Title
Invariant Recognition of Vocal Features

Author
Moore, Richard Channing

Publication Date
2011

Peer reviewed|Thesis/dissertation

Invariant Recognition of Vocal Features

By

Richard Channing Moore, III

A dissertation submitted in partial satisfaction
of the requirements for the degree of
Doctor of Philosophy
in
Biophysics
in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:

Professor Frédéric E. Theunissen, Chair
Professor Jack L. Gallant
Professor Michael R. DeWeese
Professor Fritz T. Sommer

Fall 2011

Abstract

Invariant Recognition of Vocal Features

by

Richard Channing Moore, III

Doctor of Philosophy in Biophysics

University of California, Berkeley

Professor Frédéric E. Theunissen, Chair

Animals and humans are able to communicate vocally in very challenging acoustic conditions. Background noise, especially from other individuals of the same species, may mask the relevant signal and propagation can introduce significant distortions to the sound waveform. While our brains are able to extract meaningful information from heavily degraded communication sounds, the mechanisms by which the auditory system performs this task are not well understood. This thesis shows how neural systems can and do handle signal degradations. by examining how auditory neurons in an animal model of communication, the Zebra Finch *Taeniopygia guttata*, process degraded and undegraded signals, and demonstrates that these principles can be used to perform noise reduction on voice recordings. I discuss how the notion of *invariance*, common in studies of sensory perception for roughly a century, has more recently been helpful in the analysis of sensory systems at the neural level. I discuss how to characterize the invariant properties of vocal sounds, and how to connect this analysis to the mathematical theory of invariants.

To characterize invariance at the neural level, I construct a novel metric using spike-train cross-correlation between neural responses to the same signal obtained under various conditions. Using this measure, I show that a subset of neurons in avian secondary auditory forebrain area NCM can extract a representation of birdsong that is robust to background noise interference. Spectro-temporal receptive field (STRF) and modulation transfer function (MTF) analysis show that these invariant neurons are sensitive to slowly changing pitch features. Then, using stimuli that have been degraded systematically along spectral and temporal features, I further characterize the nature and origin of invariant response properties in neurons throughout avian auditory forebrain. The response of auditory neurons to spectral degradation is well explained by their MTF, but results in the temporal domain show that some neurons show invariance properties beyond those expected from this model.

Finally, I use the insights from these experiments to construct a noise-reduction algorithm that can be implemented in real-time on digital systems. The system performs well when compared to state-of-the-art algorithms for noise reduction and I discuss how these systems interrelate in terms of processing the statistics of vocal sounds. Using these comparisons and interpretations, I show how we might improve the performance of such noise reduction algorithms.

To my loving partner Jordan, for supporting me throughout this process;

and to my parents, for their love and for always encouraging my intellectual pursuits.

Table of Contents

Acknowlegements

Thanks are due to many, many people for helping me get this far.

First and foremost, to my advisor, Frédéric, who has guided this project, who helped me through good times and bad, and whose insight sent me chasing after the different notions of invariance that I cover in the Introduction.

I also owe a great deal to my committee: Jack Gallant, Mike DeWeese, and Fritz Sommer, who have all helped me understand my results and hone my analyses.

Every one in the Theunissen lab has helped me in some way, but particularly Yuka Minton, who has kept this running and kept us in birds and reagents. Also especially to Tyler Lee, who has been a great help both with implementing and with understanding the noise reduction algorithm described in Chapter 3.

To the members of the Gallant lab who have provided some wonderful ideas… and done a lot of work on the computer infrastructure that analyzed all of my data.

Introduction

One of the unifying themes in my thesis is how neurons construct representations of the stimulus that are robust to degradations. Equivalently, this can be formulated as an investigation of how these neurons are invariant to stimulus transformations. This introduction is intended to provide a framework in which to consider this problem: I will provide an overview of what is currently known about invariant sensory representations both at the perceptual level and at the neural level, and I will show how a theoretical treatment of invariance could help us understand sensory encoding. In chapter 1, I will analyze how neurons in higher areas of the avian auditory system produce invariant representations of important communication signals in spite of background noise. In chapter 2, I will investigate how neurons process systematic degradations of the spectral and temporal characteristics of sounds. In chapter 3, I will use the insights about vocal signal processing from these experiments to build a noise-reduction algorithm.

One of the most striking features in sensory perception is that of invariance, in which the underlying stimulus may change substantially, but the percept changes little. This manifests in different ways in behavioral and neural experiments, but a basic formulation involves differing stimuli being classed as similar. In some cases, it can mean that the different inputs are in fact indiscriminable. These two options form a positive and a negative definition, respectively. In behavioral experiments, the positive sense means that while the stimuli can be classed as same, information is still available to discriminate among classes and that this is the discrimination that is behaviorally relevant. The negative sense in these cases means that the information has been lost and is no longer available.

Invariance is directly related to object recognition. In order to produce persistent percepts of physical objects in the world around us, neural representations at some level must be able to disregard variations in the raw sensory input. In his monograph on the concept of an "auditory object", Griffiths describes invariance as "the abstraction of sensory information so that information about an object can be generalized between particular sensory experiences in any one sensory domain," (Griffiths & Warren, 2004). This separation—foregrounding—of objects against the scene or background inherently requires an invariant representation of the world. Higher-level sensory systems create this sort of representation by grouping or binding together information from lower levels. Paradoxically, these lower level systems must encode the stimuli in a highly variable fashion in order for the higher level systems to produce an invariant representation (Okada et al., 2010). Experimental evidence of invariant representations has provided important information about the function of higher-level sensory processing areas in vision and, to a lesser extent, in audition.

The presence of invariance places strong constraints on the underlying computations. In this work, I will use the root word "invariant" or "invariance" to refer to two related concepts. The first is the *invariant*: this is a percept that remains unchanged despite changes in the sensory information available to the observer. The second is the *transformation*, always preceded by the words "invariant under" or "invariant to": these are the specific changes made to the stimuli, in spite of which the percept or response remains unchanged. So were I to claim that, "face identity is invariant under rotation," *face identity* would be the invariant percept, and *rotation* the transformation that does not change the perceived identity.

Historical context

Cutting (Cutting, 1983) traces the origins of the term to algebraic and mechanical analysis in the early 19th Century, and its formalization in the mathematics of group theory in the late 19th Century. He puts the use of the term in the psychology of perception a bit later, when early English translations of Helmholtz' work on perception rendered certain observations about constancy using the word *invariance*. Helmholtz is credited with one statement of the conservation of energy, and it is conceivable that he saw some parallels between physical and perceptual invariances. It seems unlikely to me that he made a direct connection with the mathematics: Noether's beautiful proof that conservation laws in mechanics are the direct result of group-theoretic invariance came more than two decades after his death. It is in principle conceivable that Helmholtz would have connected his observations in perception with mathematical formalisms like algebraic invariances, and this would be corroborated if he used the same words in perception that contemporary mathematicians and physicists employed in their work. Such an exact history of the term in Helmholtz' work is a separate project for a historian of science, and certainly one with a better grasp of German than I possess.

Invariance in behavior: human psychophysics and perception

In any event, the word invariance certainly had some use in psychology in the early 20th Century: the word appears in Koffka's monograph *Principles of Gestalt Psychology*, written in English and first published in 1935. Such modern use of the term primarily relates to our experience of certain percepts remaining the same under drastically different conditions. One of the classical invariances in vision, the perception of absolute size, is described by Boring, who points out that while on the one hand railroad tracks appear to converge as they recede into the distance, on the other hand we retain a strong sense that the distance between them is constant. By the same token, we perceive that a person's height and build remain unchanged as they walk away from us, even as they appear smaller (Boring, 1952).

Many of these consistent percepts correspond to things we otherwise know to remain the same: objects, people in our life; or from clearly changing circumstances, e.g. rotations or illumination in vision. Because this thesis is focused on auditory processing, I will start by reviewing invariance in auditory perception. The literature here is not broad, but we can identify a few important invariants: pitch, intensity, physical object, speaker, and semantic content.

Invariance in auditory perception

One of the most studied invariants in audition is pitch. To some degree the existence of the strange negatively-defined percept of timbre is a consequence of just how strong the percept of pitch is. Humans and some animals (Cynx, Williams, & Nottebohm, 1990; Lohr & Dooling, 1998) can order tones according to whether they are higher or lower, even when the tones are as dissimilar as a bandpass noise burst and a note from a flute. Another strong piece of evidence for invariant perception of pitch is the illusion of the missing fundamental. For the most part, pitch maps well to the axis of fundamental frequency; but if sufficient harmonic information is present, the fundamental itself can in fact be completely absent without altering the apparent pitch for humans and some animals (Cynx & Shapiro, 1986).

Humans are able to perceive small increments of relative loudness over a range exceeding 115dB (Viemeister & Bacon, 1988), indicating that relative level is independent of absolute level, and this level-invariant percept of relative loudness exists across a wide range of

frequencies (Buus & Florentine, 1991). This ability to hear, and recognize sounds, over a large range of loudnesses suggests that invariant neural representation must exist at some point in the auditory processing hierarchy (Sadagopan & X. Wang, 2008).

Complex stimuli like music and language exhibit other invariances. Listeners can extract speaker-invariant meaning from a word, phoneme, or sentence that is uttered by different people (Aulanko, Hari, Lounasmaa, Näätänen, & Sams, 1993; Blumstein & Stevens, 1979). Conversely, listeners can perform content-invariant identification of the speaker across different utterances. In music, we see a representation of relative pitch that is invariant to absolute pitch. Listeners can identify a melody despite its being shifted an octave or transposed into a different key (Bharucha & Mencl, 1996; Paavilainen, Jaramillo, Naatanen, & Winkler, 1999), and can group instruments despite differing listening conditions and musical pieces.

