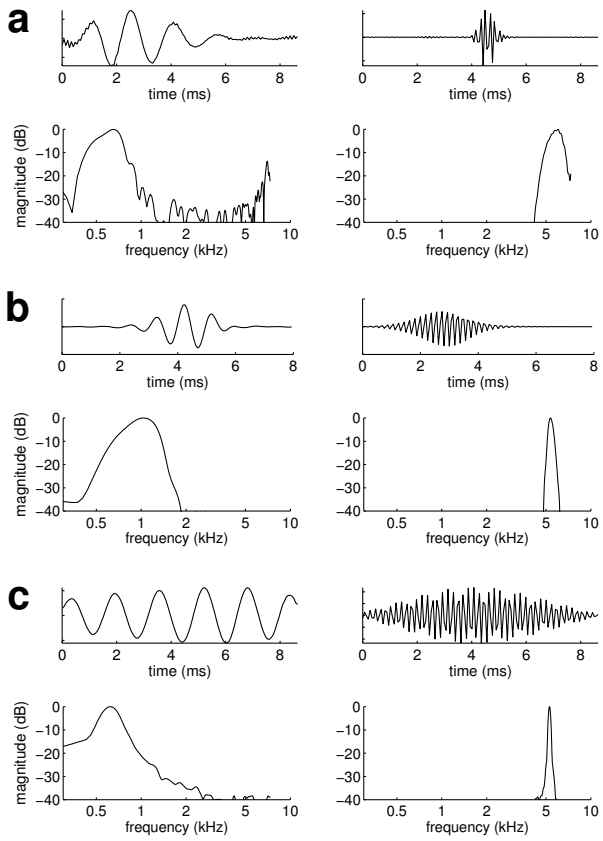


Figure 2: Power spectra of typical filters derived from environmental sounds (**a**), human speech (**b**), and animal vocalizations (**c**). The power spectra are shown below each filter.

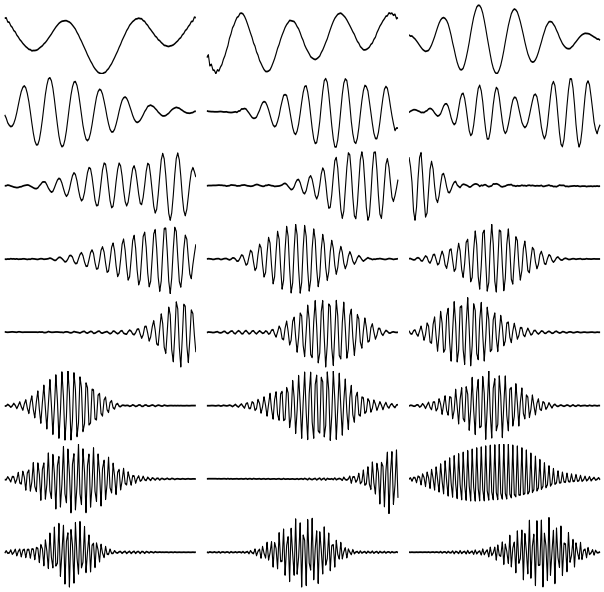


Figure 3: Efficient coding of a combined sound ensemble. Auditory filters derived from the efficient coding of a sound ensemble consisting of environmental sounds and vocalizations in a proportion of 2:1 yields filters similar to those for speech and have temporal envelopes that are intermediate between those for environmental sounds and animal vocalizations.

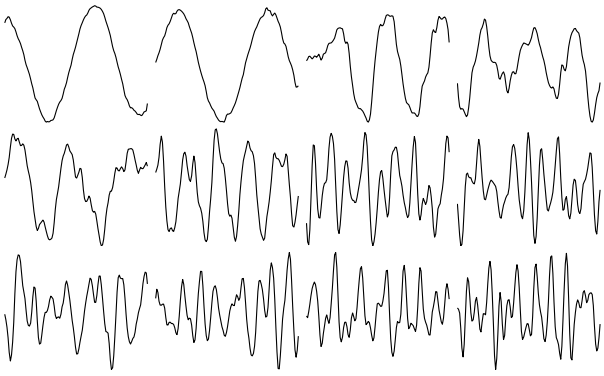


Figure 4: A subset of filters derived from principal component analysis (PCA) of environmental sounds. PCA, which can only model second-order correlations, does not yield filters that are localized in time, and only the largest are sinusoidal.

