

Efficient coding of natural sounds

Introduction

Many previous studies in auditory neurophysiology have used simple tonal stimuli to understand how neurons encode sound. While these studies have shed light on many important neural characteristics, pure tones do not typically occur in our environment. Instead, we are often surrounded by multiple sound sources with complex harmonic and transient components such as speech, environmental sounds, animal vocalizations, and background noise. These natural sounds only make up a small subset of the sample space of all possible acoustic stimuli, yet they still consist of a wide range of spectral and temporal structures. It is reasonable to hypothesize that our brains have evolved to optimally process these naturally occurring sounds in order to efficiently extract relevant acoustic cues. In recent years, an increasing number of modeling and physiological studies have used natural or natural-like stimuli to explore the validity of this prediction. With this better understanding of natural sound statistics and methods of decomposing their signals, future studies can create synthetic stimuli with similar statistics to investigate (and possibly differentiate) between natural and “naturalistic” sounds.

This brief overview of the coding of natural sounds suggests three papers for discussion. The first demonstrates that the statistics of natural sounds are redundant in the peripheral auditory system representation (Attias and Schreiner 1997). The remaining two studies implement two different ways of decomposing a sound signal. One study shows that a Fourier analysis may be sufficient for animal vocalizations, while wavelet transforms are optimal for encoding speech and environmental sounds (Lewicki 2002a). The other uses modulation spectra to encode natural stimuli and demonstrates differences between groups of natural sound ensembles (Singh and Theunissen 2003). It is also useful to interpret the findings of these studies in terms of neural responses to natural sounds. Two physiological studies in zebra finches and grasshoppers are also proposed for further reading (Hsu et al 2004, Machens et al 2005).

Information theory and sensory neural systems

Information theory was first introduced by Shannon in 1948 and provided a model for representing reliable data transfer communication systems. The essential aspects of information theory lie in source coding (which defines entropy as the least number of bits required to represent a piece of information) and channel coding (which defines channel capacity as the maximum allowable rate of information transfer). Coding theory looks for ways to increase the efficiency while reducing the error of data communication.

In 1961, Barlow applied these principles to model the behavior of neurons along the sensory pathways. Specifically, he wanted to understand how visual and audio information was processed in the brain. His efficient coding hypothesis proposed that the spiking activity of neural populations was optimized to best represent images and sounds that occur in our natural environment. He further predicted that one of the roles of early processing would be to reduce the redundancy of the represented information. Statistical independence across channels (or

neurons) would allow the efficient encoding of as much information about the stimulus as possible (Field 1987 1994, Linsker 1990, Atick 1992).

Barlow's predictions have been largely confirmed in the early stages of visual processing. Responses of neurons in the peripheral visual pathway are consistent with an optimal-code prediction (Atick 1992, Dan et al 1996, Olshausen and Field 1996, Bell and Sejnowski 1997, van Hateren and Ruderman 1998, Lewicki and Olshausen 1999). Many studies suggest that the visual system has been designed to exploit the statistics of natural images in order to maximize the efficiency of the neural representation of these visual scenes to the brain. There is an increasing amount of evidence for an analogy to be made for the auditory system. For example, the auditory nerve is better able to code natural sounds compared to white noise (Rieke 1995).

Redundant representation of natural sounds in the periphery

To model the peripheral auditory system, Attias and Schreiner (1997) passed a sound stimulus $s(t)$ through a set of overlapping bandpass filters, resulting in a set of band-limited signals $s_v(t) = x(t)\cos(v_t + \phi(t))$, where v denotes the center frequency of the filter. They measured the amount of redundancy in the information available in adjacent filters by looking at the low-order statistical properties of the amplitude ($x(t)$) and phase ($\phi(t)$) of the output signals.