Auditory perception of innate properties is not limited to musical instruments: while even untrained human listeners can group pieces of music by the instrument that produced them, a similar result applies for objects in general. Listeners can identify qualities like hollowness, material, shape, and size from impact sounds over a range of distances, constructing a distance-invariant representation of physical properties of the object (Lutfi, 2007).

One interesting special case occurs when a sound has a substantially fractal quality. This condition holds, for instance, for many textured natural sounds like running water, and can be created in synthetic sounds. In this case, the stimulus will be perceived as similar even when played back at a different rate, even across the range from ¼ to 4x the original recording speed (Geffen, Gervain, Werker, & Magnasco, 2011).

Invariance in visual perception

The original sense in which invariance was been investigated was vision. While only a few of the psychophysical invariances in audition have been observed at the neural level, many have been observed and investigated in vision.

An early study of invariance in vision focused on invariant perception of size. Observers can, to a degree, estimate the absolute size of objects and people at a distance, across a wide range of retinal image sizes, creating a "size-distance" invariance (Boring, 1952). Absolute size is perceived separately from the actual retinal size, though proportionality is not strictly preserved: to some degree, observers will actually overstimate the size of an object as it grows more distant (Gilinsky, 1955). This effect plays some a role, for instance, in making the moon look very large on the horizon and small at its zenith in the sky: the moon subtends the same arc in both cases, but is judged to be much larger in the former. That phenomenon may be due to having a referent in the objects on the horizon to which the visible features on the moon can be compared (L. Kaufman & J. Kaufman, 2000).

Many of the classical mechanical invariances—to translation, rotation, and scaling—have been studied in vision, and for the most part objects can be recognized over a range of positions, angles, and sizes. Faces form a particularly well studied case of this, in large part because of the possibility that they may occupy a dedicated chunk of cortical real estate (Freiwald & Tsao, 2010).

Neural Invariance

Given that the brain can develop representations like these, it would be nice to know just how the neural circuitry constructs them. Some of the abovementioned phenomena have been

observed in action at the neural level, either through direct electrophysiological recordings, or by imaging. The computational mechanisms by which they are produced, though, remain mostly undiscovered.

Audition

The most basic neuron-level invariance in auditory processing is center-frequency tuning at the auditory periphery. Inner hair cells remain most sensitive to a particular frequency regardless of sound level, although the bandwidth is level-dependent.

While psychophysical experiments suggest that the brain can process sounds in an intensity-invariant manner, the phenomenon is not well understood, or well documented, at the neural level. To date the only examples of intensity invariance have been observed with broadband sounds, not the simple tones commonly used to probe auditory regions (Barbour, 2011). For instance, while in general the fibers in the auditory increase bandwidth and rate with increasing sound level, many neurons in Marmoset A1 exhibit tuning that is more level independent. In these cells, termed "I" and "O" cells, the bandwidth and temporal activation profiles remain relatively constant over all levels (Sadagopan & X. Wang, 2008). While these cells themselves are not strictly invariant, the authors demonstrate through modeling that a relatively simple computation can construct a level-invariant representation from such outputs.

The only clear neural observation of truly level-invariant neurons comes from recordings of responses to conspecific song in Field L of the Zebra Finch, roughly an analogue to mammalian A1 (Billimoria, Kraus, Narayan, Maddox, & Sen, 2008). As mentioned above, this is a broadband sound, not a narrowband one; further, it is highly complex. This computation seems necessary to handle the sort of content recognition performed by higher level areas that receive information from Field L like HVC, which produces very selective responses to certain sound features (Margoliash & Fortune, 1992).

Pitch invariance is less well defined, though no less studied. In particular one issue is the definition of pitch in the first place: studies looking for pitch have to provide a definition of how they measure it. Pitch can be defined in terms of the first peak in the spectrum; in terms of the separation of two peaks in the spectrum; and in terms of how listeners describe or classify it. The former two are clearly easier to extract from the raw waveform, but there are cases in which they do not accurately describe what listeners perceive. Pitch as defined by human listeners can be extracted from interspike intervals in the auditory nerve of anesthetized cats (Cariani & Delgutte, 1996a), and this encoding accurately predicts how pitch changes due to modifications in complex tones (Cariani & Delgutte, 1996b). Thus neurons in basal auditory areas could, in principle, construct invariant representations of what humans perceive as pitch. At the level of cortex, harmonic pitch is represented by a more traditional rate-coding mechanism (Bendor & X. Wang, 2005), though the more careful representation of how perceived pitch varies with modifications has not been investigated (Bendor & X. Wang, 2006).

Linguistic invariances like speaker-invariant encoding of semantic content and the reverse are difficult to measure at the single-neuron level because we have only simple animal models of langue. Functional imaging studies in humans have demonstrated that some areas in auditory cortex are invariant to transformations of spoken phrases. Certain areas are invariant to time-reversal of speech sounds, which preserves their spectral qualities; others are invariant to spectral reversal, which preserves the temporal qualities (Okada et al., 2010). The linguistic correlates of these transformations is not immediately clear, but this paper represents a step

forward in understanding the function of auditory belt regions and shows how studies of invariance can provide useful insights into auditory processing.

Vision

As suggested by the behavioral evidence, neurons in visual cortex produce representations of objects that are invariant to physical transformations like translation, rotation, scaling, and mirroring. In the most striking case, some neurons in human amygdala appear to discriminate between different individual people and places (Quiroga, Reddy, Kreiman, Koch, & Fried, 2005). This finding provides a bound on where such representation can be constructed. In macaque cortex, clusters of face-selective neurons as measured by fMRI represent faces with invariance to mirroring and rotation. While neurons in the middle lateral (ML) and middle fundus (MF) face patches were sensitive to view angle, those in anterior lateral (AL) were invariant to horizontal mirroring, and those in anterior medial (AM) tended to be insensitive to viewing angle (Freiwald & Tsao, 2010).

At the single-cell level, a portion of neurons in cortical area MT of the macaque respond to the speed of sine gratings across a wide range of spatial frequencies. Other neurons in the area decrease their sensitivity to spatial frequency when multiple frequencies are presented simultaneously, indicating that some level of content-invariant encoding of movement speed is present in this area (Priebe, Cassanello, & Lisberger, 2003). Conversely, neurons in area IT of the macaque show selectivity for particular simple silhouette shapes over a range of positions and sizes, demonstrating a type of scale- and translation-invariance for object recognition (Ito, Tamura, Fujita, & Tanaka, 1995). Unlike face images, rotations of simple objects diminishes response in cortical neurons (Freiwald & Tsao, 2010).

Computations needed for invariance

The preceding evidence suggests that invariant representations can be built up, in a hierarchical process, from low-level features in the brain. If such a process is indeed happening, we would like to know precisely *how* these neurons are going about the process. To examine this, I will now step back and look at how invariance can be achieved from a mathematical point of view.

Intuitively, the primary computation required for invariance is that some function f must remain unchanged over some domain. E.g., if f extracts the pitch of a sound $\mathbf{x}$, the value of f must be constant throughout the set U of all sounds with the same pitch:

$$f(\mathbf{x}) = c : \mathbf{x} \in U \subset D \qquad \textbf{(i.1)}$$

where D is the set of all sounds. Inspection shows that, so long as D has more than one member for each pitch, f must be non-invertible on D. The solutions to this problem, and the problem more generally, are the subject of mathematical group theory. While the non-invertibility of f might seem at first to be troubling, one task of formal invariant theory is to describe how f reparameterizes the space $\mathbf{D}$ by a the value of c and a smaller number of local variables. In the auditory context, we might think of these local variables as loudness and duration plus the timbral parameters.

Groups and Lie algebra

Mathematically, the term *invariance* is associated with group theory and particularly with Lie groups. A *group* comprises the pairing of a set, and an operation that maps between elements in the group. The Lie groups are a special case of this with the additional constraint that both the operation and the inverse of the elements must be analytic (continuously differentiable).

Many formal treatments exist, though one of the more accessible is (Olver, 1999). A full treatment of the subject is beyond the scope of this introduction, buts the notions of groups and their invariants will makes a fair amount of sense given a few examples. A particularly simple group is the set of the real number line $\left\{\mathbf{R}^1\right\}$, and the operation of addition. If we take two elements $h = 1 \in \left\{\mathbf{R}^1\right\}$ and $g = -4.23 \in \left\{\mathbf{R}^1\right\}$, then the combination is also within the set:

$$k = h \cdot g \equiv h + g = -3.23 \in \left\{\mathbf{R}^1\right\}$$

The theory of invariants depends on a subset, that of the *transformation groups*. These are groups that map from one space to another. For instance, the set of linear transformations from $\left\{\mathbf{R}^2\right\} \mapsto \left\{\mathbf{R}^2\right\}$, i.e. the set of 2×2 matrices, forms a Lie group: the inverse is continuously differentiable, as is the operation, in this case, matrix multiplication. The invariants of Lie transformation groups are properties of the transform space that remain unchanged.

A simple example to consider is rotation: distances between points are unchanged, invariant, under rotation of the coordinate system. This corresponds to the set of rotation groups, the set of square, orthogonal matrices with determinant 1:

$$R = \left\{A \middle| A^T A = kI, |A| = 1\right\} : \mathbf{R}^N \mapsto \mathbf{R}^N$$

The point-to-point distance constraint can equivalently be stated that the distance from a point represented by X to the origin remains constant, i.e. that vector magnitude remains unchanged. The set of all points to which a given vector can be rotated traces out the surface of a hypersphere, an N-1 dimensional manifold embedded in N-space, called the *orbit* of this group.