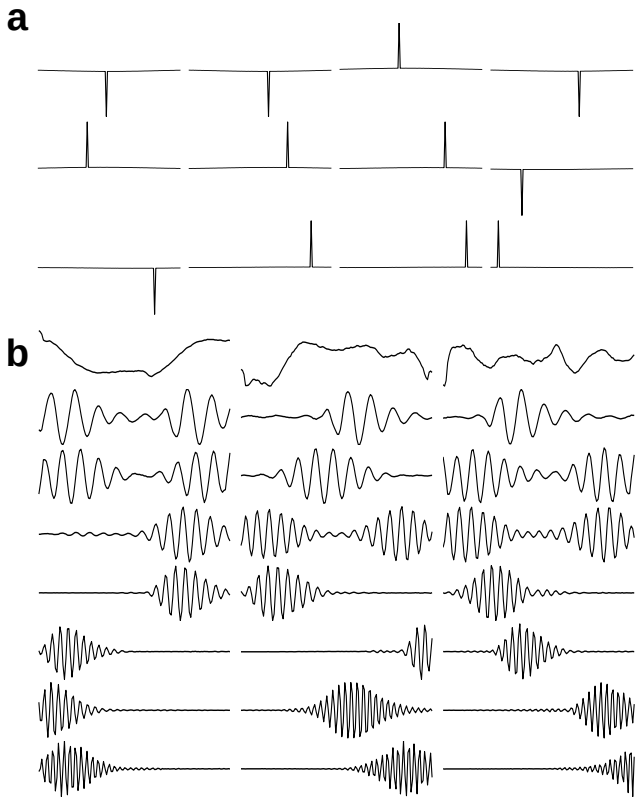


Figure 5: Sinusoidal filters that are localized in frequency are not inherently preferred by the algorithm. (**a**) Applying the same algorithm to sparse noise, where the samples are drawn independently from a sparse non-Gaussian distribution, results in an impulse representation where the filters are maximally localized in time. (**b**) Filters that are not localized in time result when an efficient code is derived for a single speaker. In this case, the non-localized filters occur at the lower frequencies for the harmonics of the speakers voice.

the algorithm was run on a data ensemble in which the samples were drawn independently from a sparse distribution ($p(x) \propto \exp(-|x|^{0.5})$). As would be expected for a data set that contains no temporal structure, the resulting filters were maximally localized in time, with each representing a different temporal position (fig.5a). If the algorithm is run on a single speaker from the speech dataset, the filters are not localized in time, and adapt to encode particular harmonics of the speaker's voice (fig.5b).

Analysis and Characterization of the Derived Codes

How the filter populations encode the three sound classes can be characterized using time frequency analysis, or the distribution of the filters in terms of their temporal envelopes and spectral power. For example, a Fourier transform represents a signal by a superposition of sinusoids. Thus, the filters are localized in frequency, but not in time (fig.6a). A wavelet transform, by contrast, is composed of filters