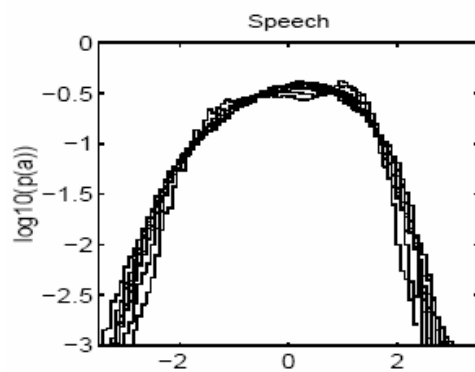


Figure 1: Amplitude probability distributions across the set of cochlear filters for speech.

Figure 1 shows the amplitude probability distributions across the filter set for human speech. The statistics are nearly identical across the filters. Similar distributions resulted for different sound types, including music, cat vocalizations, and environmental sounds. Increasing the bandwidths of the filters did not change the distributions, and the autocorrelation of $x(t)$ at different temporal resolutions also resulted in nearly identical distributions. These results suggest that natural sounds have certain statistical properties that distinguish themselves from other acoustic stimuli. Specifically, the last observation suggests bandwidth invariance may be associated with natural sounds. Furthermore, the information is highly redundant across the filters, suggesting translation invariance across the cochlear axis.

Optimal code requirements

An efficient code for representing signals will reduce redundancy and represent only the desired information. Traditional representations of signals have mostly used block-based methods, where the signal is broken down into a set of discrete blocks. For sounds with transient cues such as speech, using this form of representation may obscure the cue by causing it to depend on the weighting and the length of the blocks. Furthermore, temporal shifts in the signal can lead to very different representations. Using many short blocks partially mitigates this effect

but causes a decrease in computational efficiency. An optimal code for processing sounds needs to be both time shift-invariant and efficient (Smith and Lewicki 2005).

Blind source separation separates a set of signals such that the new set of signals has maximal statistical independence. Speech signal coding has primarily used principal component analysis to reduce a multi-dimensional (possibly correlated) data set into a small set of uncorrelated variables (Zoharian and Rothenbert 1981). However, extraction of the principal components of environmental sounds was largely unsuccessful in temporally localizing the transient sounds. In contrast, Lewicki (2002a) showed an independent component analysis, where the mutual statistical independence of the signals is assumed, can result in filter shapes that are localized in both frequency and time.

It is common to interpret the peripheral auditory system as a Fourier analyzer. However, the sharpness in auditory nerve fiber tuning is not constant across frequency. This may suggest that the distribution of cochlear tuning is actually optimized for coding natural sounds efficiently. Figure 3 illustrates the overall filter shapes for Fourier and wavelet analysis and the derived optimal filter shapes for natural sounds. The figure suggests that the Fourier transform, which gives no temporal localization information, could be optimal for efficiently coding animal vocalizations. However, a wavelet transform, which provides some temporal resolution in exchange for some frequency resolution, would be optimal for the coding of environmental sounds and human speech.

Representing the sound pressure waveform as a sum of kernel functions

A signal $x(t)$ can be decomposed into a set of weighted independent kernel functions ϕ_1 to ϕ_M (Lewicki and Sejnowski 1999, Lewicki 2002b), which are arbitrarily scaled and positioned in time such that $x(t)$ can take on any shape.

$$x(t) = \sum_{m=1}^M \sum_{i=1}^{n_m} s_i^m \phi_m(t - \tau_i^m) + \varepsilon(t)$$