The N-1 dimensional subspace is now parametrized by the invariant (the radius of the hypersphere) and N-1 local coordinates. Defining this group is equivalent to defining a function f that computes the invariant; in the case of the rotations,

$$f(x) = \sqrt{x^T x} : \mathbf{R}^N \mapsto \mathbf{R}^1$$

Applicability

Strict mathematical descriptions may or may not map well to experiments on perception. In certain cases, the mathematical formalisms elegantly describe physical phenomena that the brain must account for: the classical coordinate transformations like rotation and translation preserve the relationships between points in an N-dimensional space. For example, the actual relationship between the vertices of a wooden block will remain unchanged as we move it around a desk, or turn it, or move it closer and further from our eyes, even as the representation at the retina changes greatly (Cutting, 1983). The relationship is straightforward enough, in fact, that the 2D coordinates of vertices in an image (or the area contained between them) can be

transformed directly into relevant coordinates in 3D space in a fashion that holds even if the camera, lens train, or viewing angle changes (Van Gool, Moons, Pauwels, & Oosterlinck, 1995). This experiment does require substantial preprocessing of the images, however, to extract the positions of the vertices.

It should be possible for the brain to learn principles like the connections inherent in a solid object. Recent work on Slow Feature Analysis demonstrates that following a system through time can reveal intrinsic dependencies like point-to-point distances (Wiskott & Sejnowski, 2002). Those relationships, constraints on the distance between sets of points in 3D space, are among the elementary relationships that group theory treats.

The application of these to a perceptually-defined space, though, is not always so straightforward. Visual invariants like the edges of a solid object obey well-defined laws with respect to mechanical and optical transformations. Rigid bodies can translate and rotate, and their representation in the brain is governed first of all by the optical projection onto the 2D manifold of the retina. There is no general law governing all representations, though. Face recognition is a much trickier problem, as demonstrated by the difficulty of performing it with a computer, though the human brain performs the task seemingly as effortlessly. Many invariances in audition would fall into this latter, more complex category: the mechanics that link the identity of the speaker to the perception, through the production of their voice, the propagation through the air, to the mechanical transduction at the cochlea, are more complicated than a few 1×3 (translation), 2×3 (projection), and 3×3 (rotation) matrices.

A strict mathematical analysis of such other invariants quickly becomes more complicated. As an example, consider pitch: humans can recognize, discriminate, and match different sounds along the single dimension of pitch. There is a fair amount of debate, though, about what physical events, i.e. what features of the sound-pressure waveform, create this percept, and about how the neurons of the auditory system represent it (Cariani & Delgutte, 1996a). Just the proliferation of terms for pitch listed in the aforementioned study gives evidence that this case is somewhat less straightforward than Cutting's wooden block. Pitch is thus very much defined by the percept, rather than by some clear physical law: it has more in common with US Supreme Court Justice Potter Stewart's maxim of "I know it when I see it" than with Noether's theorem.

The origin of vocal features

Perceptual invariants should, somewhere in the brain, have a neural correlate. That fact, though, is not of immediate use when attempting to understand how networks of neurons construct such a representation out of sensory inputs from the periphery. One possible way forward, then, is to try a representation that looks more like Cutting's wooden block: precise, and measurable. In this section, I will describe how the modulation power domain may provide a way forward.

If the computation of invariants works best with explicit representations of physical space, then auditory invariants would need some sort of features to represent. One option for extracting coordinates like Van Gool's from vocal sounds is to compute an explicit representation of physical production variables. For instance, information about tongue position can be inferred from clean recordings of stop consonants (Blumstein & Stevens, 1979). In practice, though, it may not be necessary to compute such explicit representations if the underlying information about the vocal tract is present in a clear enough fashion. In this section, I

will demonstrate how one representation, the Modulation Power Spectrum or MPS, is determined directly by the vocal production apparatus and contains accessible representations thereof.

The Modulation Power Spectrum

Spectrograms show characteristic spectral and temporal features that can be visualized by taking a 2D Fourier transform. The MPS for human speech and zebra finch song are similar and are distinctive because they are strongly nonseparable. Power for both classes of sound is distributed primarily along either the temporal modulation axis or the spectral modulation axis, and there is relatively little power for cross spectrotemporal modulations. This shape is not determined by any laws of physics or waveforms; in fact, the allowed region of the modulation power space includes these features, and they are present in white noise and many synthetic sounds (Singh & Theunissen, 2003).

Rather, this property is particular to animal vocalizations, and the different portions of the space represent different articulations of the vocal apparatus. The limited area of the MPS along the two axes contains virtually all of the information required for speech processing. In fact, other areas of the MPS can be removed with little effect on the intelligibility of speech (Elliott & Theunissen, 2009).

The cepstrum and the source filter model

These properties of the MPS have their roots in the physical origin of animal vocalizations including human speech. This system is commonly modeled as a two-part system: vocalizations begin with a source, which is then filtered with a linear-time-invariant (LTI) filter to create the final sound. In the case of voiced sounds the source is a harmonic signal generated by the larynx or syrinx; in whisper speech, the source is broadband noise with little spectral character. In all cases the filter is the upper respiratory tract (Taylor & Reby, 2010). Although early models

For linear filters, the source and filter components can be separated by deconvolution, which can be performed elegantly using the cepstrum (Gold & Morgan, 2000). The underlying principle behind the deconvolution is simple: linear filters are convolutional in the time domain, and convolution in the time domain is multiplication in the frequency domain. Following eq. 20.1 and 20.2 from Gold & Morgan,

$$x(t) = \phi(\tau) * s(t) \qquad \textbf{(i.2)}$$

$$\left|X(f)\right| = \left|\Phi(f)\right| \cdot \left|S(f)\right| \qquad \textbf{(i.3)}$$

$$\log\left(\left|X(f)\right|\right) = \log\left(\left|\Phi(f)\right|\right) + \log\left(\left|S(f)\right|\right) \qquad \textbf{(i.4)}$$

where $x(t)$ is the emitted sound-pressure waveform at time t, $\phi(\tau)$ the linear vocal tract filter at delay τ, and $s(t)$ is the source at time t. $X(f)$, $\Phi(f)$, and $S(f)$ are the Fourier transforms of these signals, respectively, parameterized by the frequency f.

The final key to the utility of the cepstrum is that the vocal tract filter and the laryngeal or noisy source vary in amplitude along the frequency axis with vastly different periods. The vocal tract filter tends to have slow features, meaning that Φ varies slowly with f. Φ typically has

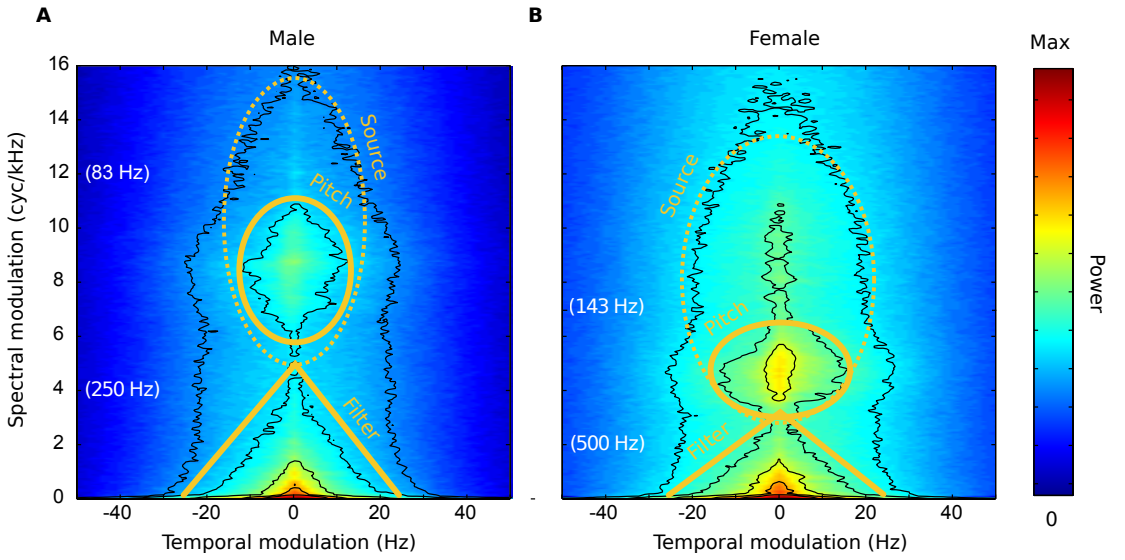


Fig. i.1: Modulation Power Spectrum for Native English Speech. Both panels contain the MPS of "emotionless" English sentences spoken by American native talkers taken from the Iowa Corpus, and several characteristic features are labeled in both plots. The triangular region labeled *filter* contains most of the energy for the formants, and corresponds to the "core" modulations found by Elliott (Elliott & Theunissen, 2009). The region of energy from the harmonic pitch is marked *pitch*, and corresponds to a sinusoid in log-power frequency space. A larger region, marked *source*, contains the pitch as well as other energy that represents, among other things, the shape of the spectral lines making up the harmonic, i.e., the goodness of pitch. **A: MPS for male speakers.** The energy from the fudamental pitch falls entirely between 4 and 12 cycles/kHz, corresponding to f_0 between 250 and 83Hz. **B: MPS for female speakers.** The fundamental pitch energy is between 2 and 7 cycles/kHz, corresponding to f_0 between 500 and 143Hz. This difference is lower than for males, though there is significant overlap. The filter/formant region covers a smaller region for female talkers, likely due to the shorter overall length of the upper respiratory tract.

only a few peaks, which correspond to the formants. Harmonic sounds S from the larynx or syrinx vary quickly in amplitude along the spectral axis f, with peaks spaced by the fundamental f_0. Noisy sources, for example from whisper speech, behave differently and cause S to vary very little, having roughly equal power at all f. Distally-produced sounds such as unvoiced consonants constitute a special case, where S and Φ are both similarly flat.