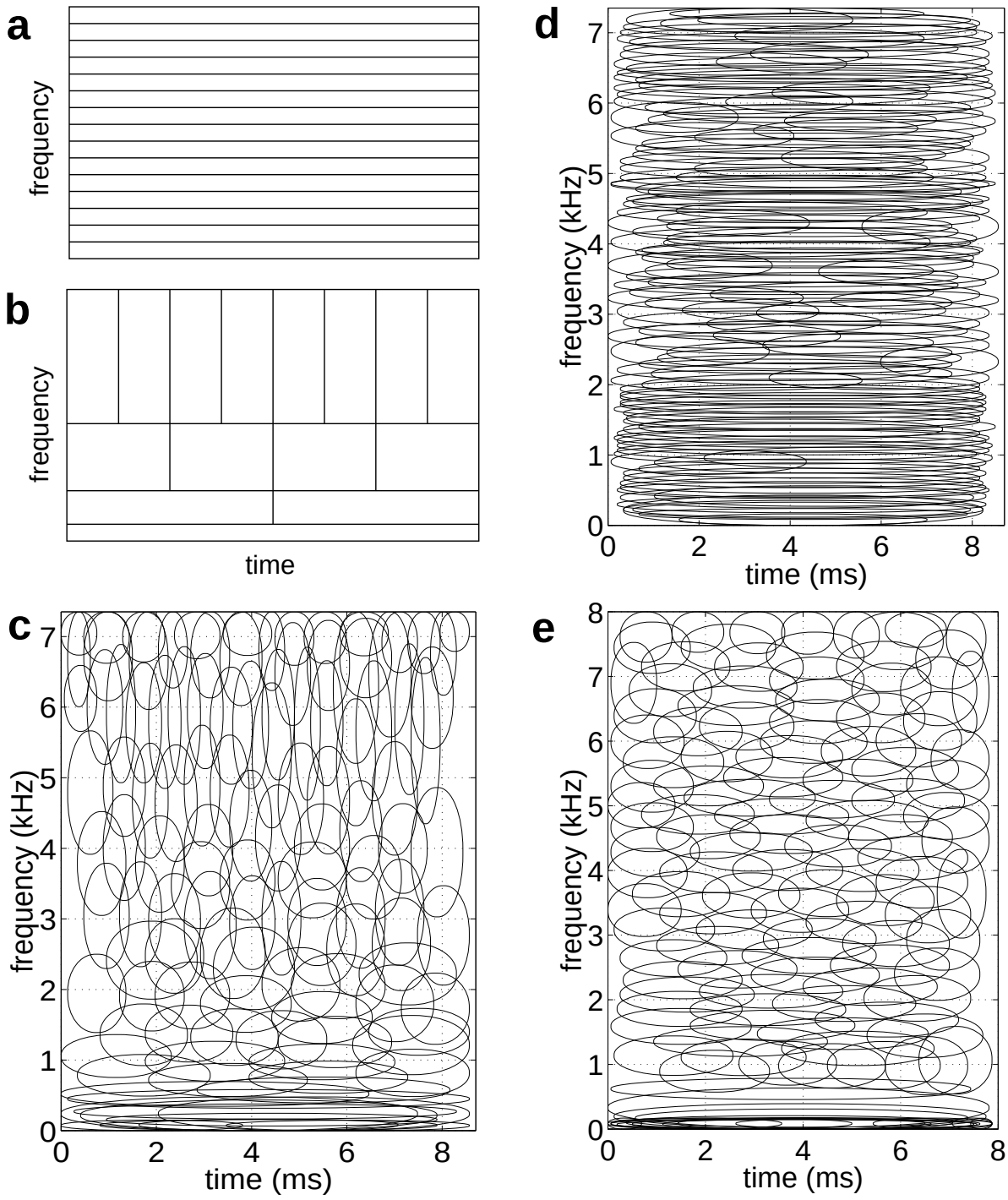
that are localized in both time and frequency (fig.6b). Because, the two codes cover the time-frequency space with the same number of functions, a wavelet representation while providing improved resolution in time, necessarily sacrifices resolution in frequency. The tiling of time-frequency space plays a central role in the design of wavelet transforms [21]. Because the individual filters derived from efficient coding are localized in their amplitude envelope and their spectral power, it is possible to plot how each population covers time-frequency space. For purposes of coding, which tiling is best depends on the statistical structure of the signals.

Time-frequency analysis (see methods for details) of the filters described in fig.1 shows that the majority of filters are localized (fig. 2). The time-frequency distribution for the environmental sounds code is similar to a wavelet representation with bandwidth increasing as function of frequency and temporal width decreasing (fig.6c). One difference, however, is that rather than having discrete increases in bandwidth and decreases in temporal width, which is common for many types of wavelets, bandwidth and temporal width change gradually with frequency. For animal vocalizations, the efficient code most closely resembles a Fourier representation, and the bandwidths are much narrower than the codes for the other datasets. The time-frequency distribution for the efficient code of speech falls in between the two others, both in terms of temporal extent and filter bandwidth. Deriving an optimal code provides a solution to the choice of how to tile time-frequency space, but it is more general, because the filters are not restricted to be localized in time-frequency space. Harmonic structure would be one example of non-local time-frequency structure, which is what one obtains for an efficient code of a single speaker (fig.5b) [22]. That the majority of the filters are localized reflects the statistical structure of the signals.

The differences between the time-frequency tiling for the three sound classes can be summarized by plotting characteristics of the filters in time frequency space as a function center frequency (fig.7). For comparison to auditory nerve filtering properties derived physiologically and psychophysically, we analyze bandwidth, filter sharpness (center frequency divided by bandwidth or Q), and the temporal envelope [23]. Bandwidth is measured 10 dB down from the spectral peak. Filters that did not have a full 10 dB drop on both sides of the spectral peak (the lowest and highest frequencies) were omitted from

---

Figure 6 *(following page)*: Time-frequency analysis. (**a**) The filters in a Fourier transform are localized in frequency, but not in time. (**b**) Wavelet filters are localized in both time and frequency. (**c-e**) The statistical structure of the signals determines how the filter shapes derived from efficient coding of the different data ensembles are distributed in time-frequency space. Each ellipse is a schematic of the extent of a single filter in time-frequency space. (**c**) Environmental sounds. (**d**) Animal vocalizations. (**e**) Speech.


a
frequency
b
frequency
time
d
frequency (kHz)
time (ms)
c
frequency (kHz)
time (ms)
e
frequency (kHz)
time (ms)

the plots to avoid artifacts introduced from the limited size of the analysis window. The filters optimized for environmental sounds show the steepest increase in bandwidth as a function of frequency, similar to a wavelet representation (fig.7a). By contrast, the filters derived for vocalizations have bandwidth that is nearly constant across frequency, as in a Fourier representation. The curve for speech lies intermediate between the other two. The corresponding curves for the temporal envelope necessarily show the same pattern due to time-frequency trade-off (fig.7b). The filters for all three sounds classes show an increase in sharpness with center frequency (fig.7c). All of the curves approximately follow a power law.