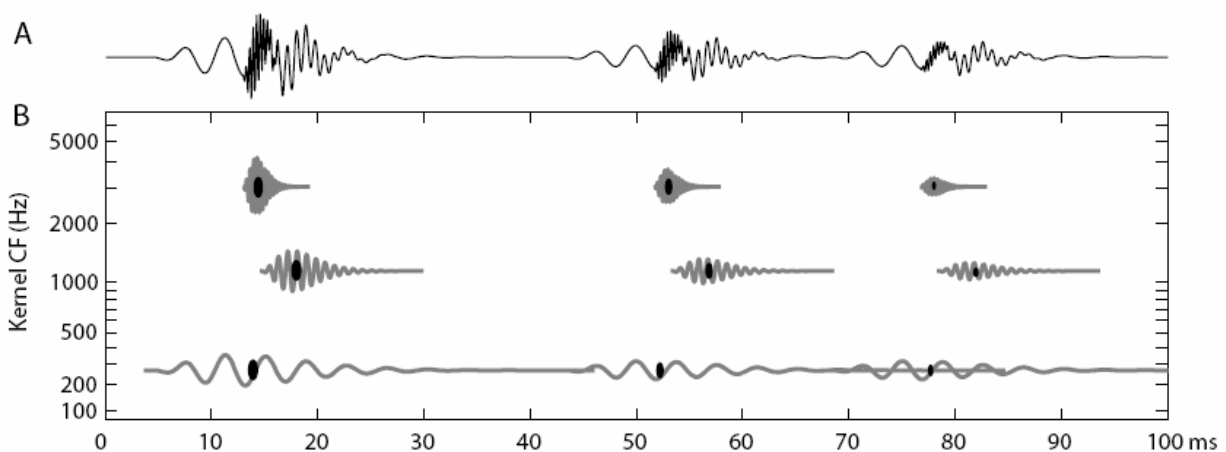


Figure 2: Illustration of signal decomposition into kernels. Black ovals indicate amplitude and spectral and temporal position of each of the nine components, and gray waveforms are their corresponding gammatone kernel functions.

The kernel functions are gammatone functions, which are commonly used to model the cochlear filters. The set of weights s_i and time shifts τ which minimize the error $\varepsilon(t)$ maximizes the

efficiency of the representation and forms the optimal code for the sound. Figure 2 illustrates a sparse (only three kernels) spike code (spikegram) of three chirps, which have the same spectral and temporal positions but different individual component amplitudes. Unlike a spectrogram representation, which represents each point on the frequency-time space as an amplitude (or pixel shade), decomposing the signal in this format retains the phase information of the stimulus.

Divisive normalization is a similar method used to reduce redundancy. Filter responses are each divided by a weighted sum of all other filter responses. This method has been demonstrated to work well in the visual system (Ruderman and Bialek 1994, Simoncelli and Schwartz 1998, Wainwright et al 2001). Schwartz and Simoncelli (2000) applied divisive normalization to model filter responses to groups of natural sounds. Their model was able to account for nonlinearities in the rate-level functions of two-tone suppression data and frequency tuning curves of auditory nerve fibers.

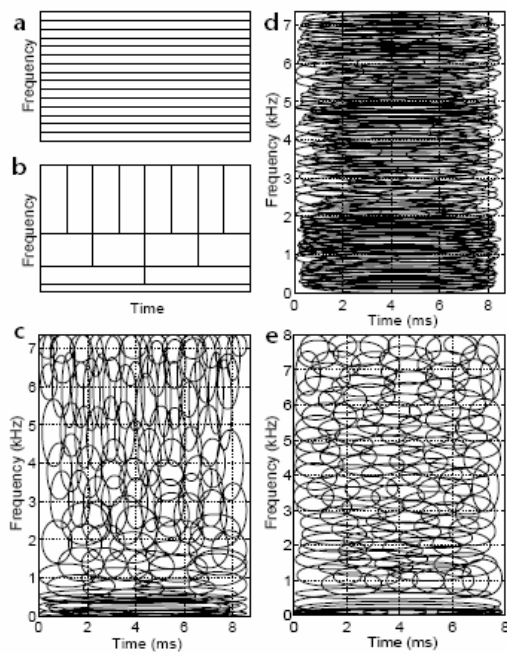


Figure 3: (a) Filters in Fourier transform; (b) Wavelet filters; (c-e) optimal filter shapes for (c) environmental sounds, (d) animal vocalizations, and (e) speech.

Representing the signal as a sum of weighted ripple components (modulation spectra)

Singh and Theunissen (2003) represented the spectrograms of natural sounds as the sum of weighted independent ripple components (where the direction and frequency of the ripple is mapped to a point in the frequency modulation – temporal modulation space). Furthermore, the relative weights of each ripple component can be expressed on this modulation space, resulting in a modulation spectrum.