We can exploit these different scales by taking a second Fourier transform along the spectral axis, yielding the cepstrum, $c(\omega_f)$:

$$c(\omega_f) = \mathrm{FT}\left(\log\left(\left|X(f)\right|\right)\right)$$

$$c(\omega_f) = \mathrm{FT}\left(\log\left(\left|\Phi(f)\right|\right)\right) + \mathrm{FT}\left(\log\left(\left|S(f)\right|\right)\right) \qquad \textbf{(i.5)}$$

So long as the assumption that the source and filter have different scales holds, their contributions will now be completely separate as illustrated by figure i.1. For many animal vocalizations, including songbirds and humans, this is a reasonably good approximation. For this reason the cepstrum is well known in the computational speech processing literature, and in fact forms the basis of most modern Automatic Speech Recognition (ASR) and many telecommunications data compression systems (Gold & Morgan, 2000).

The Modulation Power Spectrum and the cepstrum

In fact, all of these analyses are the same. The MPS contains the information as the cepstrum, because up to this point the calculation is the same. To obtain the MPS, we simply take a series of spectral samples $x(t')$ over time, parameterized by absolute time t'. The linear spectrogram is then $X(f,t')$, and we can compute a "cepstrogram" $c(\omega_f,t') = \mathrm{FT}_t\left(\log\left(\left|X(f,t')\right|\right)\right)$. The MPS is then

$$M(\omega_f,\omega_t) = \mathrm{FT}_t\left(c(\omega_f,t')\right) \qquad \textbf{(i.6)}$$

This simple derivation means that the MPS performs the same task as the cepstrogram: separating the source and filter based on their characteristic spectral modulations.

Nonseparability of the MPS

The MPS, though, does something important that the cepstrum does not. For many vocalizations—certainly at least for Zebra Finch song and for English spoken by native speakers—structure of the MPS is nonseparable. This lab has previously demonstrated that this is not the result of any physical restriction on possible waveforms (Woolley, Fremouw, Hsu, & Theunissen, 2005), but rather, a characteristic either of these particular vocal repertoires or of the physical and neural vocal mechanisms. It is possible that this is the result of constraints imposed by the physical vocal apparatus. Limits on the speed of muscle movement in the upper respiratory tract or the diaphragm could in fact restrict what modulations can be produced, though this remains to be proven.

Nonseparability means simply that the whole modulation power spectrum $M(\omega_f,\omega_t)$ cannot be decomposed as the product of the spectral and temporal marginals:

$$M(\omega_f,\omega_t) \neq F(\omega_f)\cdot G(\omega_t) \qquad \textbf{(i.7)}$$

$$F(\omega_f) = \tfrac{1}{n_f}\sum_{i=1}^{n_f} M(\omega_f,\omega_t)$$

$$G(\omega_t) = \tfrac{1}{n_t}\sum_{i=1}^{n_t} M(\omega_f,\omega_t)$$

This nonseparability is the result of the fact that certain cepstral bands have strong temporal correlations while others do not. This in turn means that the characteristic timescales of the various parts of speech have drastically different timescales. Besides the vowel-formant separation achieved by the cepstrum, the MPS also separates formants from formant transitions and plosives. That separation is in fact quite challenging for many speech-processing systems (Hu & D. Wang, 2004): using the cepstrum or the spectrogram as a representation tends to make an implicit assumption that all features are modulated at the same timescale. In fact, the empirical shape of the MPS of speech shows that this is a very poor approximation (Woolley et al., 2005).

Sensitivity to these particular modulations suggests that animal brains are exploiting these dependencies to identify and process vocalizations. The results that I will present in the rest

of this thesis further suggest a specific role for this sort of processing, both in brains and in audio processing.

Chapter 1

Noise-invariant Neurons in the Avian Auditory Cortex: Hearing the Song in Noise.

Abstract

Robust neural representations of sounds are essential for generating invariant percepts of behaviorally relevant acoustical signals distorted by propagation or noise. I found that a subset of neurons in a secondary area of the avian auditory cortex exhibit noise-invariant responses to bird song in the sense that they responded with similar spike patterns to song stimuli presented in silence and over a background of masking noise. By characterizing the neurons' tuning in terms of their responses to modulations in the temporal and spectral envelope of the sound, I also show that noise-invariance is partly achieved by selectively responding to slow sound features with high spectral structure.

Introduction

Invariant neural representations of behaviorally relevant objects are a hallmark of high-level sensory regions and are interpreted as the outcome of a series of computations that would allow us to recognize and categorize objects in real life situations. For example, view-invariant face neurons have been found in the inferior temporal cortex (Freiwald & Tsao, 2010) and are thought to reflect our abilities to recognize the same face from different orientations and scales. The representation of auditory objects by the auditory system is less well understood although neurons in high-level auditory areas can be very selective for complex sounds and, in particular, communication signals (Mooney, 2001; Rauschecker, Tian, & Hauser, 1995). It has also been shown that auditory neurons can be sound level invariant (Billimoria, Kraus, Narayan, Maddox, & Sen, 2008; Sadagopan & Wang, 2008) or pitch sensitive (Bendor & Wang, 2005). As is the case for all neurons labeled as invariant, pitch sensitive neurons respond similarly to many different stimuli as long as these sounds yield the same pitch percept. In particular, they respond equally well to harmonic stacks with or without power in the fundamental frequency. Both sound level invariant and pitch sensitive neurons could therefore be building blocks in the computations required to produce invariant responses to particular auditory signals subject to distortions due to propagations or corruption by other auditory signals. However, the existence of "recorder" invariant (in analogy with "view" invariant in vision) auditory neurons remains unknown. Similarly, the neuronal computations required to recognized communication signals embedded in noise are not well understood although it is known that humans (Bronkhorst, 2000) and other animals (Bee & Micheyl, 2008) excel at this task.

Birds, like humans, communicate vocally. In general, however, these sounds do not occur against a silent background. Environmental noise may impinge, but in more gregarious species other simultaneous voices can pose an even larger problem. How, then, does one pick out a single speaker against a background of very similar sounds? This is precisely the cocktail party problem describedy by Cherry (Cherry, 1953).

In this study, I examined how neurons in the secondary avian auditory cortical area NCM (*CaudoMedial Nidopalium*) responded to song signals embedded in background noise to test whether this region presents noise-invariant characteristics that could be involved in robust song recognition. I chose the avian model system because birds excel at recognizing individuals based on their communication calls (Stevenson, R. E. Hutchison, J. B. Hutchison, Bertram, & Thorpe, 1970; Vignal, Mathevon, & Mottin, 2004b), often in very difficult situations (Aubin & Jouventin, 2002). Moreover, the avian auditory system is relatively well characterized and it is known that neurons in higher-level auditory regions can respond selectively to particular conspecific songs (Knudsen & Gentner, 2010). I focused my study on NCM because a series of neurophysiological (Phan, 2006; Stripling, Volman, & Clayton, 1997) and immediate early gene studies (Bolhuis, Zijlstra, Boer-Visser, & Van der Zee, 2000; Mello et al., 1995; Mello, Nottebohm, & Clayton, 1995) have implicated this secondary auditory area in the recognition of familiar songs. In addition, although neuronal responses in the primary auditory cortex regions are systematically degraded by noise (Narayan et al., 2007), Zenk expression suggested that NCM neural activity in response to conspecific song was relatively constant for a range of behaviorally relevant noise levels (Vignal, Attia, Mathevon, & Beauchaud, 2004a).

Several computational modeling projects have examined the cocktail party problem from an explicitly neural perspective. Von der Malsburg and Schneider demonstrated that they could extract information from two streams from the output of a model of ensemble spiking neurons (Malsburg & Schneider, 1986). Following the literature on sparse independent neural representation (Olshausen, 2002; Olshausen & Field, 1996; Smith & Lewicki, 2006), Asari et. al. showed that decomposition of the spectrogram along those lines can help with sound source separation if the positions of the sources are known (Asari, Pearlmutter, & Zador, 2006). Projecting the spectrogram into a neural-ensemble-like space using a basis inspired by STRFs can similarly help with stream segregation (Elhilali & Shamma, 2008). A smaller number of experiments have found direct evidence that neurons in primary auditory cortex are involved in this sort of stream segregation (Narayan et al., 2007).

Methods

I recorded neural responses from single neurons in the NCM of anesthetized adult male Zebra Finches. By carefully orienting the electrode angle, I was able to sample NCM along its entire dorsal to ventral extent (Fig. 1K). I obtained responses to 40 different unfamiliar conspecific songs played back at 70 dB SPL and to the same songs embedded in a synthetic naturalistic noise with a signal to noise ratio of 3 dB. The naturalistic noise was obtained by low-pass filtering white noise is the space of temporal and spectral modulations to obtain modulation-limited noise (ml-noise). I quantified the noise invariance of each neuron calculating a de-biased correlation coefficient between the post-stimulus time histograms (PSTHs) obtained for the song alone and song + ml-noise stimuli. I called this correlation coefficient the *noise invariance.*

Materials

Animals

I recorded from four urethane-anesthetized Zebra Finches. Animal rearing, surgical preparation, anaesthesia, euthanasia, and histological reconstruction were all performed as reported in my lab's earlier work (Woolley, Gill, Fremouw, & Theunissen, 2009), as approved by the UC Berkeley Animal Care and Use Committee.

Sound Stimuli

My stimuli were divided into three classes: zebra-finch songs, roughly 1.6-2.6 seconds in length; synthetic modulation-limited noise (ml-noise) sounds, each exactly four seconds long; and combinations of the two. I played four trials at each recording location, each consisting of a randomized sequence of 40 songs, 40 masking noise stimuli, and 40 combined stimuli. Stimuli were separated by a period of silence with a length uniformly and randomly distributed between five and seven seconds.