Deriving an efficient code for the combined set of environmental sounds and animal vocalizations (fig.3) yields similar bandwidth and sharpness curves. The curves for these filters can be varied from one extreme to the other by changing the relative composition of the two types of sounds in the dataset (unpublished observations). The curves most consistent with with physiological measurements [24] are those for the speech data set and the combined sound ensembles (fig.7d).

One argument for explaining the increase in filter bandwidth with center frequency is based on the observation that the average power spectrum for speech, music, and some natural sounds is approximately $1/f$ [25, 26]. In frequency bandwidths are chosen so that each band has equal average power, the bandwidth must increase exponentially with frequency [27]. The predictions of this model, which assumes a spectral representation, can be obtained from the average power spectra for the three sound classes (fig.8a). The functions of equal power bandwidth versus center frequency (fig.8b) are less in agreement with physiological and psychophysical observations and are different from those derived from efficient coding because a spectral representation ignores temporal structure in the signal.

Discussion

The main insight provided by this analysis is not into the response properties of individual cochlear fibers, but into how sound is encoded by the specific distribution of response properties of the population. That the filters in an efficient code of sound are localized in time-frequency space should perhaps come as no surprise for general classes of sounds. If particular harmonic structures do not dominate the statistics, then one would expect as a consequence localized filters. Aside from the details of the filter form, given localized filters, the only remaining degrees of freedom for the auditory code is the particular trade-off between time and frequency in the population. A common mischaracterization of the peripheral auditory system is that it performs a Fourier analysis of the auditory signal. If this were true, filter bandwidth would remain roughly constant as a function of center frequency, which is not what is observed experimentally. Another approximation of the auditory system is that the filters in the auditory filter bank have constant sharpness as in a wavelet representation. This too is inconsistent