Figure 4 shows the contour of a white noise modulation spectrum, which is essentially a representation of the shape and bandwidth of the filters in the ripple components used. A lot of the energy in the original stimulus has thus been filtered out. In contrast, natural sounds should have spectral and temporal structures that modulate on these frequency and time scales, such that most of the energy would be represented in their modulation spectra. Figure 4 shows that the spectra for natural sounds have a “+” shape, indicating that these stimuli do not have rapid temporal and spectral modulations at the same time. Furthermore, songs and speech have a lot of

high spectral modulation occurring at low temporal modulation, while the environmental sounds have more oval contours, similar to that found for white noise. These results provide insight into how to choose the appropriate time-frequency scales for decomposing different sounds for preprocessing strategies necessary for hearing aids or cochlear implants.

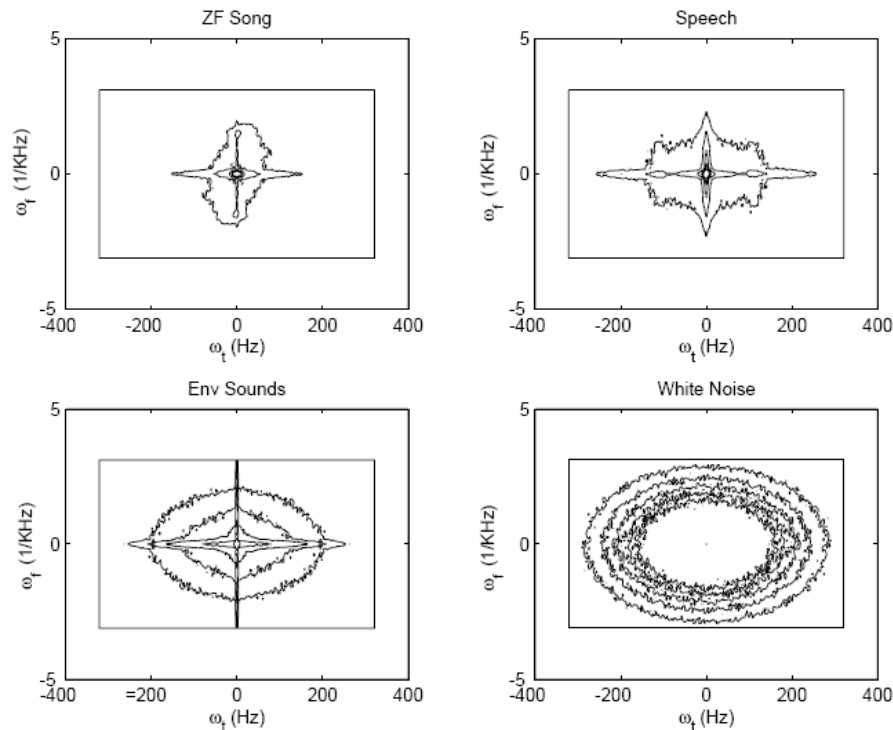


Figure 4: Modulation spectra of three sets of natural sounds and white noise.

Neural responses to natural sounds

The discussion thus far as mostly consisted of analyzing the statistical properties of natural sounds. A number of physiological studies have recorded neural responses to natural or naturalistic stimuli. In particular, several studies have analyzed neural responses of zebra finches to stimulus ensembles consisting of songs from the same species. For example, a hierarchical study demonstrated increasing selectivity for the natural songs (as opposed to synthesized songs with similar spectral-temporal modulations) along the ascending auditory pathway (Hsu et al 2004). Another study found that their auditory central neurons carry information in their phase locking to the stimulus or modulation rate, as well as in their temporal spiking patterns (Wright et al).

While it seems appropriate to think of the auditory system as optimal for efficient coding of sounds that are in our natural environment, perhaps our neural coding strategies are also affected by the relative importance of a sound. For example, the optimal stimulus set for the auditory neurons in grasshoppers does not directly coincide with sounds in their natural environment. Instead, the neurons appear to be optimized for coding a subset of these natural sounds which are behaviorally relevant (Machens et al 2005).