Each of the combined stimuli consisted of one ml-noise sound, randomly paired with one of the songs. The noise stimulus began the standard five to seven seconds after the previous stimulus, and the song began after a random delay of 0.5 to 1.5 seconds. In these combined presentations, I attenuated the noise stimuli 3dB below their normal level.

I created the ml-noise stimuli using Matlab by low-pass filtering Gaussian white noise in the modulation domain using the modulation filtering procedure described in Chapter 2. This modulation low-pass filter had cutoff frequencies of $\omega f = 1.0$ cycles/kHz and $\omega t = 50$ Hz and gain roll off of 10dB/(cycle/kHz) and 10dB/10Hz. The cutoff modulation frequencies were chosen in order to generate noisy sounds with similar range of modulation frequencies found in environmental noise. In addition, most of the modulations found in zebra finch song are well masked by this synthetic noise although it should be noted that song also includes sounds features with high spectral modulation frequencies (above 2 cycles/kHz) and high temporal modulation frequencies (above 60Hz).

The frequency spectrum of the ml-noise was flat from 250 Hz to 8 kHz completely overlapping the entire range of the band-passed filtered songs I used in the experiments. Thus, although different results could be found with noise stimuli with different statistics, I carefully designed my masking noise stimulus to both capture the modulation found in natural environmental noise while at the same time completely overlapping the frequency spectrum of my signal.

All stimuli were processed to equalize loudness using custom code in Matlab and presented using software and electronics (Tucker-Davis Technologies, Alachua, FL, www.tdt.com). Stimuli were stored and presented using two RP2 processors, amplified with an SA1 amplifier, and played over a speaker (Blaupunkt) at 72dB C-weighted average sound pressure level.

Electrophysiology

Because I was looking for cells in NCM, I used more medial coordinates than in my lab's previous experiments. With the bird's beak fixed at a 55° angle to the vertical, electrodes were inserted roughly 1.2mm rostral and 0.5mm lateral to the Y-sinus.

After preparatory stereotactic surgery, I positioned the bird 20cm in front of the loudspeaker inside a double-walled anechoic chamber (Acoustic Systems, Inc., now part of ETS

Lindgren, Cedar Park, TX, www.acousticsystems.com). All electronics were either grounded (microdrives) or shielded by a grounded enclosure (speaker).

I made extracellular recordings from tungsten-parylene electrodes having impedance between 1 and 3 mega Ohms (A-M Systems, Sequim, WA www.a-msystems.com). Electrodes were advanced in 0.5µm steps with a microdrive (Newport, Irvine, CA, www.newport.com), and extracellular voltages were recorded with a system from Tucker-Davis Technologies. Signals were amplified with an RA4 headstage, digitized with an RA4PA Medusa four-channel preamp, and collected with an RA16BA Medusa Base Station. All data were saved asynchronously to a Dell computer running Windows XP (Microsoft) using OpenEx software (TDT).

In all cases, the extracellular voltages were thresholded to collect candidate spikes. Each time the voltage crossed the threshold, the timestamp was saved along with a high-resolution waveform of the voltage around that time (0.29ms before and 0.86ms after for a total of 1.15ms). After the experiment, these waveforms were sorted using SpikePak (TDT) to assess unit quality.

In each bird, I advanced the electrode until I found auditory responses, then recorded a full protocol as described in the previous section. When I no longer found auditory responses, I moved the electrode 300µm microns further, made an electrolytic lesion (2uA x 10s), advanced another 300µm, and made a second identical lesion.

Data Analysis

I used custom code written in MATLAB and Python for all of these analyses, including TDT's OpenDev suite with Matlab to export data for processing; code in Matlab for most of the numerical analysis; and a MySQL database and code in Python using SQLAlchemy, NumPy, and SciPy for meta-analyses.

Responsiveness

I assessed responsiveness using an average z-score metric for each stimulus class (Amin, Grace, & Theunissen, 2004). The z-score for the *i*th stimulus is

$$z_i = \frac{\tilde{r}_i}{\sqrt{\frac{1}{n_{trials}-1}\sum_{j=1}^{n_j}\left(\tilde{r}_{ij}-\tilde{r}_i\right)}}$$

$$\tilde{r}_i \equiv \sum_{j=1}^{n_{trials}}\left(\tilde{r}_{ij}-\tilde{r}_{ij0}\right)$$

$$\tilde{r}_{ij} \equiv \frac{1}{n_{samples}\,t_{sample}}\sum_{k=1}^{n_{samples}}\left(r_{ijk}\right) \qquad \textbf{(1.1)}$$

where $\tilde{r}_i$ is the background-subtracted mean rate for the *i*th stimulus; n_j is the number of trials for the *i*th stimulus; $\tilde{r}_{ij}$ is the raw mean rate and $\tilde{r}_{ij0}$ the background rate for the *j*th trial of the *i*th stimulus; r_{ijk} the number of spikes in time bin *k*; n_k the number of time bins; and t_k the width of the bin.

The average z-score for that class of stimuli is then

$$\bar{z} = \frac{1}{n_{stims}} \sum_{i=1}^{n} z_i \qquad \textbf{(1.2)}$$

Using a cutoff of $\bar{z} \geq 1.5$ for either ml-noise or song stimuli, 32 of the 50 single units were responsive.

Invariance

To measure invariance, I evaluated the similarity between the responses to masked and unmasked song by computing the correlation coefficient between the PSTH (Peri-Stimulus Time Histogram) for each response. Each PSTH was first smoothed using a 31 millisecond Hanning window. This method in many ways resembles the Rcorr method of Schreiber (Schreiber, Fellous, Whitmer, Tiesinga, & Sejnowski, 2003), except the pairing is done between two different sets of spike trains.

Mean PSTH

I started with the average response from all trials for each stimulus, averaged across trials to produce a PSTH and smoothed:

$$\bar{r}_i \equiv \frac{1}{n_{trials}} w * \sum_j r_{ij} \qquad \textbf{(1.3)}$$

where $\bar{r}_i$ is the PSTH for the *i*th song stimulus, n is the number of trials, w is the window, and r_{ij} is again the response to the *j*th trial of the *i*th song stimulus.

Responses to masked stimuli

We can construct a similar PSTH $\bar{r}_i^m$ for the responses to the masked stimuli by substituting the single-trial response to the *j*th masked presentation of the *i*th stimulus, r_{ij}^m, for the unmasked single-trial responses r_{ij} in equation 1.3. I will extend the superscript 'm' notation to refer to variables collected from the masked trials.

Jackknife bias-correction of invariance

To correct for bias introduced by the small number of trials used to compute each PSTH (four), I used leave-one jackknifing (Efron & Tibshirani, 1994; Quenouille, 1956; Tukey, 1958)The single-stimulus results indicate a small but consistent negative bias in the four-trial estimates. I then computed the invariance as the mean of the individual bias-corrected correlations for each stimulus. This again diverges from Schreiber's Rcorr in that I report a bias-corrected estimate of the mean rather than the raw value.

Jackknife holdout sets

Following the notation of Efron, we can define leave-one a jackknife estimate of the PSTH, $\bar{r}_{i(p)}$:

$$\bar{r}_{i(p)} = \frac{1}{n_{trials}-1} w * \sum_{j \neq p} r_{ij} \quad \textbf{(1.4)}$$

As an extension, we can construct a delete-d jackknife PSTH $\bar{r}_{id}$:

$$\bar{r}_{idk} = \frac{1}{n_{trials}-d} w * \sum_{j \in q_k} r_{ij} \quad \textbf{(1.5)}$$

$$q_k \subset \{1,2,\ldots,n_{trials}\}, |q_k| = n_{trials} - d \quad \textbf{(1.6)}$$

q_k is defined as the *k*th subset of the trial indices with *d* items removed. I will use Effron's "dot" notation to mark the null holdout with no trials omitted, i.e.,

$$\bar{r}_{i0k} \equiv \bar{r}_{i(\cdot)k} \equiv \bar{r}_{i(\cdot)} \equiv \bar{r}_i$$

Correlation coefficients

We can now compute correlation coefficients between these sets:

$$c_{idk} = \mathrm{corr}\left(r_{idk}, r^m_{idk}\right) \quad \textbf{(1.7)}$$

where correlation is defined conventionally as

$$\mathrm{corr}(a,b) = \frac{a \cdot b}{\sqrt{(a \cdot a)(b \cdot b)}} \quad \textbf{(1.8)}$$

From this set of correlation estimates, we can use the jackknife to bias-correct the correlation coefficient, effectively estimating the infinite-trial correlation. This is done by regressing the estimates $\left\{c_{ipk}\right\}_{p=0,1,\ldots d}$ against $\frac{1}{n_{trials}-p}$:

$$\hat{\alpha}_d,\hat{\beta}_d = \underset{\alpha,\beta}{\arg\min}\left(\sum_{p=0}^{d}\sum_{k=1}^{n_k(p)}\varepsilon_{ipk}^2\right) \quad \textbf{(1.9)}$$

$$\hat{c}_{id} = \hat{\alpha}_{id} \quad \textbf{(1.10)}$$

$$\varepsilon_{ipk} = \left(c_{ipk} - \beta x_p - \alpha\right)$$

$$x_p = \frac{1}{n_{trials}-p}$$

$$n_k(p) = \binom{d}{n_{trials}-p}$$

The resulting y-axis intercept, $\hat{\alpha}_{id}$ corresponding to $n=\infty$, gives a bias-corrected estimate of $c_{\infty i}$, the expected correlation for an infinite number of trials. Here ε_{ipk} is the regression error for the jackknifed sample indexed by *i*, *p*, and *k*.