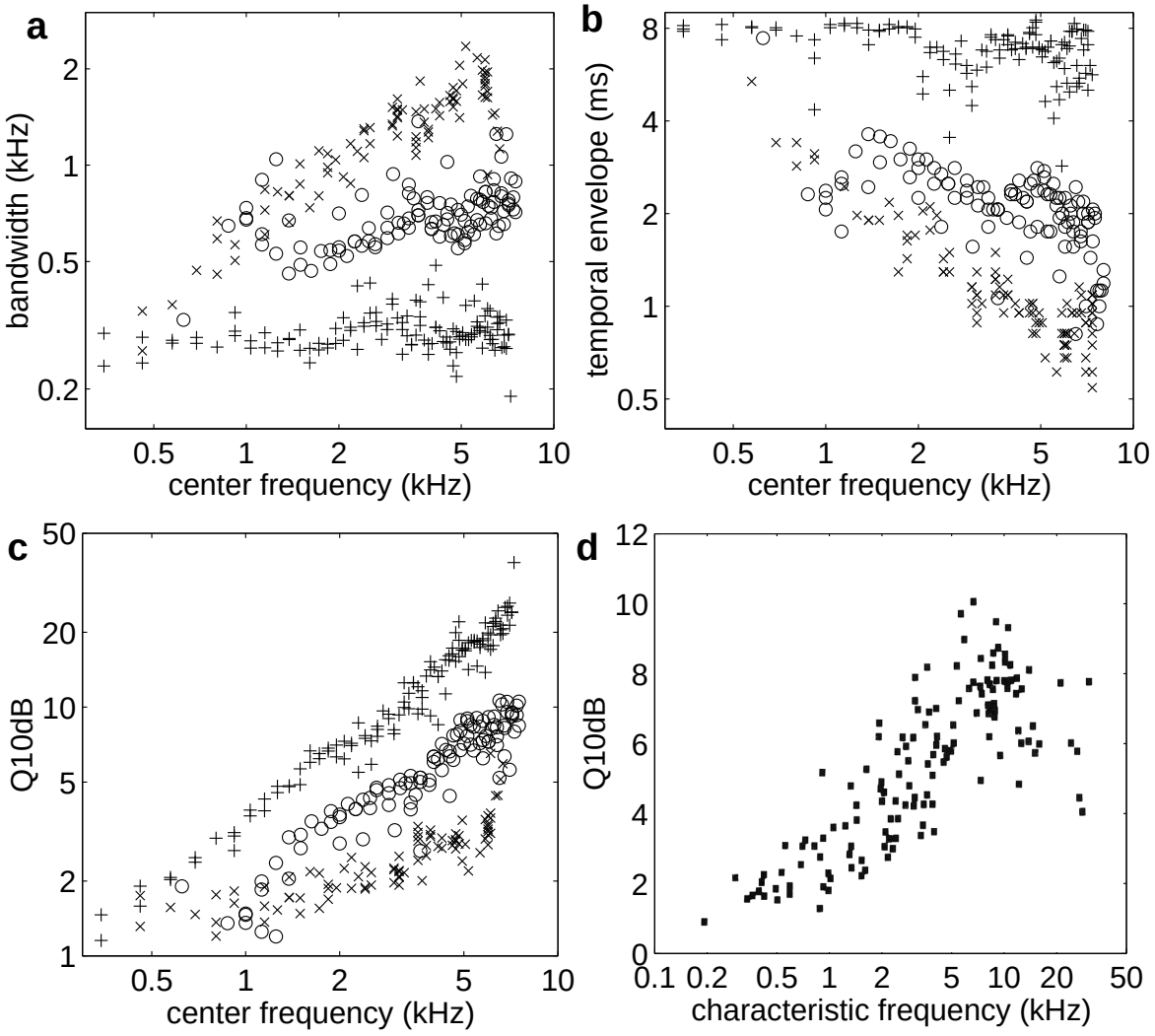


Figure 7: Filter characteristics as a function of center frequency for environmental sounds (×), speech (○), and vocalizations (+). Filter bandwidth (**a**) . Filter temporal envelope width (**b**). Filter sharpness or center frequency divided by bandwidth (Q_{10dB}) (**c**). Q_{10dB} measured from cat auditory nerve fibers (**d**) From [24].

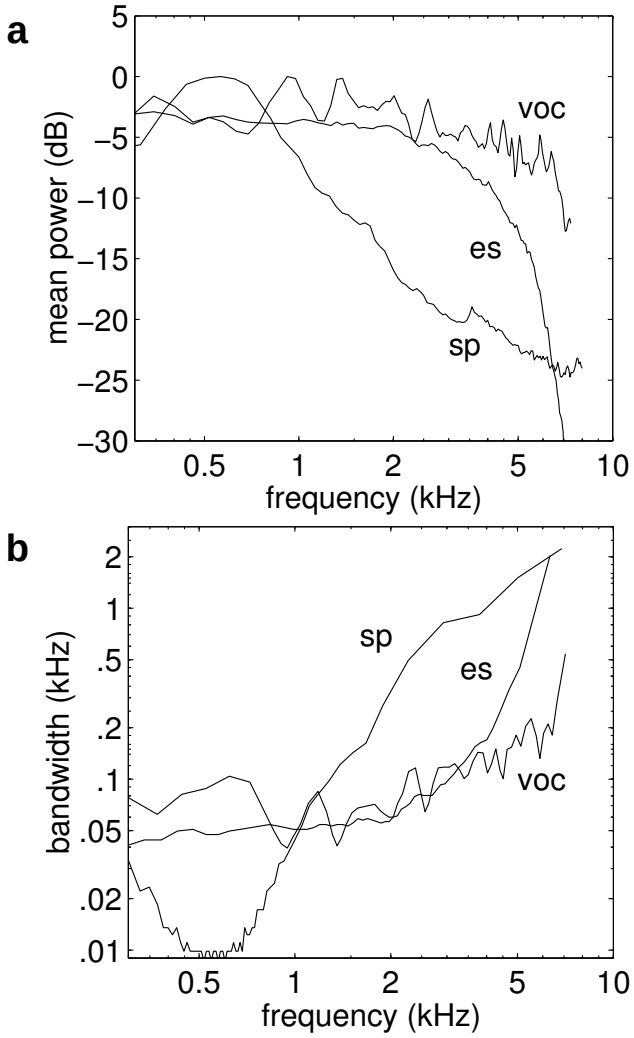


Figure 8: Predictions of bandwidth vs frequency curves assuming equalization of spectral power across bandwidths. (**a**) Average power spectra for environmental sounds (es), vocalizations (voc), and speech (sp). (**b**) The corresponding bandwidth curves chosen so that each spectral band has equal average power.

with the experimental data. The results provide an explanation for the distribution of cochlear tuning properties in terms of efficient coding, and suggest that the auditory system is adapted for the efficient representation of a broad range of natural sounds, including sounds in the natural environment and animal vocalizations.