References

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

- **Attias H., Schreiner C. E. (1997). Temporal low-order statistics of natural sounds. *Adv. Neural Info. Process. Syst.* 9:27-33.
- Barlow H. B. (1961). Possible principles underlying the transformation of sensory messages. In *Sensory Communication*. MIT Press, Cambridge MA.
- Bell A. J., Sejnowski T. J. (1997). The ‘independent components’ of natural scenes are edge filters. *Vision Res.* 37:3327-3338.
- Dan Y., Atick J.J., Reid R. C. (1996). Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *J. Neurosci.* 16:3351-3362.
- Field D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *J. Optical Soc. Am.* 12:2379-2394.
- Field D. J. (1994). What is the goal of sensory coding? *Neural Comp.* 6:559-601.
- ***Hsu A., Woolley S. M. N., Fremouw T. E., Theunissen F. E., Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons. *J. Neurosci.* 24:9201-9211.
- Lewicki M. S., Olshausen B. A. (1999). A probabilistic framework for the adaptation and comparison of image codes. *J. Opt. Soc. Am.* 16:1587-1601.
- Lewicki M. S., Sejnowski T. J. (1999). Coding time-varying signals using sparse, shift-invariant representations. In *Advances in neural information processing systems*, 11. MIT Press, Cambridge MA.
- **Lewicki M. S. (2002a). Efficient coding of natural sounds. *Nature Neurosci.* 4:356-363.
- Lewicki M. S. (2002b). Efficient coding of time-varying patterns using a spiking population code. In *Probabilistic models of the brain: Perception and neural function*. MIT Press. Cambridge, MA.
- Linsker R. (1990). Perceptual neural organization – some approaches based on network models and information theory. *Annu. Rev. Neuro.* 13:257-281.
- ***Machens C. K., Gollisch T., Kolesnikova O., Herz A. V. M. (2005). Testing the efficiency of sensory coding with optimal stimulus ensembles. *Neuron.* 47:447-456.
- Olshausen B. A., Field D. J. (1996). Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. *Nature.* 381:607-609.
- Rieke F., Bodnar D. A., Bialek W. (1995). Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proc. R. Soc. London. Ser. B* 262:259-265.
- Ruderman D. L., Bialek W. (1994). Statistics of natural images: Scaling in the woods. *Phys. Rev. Letters.* 73:814-817.
- ***Schwartz O., Simoncelli E. P. (2000). Natural sound statistics and divisive normalization in the auditory system. *Adv. Neural Info. Proc. Syst.* MIT Press. Cambridge, MA.
- Shannon C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* 27:379-423, 623-656.
- Simoncelli E. P., Schwartz O. (1998). Image statistics and cortical normalization models. In *Adv. Neural Information Processing Systems*. MIT Press. Cambridge, MA.
- **Singh N. C., Theunissen F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *J. Acoust. Soc. Am.* 114:3394-3411.
- *Smith E., Lewicki M. S. (2005). Efficient coding of time-relative structure using spikes. *Neural Comp.* 17:19-45.

- van Hateren J. H., Ruderman D. L. (1998). Independent component analysis of natural image sequences yield spatiotemporal filters similar to simple cells in primary visual cortex. *Proc. R. Soc. Lond. B Biol. Sci.* 265:2315-2320.
- Wainwright M. J., Schwartz O., Simoncelli E. P. (2001). Natural image statistics and divisive normalization: Modeling nonlinearity and adaptation in cortical neurons. In *Statistical theories of the Brain*. MIT Press. Cambridge, MA.
- Wright B. D., Sen K., Bialek W., Doupe A. J. (2002). Spike timing and the coding of naturalistic sounds in a central auditory area of songbirds. In *Advances in Neural Information Processing Systems 15*. MIT Press. Cambridge, MA.
- Zoharian A. S., Rothenbert M. (1981). Principle component analysis for low redundancy encoding of speech spectra. *J. Acoust. Soc. Am.* 69:832-845.

- * suggested for background reading
- ** suggested for discussion
- *** suggested for further reading