This method also gives an error bar, in the form of the standard error of the y-intercept, computed as

$$\hat{\sigma}_{id} = \sigma_{\hat{\alpha}_{id}} = \sigma_{\hat{\beta}_{id}} \sqrt{\tfrac{1}{m}\sum_{p=0}^{d} n_k(p) x_p^2} \qquad \textbf{(1.11)}$$

$$\sigma_{\hat{\beta}_{id}} = \sqrt{\frac{\frac{1}{m-2}\sum_{p=0}^{d}\sum_{k=1}^{n_k(p)} \hat{\varepsilon}_{ipk}^2}{\sum_{p=0}^{d} n_k(p)\left(x_p - \bar{x}\right)^2}}$$

$$m = \sum_{p=0}^{d} n_k(p)$$

$$\bar{x} = \tfrac{1}{m}\sum_{p=0}^{d} n_k(p) x_p$$

where $\hat{\sigma}_{id}$, the standard error of our estimate, is the standard error of the intercept $\sigma_{\hat{\alpha}_{id}}$. $\sigma_{\hat{\beta}\,id}$ is the standard error of the slope, and m is the total number of samples. This defines a $(1-\alpha)\%$ confidence interval δ_α

$$\delta_\alpha = \frac{t_\alpha \hat{\sigma}_{id}}{\sqrt{m}} \qquad \textbf{(1.12)}$$

$$p\left(\hat{c}_{id} - \delta_\alpha \le c_{i\infty} \le \hat{c}_{id} + \delta_\alpha\right) = 1 - \alpha$$

For this experiment, I have assumed that the error is normal and computed t_α from the Student's t distribution with $m-1$ degrees of freedom.

We can compute a final estimate of the cell's invariance $\hat{c}_d$ as the mean of the distribution of the single-stimulus correlation estimates $\{\hat{c}_{id}\}$, and similarly a variance for this population $\hat{\sigma}_d^2$:

$$\hat{c}_d = \frac{1}{n_{stims}} \sum_{i=1}^{n_{stims}} \hat{c}_{id} \quad \textbf{(1.13)}$$

$$\hat{\sigma}_d^2 = \frac{1}{n_{stims}-1} \sum_{i=1}^{n_{stims}} \left(\hat{c}_{id} - \hat{c}_d\right)^2 \quad \textbf{(1.14)}$$

Spectrogram computation

All stimuli, denoted in the time domain by $s(t)$, were preprocessed using a short-time Fourier transform and then computing the log power. Before taking the Fourier transform, I applied a Gaussian window $w(\tau)$ to the samples, with the length of the window chosen to give a 125Hz bandwidth in the final spectrogram. For a set of stimuli, I computed linear power spectrogram as the squared amplitude of each Fourier coefficient:

$$a(f,t) = \mathrm{STFT}\left(s(t), w(\tau)\right)$$

$$P(f,t) = a^2(f,t) \quad \textbf{(1.15)}$$

Because I will ultimately be taking the log, it is important to avoid zero values in the linear spectrogram. The traditional engineer's approach is simply to add a negligible quantity ε to each value before taking the log. Common values for the fudge factor ε include 10^{-6}, 10^{-9}, etc., and machine epsilon. The modified power P' is then:

$$P'(f,t) = P(f,t) + \varepsilon$$

$$\varepsilon << \min_{f,t}\left(P(f,t)\right)$$

A more principled approach, though, is to set an explicit noise floor. To do this, one first normalizes the spectrograms by dividing by a constant P_0. In this case I used the maximum power in any band across the entire set of stimuli being used for the regression. One can in principle use a different constant for each stimulus, although this eliminates any information about relative loudness between sounds. The normalized power is then $P'(f,t) \in (0,1]$:

$$P'(f,t) = \frac{P(f,t)}{P_0} \quad \textbf{(1.16)}$$

$$P_0 = \max_{f,t}\left(P(f,t)\right)$$

One can now take the log of this power without fear of zero values. Applying the noise floor requires choosing a value Φ, e.g. -80dB, and truncating P' below that value:

$$S(f,t) = \max\left(20\log_{10}\left(P'(f,t)\right),\Phi\right) \quad \textbf{(1.17)}$$

Equivalently, one can add Φ and then truncate at zero:

$$S(f,t) = \max\left(20\log_{10}\left(P'(f,t)\right) + \Phi, 0\right)$$

Spectro-Temporal Receptive Fields (STRFs)

For each responsive single unit, I computed a linear STRF using STRFLAB (strflab.berkeley.edu) using ridge regression, a method very similar to the normalized reverse correlation method described by our lab previously (Theunissen et al., 2001). This method, available as the "directfit" method in STRFLAB, provides a least-squares solution to the linear regression between the stimulus represented as a log-power spectrogram and the response represented as a smoothed PSTH:

$$\hat{h} = \mathbf{C}_{SS}^{-1}\mathbf{C}_{SR} \quad \textbf{(1.18)}$$

where $\hat{h}$ is the STRF estimate, $\mathbf{C}_{SS}$ is the stimulus cross-correlation, and $\mathbf{C}_{SR}$ is the stimulus-response cross correlation.

This fit presumes zero-mean stimulus and response, requiring that we subtract (and save) the mean stimulus and mean response. In this case, I compute the mean stimulus per band. Though not strictly necessary, I also normalized the stimuli so that each band had not only zero mean but unit variance:

$$S'(f,t) = \frac{S(f,t) - S_0(f)}{\sigma_S(f)} \quad \textbf{(1.19)}$$

$$S_0(f) = \underset{t}{\text{mean}}\left(S(f,t)\right)$$

$$\sigma_S(f) = \underset{t}{\text{std}}\left(S(f,t)\right)$$

The autocorrelation $\mathbf{C}_{SS}(f,\omega_t)$ is then computed from the Fourier transform of $S'(f,t)$ as described in (Theunissen et al., 2001).

Similarly, the mean-subtracted response is

$$r'(t) = r(t) - r_0 \quad \textbf{(1.20)}$$

$$r_0 = \underset{t}{\text{mean}}\left(r(t)\right)$$

and the crosscorrelation $\mathbf{C}_{SR}(f,\omega_t)$ is computed from $R(\omega_t)$, the Fourier transform of $r'(t)$, again as described in (Theunissen et al., 2001).

In all of my calculations, I considered the mean values $S_0(f)$ and r_0 and variances $\sigma_S(f)$ to be static parameters (as opposed to ones to be optimized) and stored them as such.

STRF performance

I assessed the performance of each STRF using coherence and the normal mutual information. First I compute the expected coherence between two single response trials; I then computed the coherence between the STRF prediction and the average response. I then compute the normal

mutual information for each (Hsu, Borst, & Theunissen, 2004a), calling the former the *response information* and the latter the *predicted information*. I call the ratio of the predicted information to the response information the *performance ratio*, and provides a measure of model performance that is independent of the variability of the neuron. In all of my receptive field analyses, I used only STRFs that predict sufficiently well, defined here as having predicted information of at least 1.2 bits/second and a performance ratio of at least 20%.

STRF cross prediction

In order to assess how the linear model (i.e. the STRF) accounts for the observed invariance, I used the STRF to predict the response to noise-corrupted stimuli. Because the noise stimuli were prerecorded, I was able to compute the exact stimulus presented during the masker trials. Because the masker-stimulus combination varied by trial, these stimuli were trial-specific.
I preprocessed the masked stimuli using the parameters from the original optimization. That is, I computed $S^{m'}$ and $r^{m'}$ using S_0, r_0, and σ_S rather than computing new means S_0^m and r_0^m and variances σ_0^m from the masked stimuli S^m and corresponding responses r^m:

$$S^{m'}(f,t) = \frac{S^m(f,t) - S_0(f)}{\sigma_0(f)} \quad \textbf{(1.21)}$$

$$r^{m'}(t) = r^m(t) - r_0 \quad \textbf{(1.22)}$$

I then computed the prediction for each trial based on this new stimulus:

$$\hat{r}^m(t) = \hat{h}(f,\tau) * S^{m'}(f,t) + r_0 \quad \textbf{(1.23)}$$

Prediction power can be assessed again using the normal mutual information as described in the previous section.

Linear model (STRF) invariance

I wanted to know how well the linear model captured the invariance of the cell. To that end, I computed a the cross-predicted response to song + noise for each trial using equation 1.23. To be consistent with the bias correction procedure used for the invariances, I computed jackknife samples for the predictions as well:

$$\hat{r}_{idj}^m = \frac{1}{n_{trials}-d} w * \sum_{j \in q_k} \hat{r}_{ij}^m$$

I then computed the jackknifed correlation estimates between the $\hat{r}^{m}_{idj}$ and $\bar{r}_{idj}$ using equations 1.7-1.14 to compute a bias-corrected correlation estimate $\hat{c}'_d$ with its standard deviation $\hat{\sigma}'_d$. To determine whether the model invariance differed from the measured invariance, I compared that distribution to the single-stimulus correlations $\hat{c}_{id}$ using Welch's approximate two-sample, two-tailed t-test for distributions with different variances:

$$t'_s = \frac{\hat{c}'_d - \hat{c}_d}{\sqrt{\frac{\sigma'_d + \sigma_d}{n_{stims}}}} \qquad \textbf{(1.24)}$$

Modulation Power Transfer Functions (MPTFs)

To interpret the STRFs, I computed the Modulation Power Transfer Function (MPTF) from each STRF by taking the 2D Fourier Transform, squaring to obtain the power, and truncating to the upper two quadrants (Hsu, Woolley, Fremouw, & Theunissen, 2004b). The MPTF expresses the power transmitted in terms of spectral and temporal modulations. I then computed the center of mass of each transfer function, yielding coordinates in spectro-temporal modulation space. I interpret these coordinates as the best spectral modulation frequency and best temporal modulation frequency for the cell.