Linear models of auditory coding based on revcor filters can only account for cochlear nerve fiber responses to stimuli within a limited dynamic range and do not capture effects such as adaptation or two-tone inhibition [4]. The limitations of this simple, but mathematically tractable, linear model used here should also be noted. First, because the form of the model precludes the possibility of modeling dynamic aspects of auditory filters, such as any form of gain control or the broadening for higher stimulus intensities. Second, statistical regularities on a larger time scale cannot be captured, because the model only optimizes the efficiency of the code within the analysis window. Incorporating these factors into a model of auditory periphery would likely increase the efficiency of the sensory code [28–30].

The predicted filter shapes typically do not show the temporal asymmetry of the auditory nerve filters which have a rapid rise followed by a slower decay. One explanation for this is that, for reasons of simplicity, the model was designed to encode only the signal within the analysis window and were not restricted to be causal. Deriving efficient codes for such models is beyond the scope of current methods of ICA. A causal model in which the encoding proceeds forward in time would lead to an asymmetry in the filters because structure preceding the current time would have already been accounted for earlier in the coding. This filters in such a code could presumably capture more closely the structure of the onset of offset envelopes of natural sounds.

The result that the optimal representation changes depending on the class of sounds suggests that similar adaptations might be used by biological systems in different ecological niches. The explanation for the organization of the auditory sensory code suggested here is that the auditory system is adapted for processing a broad class of sounds. If we allow that the human cochlear nerves follow filter bandwidth and sharpness curves similar to those measured physiologically, this analysis also suggests that the acoustic properties of speech make efficient use of the bandwidth available in auditory system. That the efficient representation of speech is nearly identical to that for natural sounds combined with vocalizations could reflect an evolutionary adaptation of speech itself to make maximally efficient use of the coding properties of a prelinguistic auditory system.

These results mirror nicely the results on efficient coding of natural images by demonstrating that efficient coding of natural sounds can explain many of the sensory coding properties of the auditory system. Efficient coding of natural scenes results in a population of localized, oriented Gabor wavelet-like filters [11, 12]. The analogue of this in the auditory system is the gammatone filter, or a gamma-modulated sinusoid. A prevalent form of structure in natural scenes is an edge which can be efficiently

encoded by a population of Gabor filters. Similarly sound onsets or 'acoustic edges' can be efficiently encoded by a population of filters that resembles a gammatone filter bank. In both cases, however, the interpretation offered by the theory is not that these filters as 'edge detectors', but rather that the code is optimized to encode a more general class of patterns efficiently: those with edges and those that vary smoothly. These results lends further support to the hypothesis that efficient coding is a general principle for sensory coding. A challenge that remains in finding workable models that can account for the many types of non-linearities and higher-order processing in perceptual systems.

Methods

Theory

The relationship between efficient coding and statistical density estimation can be seen by applying Shannon's source coding theorem which states that the lower bound the expected code length is the entropy of the data, $H(p) = -\sum p(x) \log p(x)$. The true probability density $p(x)$, however, is unknown and must be approximated by the density $q(x)$ assumed by the model. From this the lower bound on expected code length $l(x)$ becomes

$$\begin{aligned} E[l(x)] &\geq \sum_x p(x) \log \frac{1}{q(x)} \\ &= \sum_x p(x) \log \frac{1}{p(x)} + \sum_x p(x) \log \frac{p(x)}{q(x)} \,. \end{aligned}$$

The right term is the entropy. The term on the left is the Kullback-Leibler divergence between p and q and is zero if and only if $p = q$. Thus, the closer q is to p, the lower the bound on the expected code length.

To obtain an expression for the likelihood of the data under the model (1), the stimulus waveform $\mathbf{x}$ is expressed as a sum of basis functions ϕ_i weighted by coefficients a_i, $\mathbf{x} = \sum_i a_i \phi_i$, where $\mathbf{x} = [x(1), ..., x(N)]$. To obtain the filters h_i, the previous equation is written in matrix form $\mathbf{x} = [\phi_1; ...; \phi_k]\mathbf{a} = \Phi\mathbf{a}$. Then $\mathbf{a} = \Phi^{-1}\mathbf{x}$, yielding the (FIR) filters in the rows of Φ^{-1}. The data likelihood is then $p(\mathbf{x}|\Phi) = p(\mathbf{a})/|\det \Phi|$, as shown by [31–33]

Data

The ensemble of environmental sounds were obtained from a variety of sources and included rustling brush, crunching leaves and twigs, rain, fire, and forest and stream sounds, and was 45 seconds in total duration. Animal vocalizations were selected from a collection of mammalian vocalizations [34]. A representative sample of 44 vocalizations were used with each edited to remove the periods of silence.