Spike waveforms

One crucial question that arose is whether invariance is a property of particular types of cells. In some cases, different types of neurons have different extracellular spike waveform shapes, so I used this as a proxy for type of cell.

For each responsive, predictive single unit in the sample, I computed a representative waveform for that by averaging all of the spikes assigned to the unit. I then normalized that mean waveform for each unit by the maximum of its absolute value.

Results

I find that three things contribute substantially to invariance. First, cells in more dorsal areas, probably secondary auditory regions, are more invariant. Second, cells with certain receptive field properties are more invariant. Third, processing nonlinearities only contribute to invariance in a few cells; for the bulk, nonlinearity makes the cell less invariant.

Invariance and anatomy

As illustrated on the left panels in ure 1 (A-E), responses of some neurons to song signal were almost completely masked by the addition of noise. In these situations, the PSTH obtained for song only (panel C) is very different than the one obtained for song + noise (panel E) yielding small values of noise invariance (close to 0). However, some neurons also showed a striking robustness to noise degradation as illustrated in panels F-J. Those neurons had similar PSTHs for both conditions and high values of noise invariance (close to the maximum value of 1). Neurons with different degrees of invariance were found throughout NCM but the neurons in the ventral region tended to have highest degrees of invariance (panel L). NCM also exhibits some degree of frequency tonotopy along this dimension with higher frequency tuning found in more ventral regions (Ribeiro, Cecchi, Magnasco, & Mello, 1998; Terleph, Mello, & Vicario, 2006) but this cannot explain the organization for noise invariance since the song and ml-noise stimuli had

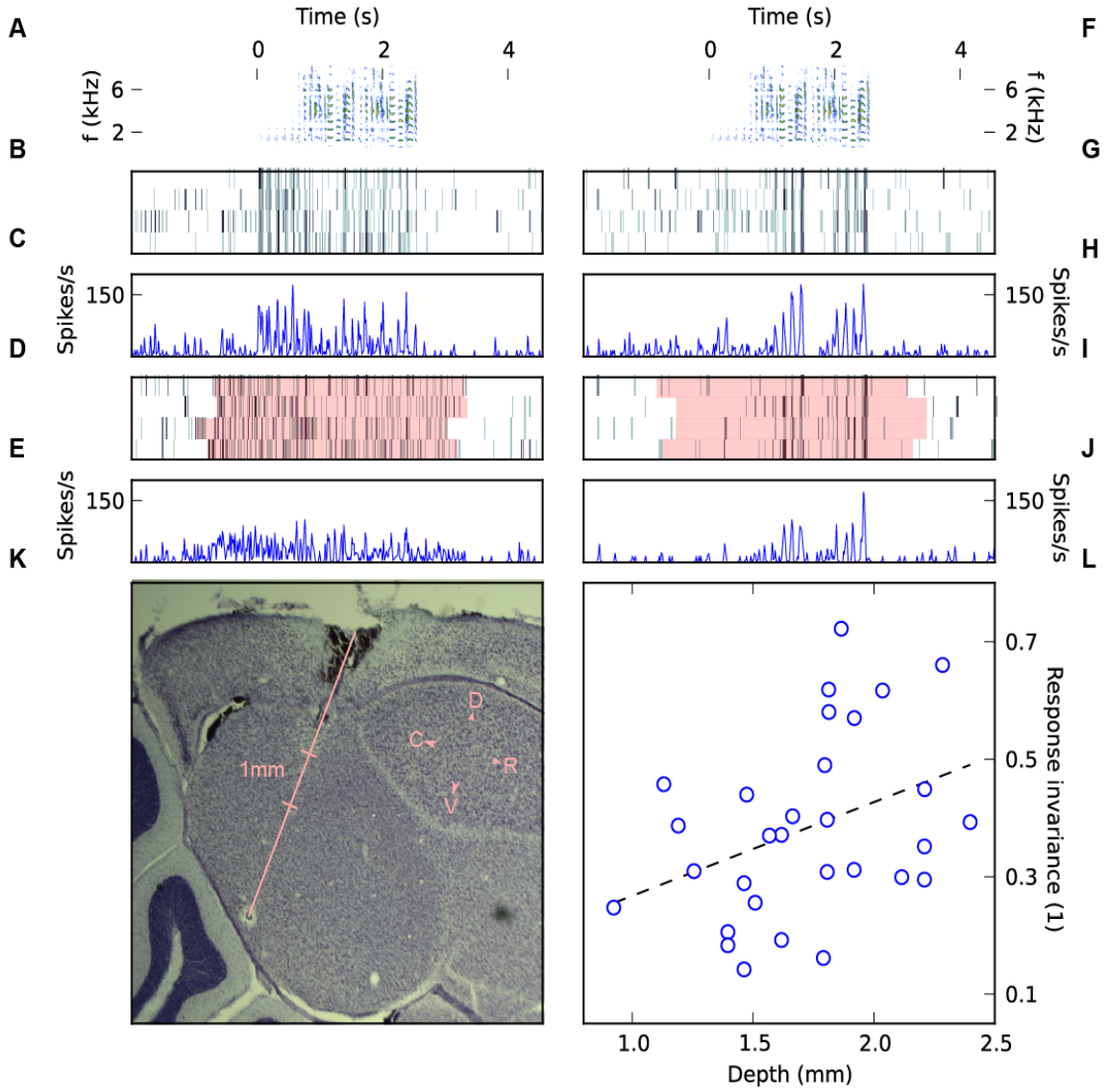


Figure 1.1: Noise-invariant Responses in the Avian NCM. **A, F. Spectrograms showing the same zebra finch song stimulus used in two separate recordings**. Song starts at 0s. The spectrogram of the song+ml-noise stimuli is not shown. **B-C, G-H. Raster plots (B, G) and corresponding smoothed PSTH (C, H) showing the response of each neuron to the song alone.** Clear temporal synchrony across the four trials can be seen for both cells, illustrative of an equally robust response to song stimuli. **D-E, I-J. Raster plots (D, I) and corresponding smoothed PSTH (E, J) showing the responses to song+ml-noise.** The pink highlights show the time when a different noise was present on each trial. This addition of noise destroys the cross-trial synchrony in the response for the neuron shown on the left column but not for the neuron shown in the right column. For this neuron, the response to song+ml-noise is very similar to the response of song alone, resulting in high noise invariance). **K. Photomicrograph of Nissl-stained brain slice in one bird showing the typical trajectory of the electrode penetration.** Electrode track, scale bar, and stereotactic axes are marked. **L. Scatter plot of noise invariance against stereotactic depth of neural recordings**. Noise invariance and recording depth were significantly correlated (slope = 0.15/mm, r = 0.38, p = 0.040): cells further along the electrode track are more invariant to masking noise. The example neurons have noise invariance values of 0.2 (left column) and 0.7 (right column)

similar spectra. NCM neurons have also been shown to respond differentially to natural and synthetic sounds and to familiar and non-familiar sounds but this is the first demonstration of

noise-invariant neurophysiological responses in this region and of a potential gradient along the dorsal/ventral axis.

Linear model invariance

To begin to understand how this system achieves noise invariance, I estimated the spetro-temporal receptive fields (STRFs) of each neuron from their responses to song (Theunissen et al., 2001). The STRF describes how acoustical patterns in time-frequency are correlated with a neuron's response. The STRF can also be as a model of the neuron to provide estimated neural responses for arbitrary sound stimuli. The STRF model is often described as "linear" but one should be aware that both static input (as the sound pressure waveform is transformed into a log spectrogram (Gill, Zhang, Woolley, Fremouw, & Theunissen, 2006)) and output non-linearities (a rectification) are part of this framework. To determine whether the STRF could explain noise-invariance, I performed a regression analysis between the mean noise-invariance $\hat{c}_d$ that I measured directly from the neurons response and the mean noise-invariance $\hat{c}'_d$ obtained from the predictions of STRF model (Fig. 1.2A). Two results come out of this analysis. First, the actual invariance and the model invariance are positively correlated showing that the neurons' STRFs could in part explain the observed noise-invariance. Second, I found that for most neurons, the degree of invariance predicted by the STRF model was actually greater than the one found in actual neurons. In other words, for a majority of neurons, additional non-linearities not captured in the STRF model make neurons less invariant. Although this result might seem surprising for an auditory region believed to be important for song recognition, it has a simple explanation. Many high-level neurons show adapting responses to sound intensity levels (Dean, Harper, & McAlpine, 2005) and this common non-linear response property is not captured in the simple STRF model. Intensity adapting neurons would exhibit a decrease in response to the song in noise relative to the song alone due to the adaptive changes in gain. The predicted response from the STRF does not incorporate this gain change yielding more similar responses for song and song+ml-noise stimuli than observed in the actual data.

Therefore, for the task of extracting the song from noise, the most effective neurons are those that are the most "linear" (the ones closest to the x=y line in Fig. 1.2A) or the few for which the non-linearity boost invariance (n=3/31). Although the specific non-linearities that could be beneficial for preserving signal in noise still need to be described, previous research have characterized higher-order non-linearities response that could play an important role: neurons in NCM exhibit stimulus specific adaptation (Stripling et al., 1997) and neurons in another avian secondary auditory area, CM (*Caudal Mesopallium*), respond preferentially to surprising stimuli (Gill, Woolley, Fremouw, & Theunissen, 2008). These non-linearities could facilitate noise invariance responses since they tend to de-emphasize the current or expected stimulus (in this case noise like sounds) without decreasing the gain of the neuron to sound at the same frequency.