Although the collection was recorded so as to isolate the vocalizations, the background sounds of the natural habitat, typically birds and insects, were audible in many of the recordings. Speech was obtained from the TIMIT continuous speech corpus using 100 male and female speakers. To reduce the amount of computation required, the environmental and animal vocalization stereo sounds were converted to mono and downsampled from 44.1 kHz to 14.7 kHz. Speech was used at the original sampling frequency of 16 kHz. All waveforms where highpass filtered using a cutoff equal to the sampling frequency divided by the window size, which was 115 Hz for natural sounds and animal vocalizations and 125 Hz for speech. Datasets were then constructed with 128-sample segments, randomly selected from the sound ensembles.

Algorithm

The details of the basic algorithm to derive efficient codes have been given previously [12,14]. We briefly summarize them here. The basis matrix Φ (or equivalently the filters) are optimized by maximizing the likelihood of the data ensemble under the model

$$\Delta\Phi \propto \Phi\Phi^T \frac{\partial}{\partial\Phi} \log p(\mathbf{x}|\Phi) = \Phi(\mathbf{I} - \varphi(\mathbf{a})\mathbf{a}^T)\,. \tag{2}$$

where the prefactor $\Phi\Phi^T$ is used for faster convergence [35], and $\varphi(\mathbf{a}) = (\log p(\mathbf{a}))'$. Independence of the coefficients is assumed, i.e. $p(\mathbf{a}) = \prod_i p(a_i)$. The distribution $p(a_i)$ is typically fixed a priori, but here we fit this distribution to the data using a generalized Gaussian [36–38], $p(a) \propto \exp(-|a|^q)$. This yields $\varphi(a) = -\theta|a-\mu|^{q-1}qc\sigma^{-q}$, where $\theta = \text{sign}(s)$ and $c = [\Gamma(3/q)/\Gamma(1/q)]^{q/2}$ (subscripts omitted for clarity). By inferring the maximum a posteriori value of q_i from the data, the model can fit a broad range of statistical distributions, including those assumed by both principal and independent component analysis, and was very well matched to the distributions observed here. It should be noted that the learning rule is not intended to be biological plausible but simply a method for deriving the theoretical predictions of efficient coding for the ensembles of natural sounds.

In all the experiments presented here, five optimizations were performed from different random initial conditions of Φ, all yielding qualitatively similar results. During optimization each gradient step was estimated using a block of 640 waveform segments. The gradient stepsize was reduced linearly from 10^{-1} to 10^{-5} over 10,000 iterations and then optimized further at the final stepsize for an additional 10,000 iterations. Each generalized Gaussian parameter q_i for the output distribution $p(a_i)$ was estimated throughout optimization, every 20,000 patterns. Instability in the gradient can result from small values q_i, so the contribution of the q_i to the gradient was limited to a minimum of 2/3 during optimization.

Time-frequency analysis

To determine the extent of a filter in time-frequency space, the temporal extent was measured using the width required to cover 95% of the filter power. Frequency width was measured using the spectral bandwidth at 10 dB down from the peak. Filters that were not localized (where the main spectral peak at accounted than 50% of total power) were omitted from the plot and totaled of seven, one, and zero out of 128 filters for environmental sounds, vocalizations, and speech, respectively.

Acknowledgements

The author thanks Carl Olson, Bruno Olshausen, and Lori Holt for helpful discussions and feedback on the manuscript.

References

1. Barlow, H. B. Possible principles underlying the transformation of sensory messages. In Sensory Communication, Rosenbluth, W. A., editor, 217–234. MIT Press, Cambridge (1961).

2. Kiang, N. Y.-S., Watanabe, T., Thomas, E. C., and Clark, L. F. *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve.* MIT Press, Cambridge, (1965).

3. Evans, E. F. Frequency selectivity at high signal levels of single units in cochlear nerve and nucleus. In Psychophysics and physiology of hearing, Evans, E. F. and Wilson, J. P., editors, 185–192. Academic Press, New York (1977).

4. de Boer, E. and de Jongh, H. R. On cochlear encoding: Potentialities and limitations of the reverse-correlation technique. *J. Acoust. Soc. Am.* **63**(1), 115–135 (1978).

5. Carney, L. H. and Yin, T. C. T. Temporal coding of resonances by low-frequency auditory nerve fibers: single-fiber responses and a population model. *J. Neurophys.* **60**, 1653–1677 (1988).

6. Field, D. J. Relations between the satistics of natural images and the response properties of cortical cells. *J. Optical Soc. Am. A* **4**(12), 2379–2394 (1987).

7. Field, D. J. What is the goal of sensory coding? *Neural Comp.* **6**(4), 559–601 (1994).

8. Linsker, R. Perceptual neural organization - some approaches based on network models and information-theory. *Ann. Rev. Neuro.* **13**, 257–281 (1990).

9. Atick, J. J. Could information-theory provide an ecological theory of sensory processing. *Network: Comp. Neural Sys.* **3**(2), 213–251 (1992).

10. Rieke, F., Bodnar, D. A., and Bialek, W. Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proc. R. Soc. Lond. Ser. B-Biol. Sci.* **262**, 259–265 (1995).

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

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

13. van Hateren, J. H. and Ruderman, D. L. Independent component analysis of natural images sequences yield spatiotemporal filters similar to simple cells in primary visual cortex. *Proc. Royal Soc. Lond. B* **265**, 2315–2320 (1998).

14. Lewicki, M. S. and Olshausen, B. A. A probabilistic framework for the adaptation and comparison of image codes. *J. Opt. Soc. Am. A* **16**(7), 1587–1601 (1999).

15. Comon, P. Independent component analysis, a new concept? *Signal Processing* **36**(3), 287–314 (1994).

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

17. Laughlin, S. B. Matching coding to scenes to enhance coding efficiency. In Physical and Biological Processing of Images, Braddick, O. J. and Sleigh, A. C., editors, 42–72. Springer, Berlin (1983).

18. Bell, A. J. and Sejnowski, T. J. Learning the higher-order structure of a natural sound. *Network Computation in Neural Systems* **7**(2), 261–267 (1996).

19. Irino, T. and Patterson, R. D. A time-domain, level-dependent auditory filter: The gammachirp. *J. Acoust. Soc. Am.* **101**(1), 412–419 (1997).

20. Zoharian, A. S. and Rothenberg, M. Principal-component analysis for low redundancy encoding of speech spectra. *J. Acoust. Soc. Am.* **69**, 832–845 (1981).

21. Mallat, S. *A Wavelet Tour of Signal Processing.* Academic Press, second edition edition, (1999).

22. Lewicki, M. S. and Sejnowski, T. J. Learning overcomplete representations. *Neural Computation* **12**(2), 337–365 (2000).

23. Moore, B. C. J., editor. *Frequency Selectivity in Hearing*. Academic Press, (1986).

24. Evans, E. F. Cochlear nerve and cochlear nucleus. In Handbook of sensory physiology, Keidel, W. D. and Neff, W. D., editors, volume 5/2, 1–108. Springer, Berlin (1975).

25. Voss, R. F. and Clarke, J. 1/f noise in music and speech. *Nature* **258**, 317–318 (1977).

26. Attias, H. and Schreiner, C. Low-order temporal statistics of natural sounds. In *Advances in Neural and Information Processing Systems*, volume 9 (Morgan Kaufmann, San Mateo, 1997).

27. Furth, P. M. and Andreou, A. G. A design framework for low power analog filter banks. *IEEE Trans. Circuits Systems I* **42**(11), 966–971 (1995).

28. Lewicki, M. S. and Sejnowski, T. J. Coding time-varying signals using sparse, shift-invariant representations. In *Advances in Neural Information Processing Systems*, volume 11, 730–736. MIT Press, (1999).

29. Brenner, N., Bialek, W., and van Steveninck, R. D. Adaptive rescaling maximizes information transmission. *Neuron* **26**, 695–702 (2000).

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

31. Pearlmutter, B. A. and Parra, L. C. A context-senstive generalization of ICA. In *International Conference on Neural Information Processing*, 151–157, (1996).

32. MacKay, D. J. C. Maximum likelihood and covariant algorithms for independent component analysis. University of Cambridge, Cavendish Laboratory. Available at `ftp://wol.ra.phy.cam.ac.uk/pub/mackay/ica.ps.gz` (1996).

33. Cardoso, J.-F. Infomax and maximum likelihood for blind source separation. *IEEE Signal Processing Letters* **4**, 109–111 (1997).

34. Emmons, L. H., Whitney, B. M., and Ross, D. L. Sounds of the neotropical rainforest mammals. Library of Natural Sounds, Cornell Laboratory of Ornithology, (1997). Audio CD.

35. Amari, S., Cichocki, A., and Yang, H. H. A new learning algorithm for blind signal separation. In *Advances in Neural and Information Processing Systems*, volume 8, 757–763 (Morgan Kaufmann, San Mateo, 1996).

36. Box, G. E. P. and Tiao, G. C. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA, (1973).

37. Mallat, S. G. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE PAMI* **11**, 674–693 (1989).

38. Simoncelli, E. P. and Adelson, E. H. Noise removal via Bayesian wavelet coring. In *3rd IEEE Int'l Conf on Image Processing, Laussanne Switzerland*, , Sept. (1996).