Linear receptive field features

Since the STRF could for the large part explain the observed noise-invariance, I could then ask what feature of the neurons spectro-temporal tuning was important for this computation. A correlation analysis of the neurons' invariance with the neurons' average spectral and temporal modulation gain estimated from their STRFs showed that invariant neurons tend to preferentially respond to high spectral modulations and low temporal modulations (Fig. 2B-C). The ensemble modulation transfer functions estimated for different subsets of neurons further illustrate how the spectral and temporal modulation tuning co-vary along the noise-invariance dimension (Fig

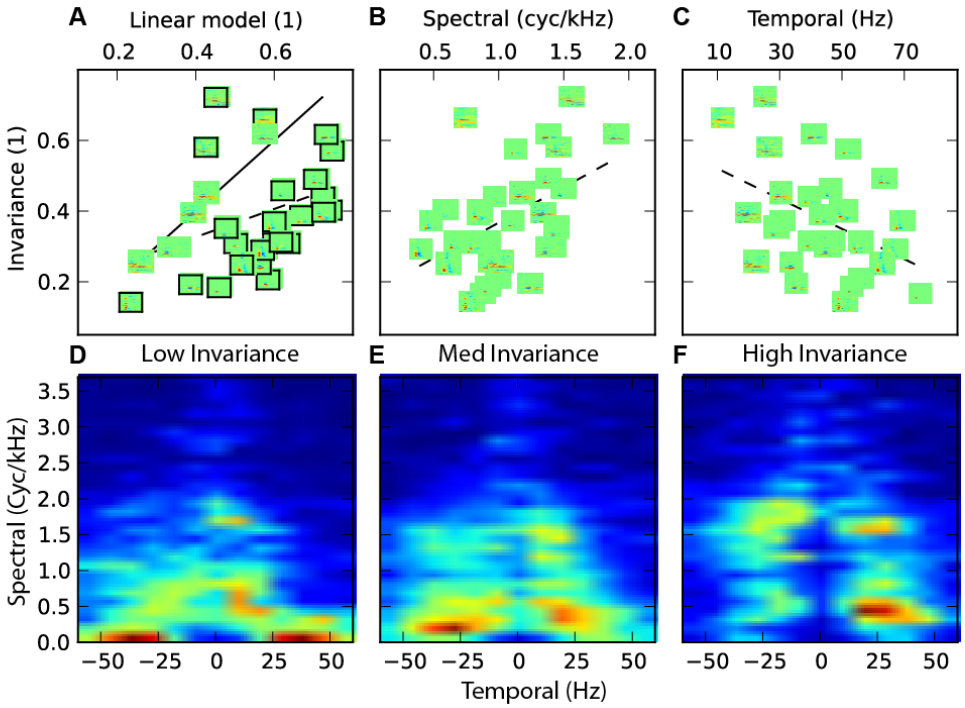


Figure 1.2: Spectro-temporal tuning and invariance. Vertical axis for A-C is the same and shows the noise invariance of the neural response. Each neuron (each point on the scatterplots) is represented by its STRF (0.25 – 8kHz on the vertical axis, 0-60ms on the horizontal). **A.** Invariance vs STRF Model Invariance. The solid line has slope 1.0, showing equal performance between the STRF model and the neural response. Neurons with significantly different performance ($p<0.05$, two-tailed t-test) have their receptive fields outlined. Dashed line shows regression fit (slope = 0.40, r = 0.38, p = 0.034), indicating the positive correlation between the invariance predicted by the STRF and the actual invariance on the neuron. **B.** Invariance vs Spectral Modulation Tuning. Neurons sensitive to higher spectral modulations are more invariant (r = 0.47, p = 0.007). The best spectral and temporal modulation frequency for each neuron was obtained from the center of mass of the modulation transfer function obtained from its STRF (SOM). **C.** Invariance vs Temporal Modulation Tuning. Neurons sensitive to lower temporal modulations are more invariant (r = -0.43, p = 0.015). **D-F.** Ensemble modulation transfer functions (SOM) for neurons grouped by invariance. Low invariance neurons (**D**, invariance<0.3, n=11) respond to high temporal frequency and low spectral frequency modulations. Neurons with moderate invariance (**E**, 0.3<invariance<0.4, n=11) transmit faster, sharper modulations. Neurons with high invariance (**F**, invariance>0.2, n=10) respond mostly to slower and spectrally sharp sounds

1.2D-F). Thus, noise invariant neurons exhibit the combination of longer integration times and sharp spectral tuning. In addition, the sharp excitatory spectral tuning was often combined with sharp inhibitory spectral tuning as well. These properties make noise-invariant neurons particularly sensitive to the longer harmonic stacks present in song (and other communication signals) even when these are embedded in noise as illustrated in the example neuron in fig 1 (F-J). Modeling studies (see Chapter 3) suggest, not surprisingly, that different set of invariant neurons would be found for different types of signal and noise but, at the same time, that similar tuning properties provide an efficient filtering of unwanted natural-like noise (e.g. colony noise) when processing a biologically relevant signal such as song.

Spike waveforms

I performed PCA on the normalized mean waveforms for the predictive, responsive neurons. I found two major clusters, and separated them with a linear discriminant on the first two principle components. Visual inspection shows that these clusters correspond to the two most common stereotypical waveforms immediately recognizable to any electrophysiologist (Fig. 1.3A). The slower waveforms (group 2) are significantly less invariant than the faster ones (group 1), and in fact no neuron in group 2 has an invariance higher than 0.5 (Fig. 1.3C). The model invariance is the same (Fig. 1.3D), meaning that the model improvement is not related to

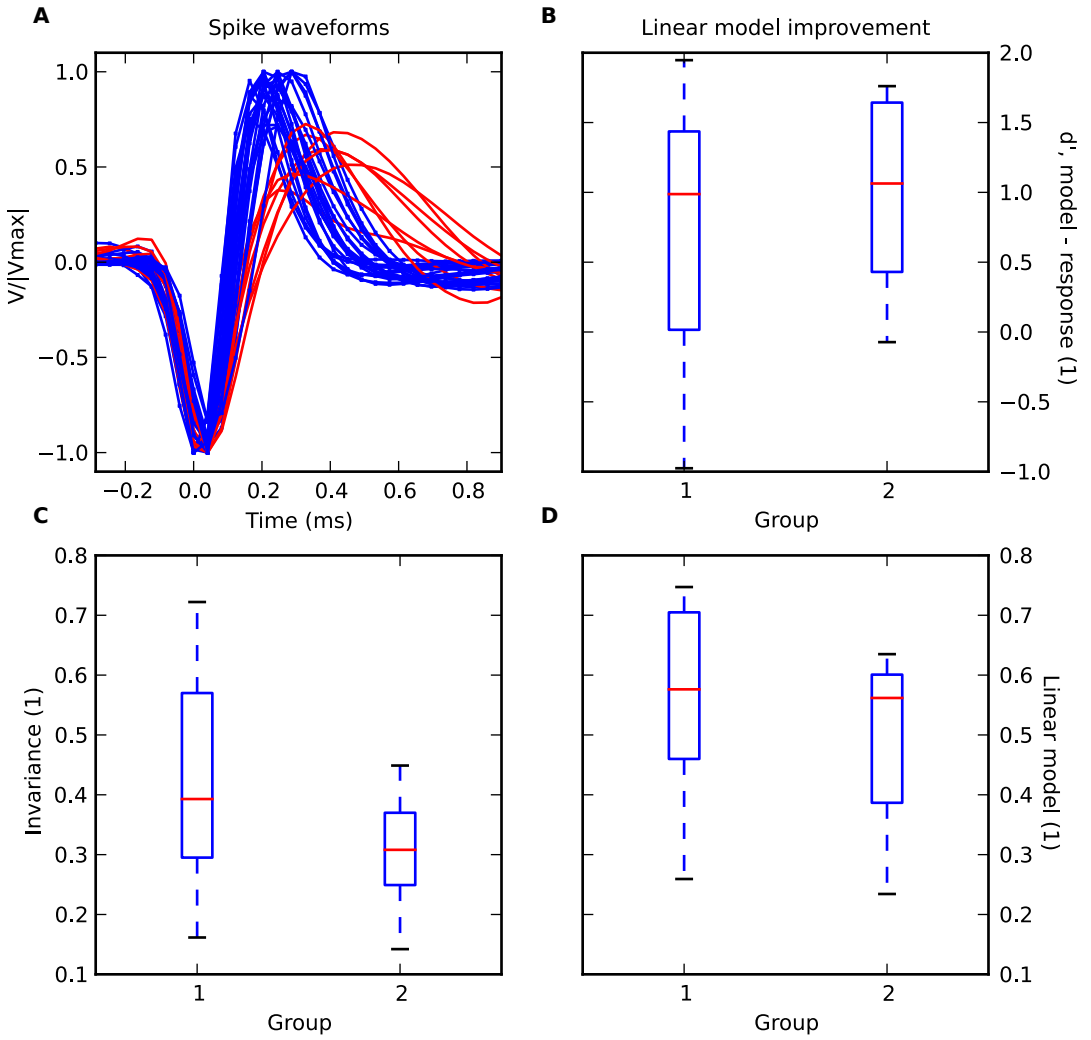


Figure 1.3. ***Spike waveforms***. **A**: Average spike waveform for each single unit in the population. Colors by PCA sort group: blue is group 1, red group 2. Group 1 has faster timescale, and has nearly equal positive and negative voltage deviations. Group 2 is slower and has a lower positive peak relative to the negative one. **B**: Improvement due to the linear model, shown as d-prime. The distribution of values is the same for both spike shape groups (p=0.663, Welch's two-tailed t-test approximation), meaning that both groups suffer similarly from the nonlinearities. **C**: Group 2 is less invariant than group 1 (p=.034, Welch's two-tailed t-test approximation). The most invariant neurons are all in group 1, while no neuron in group 2 has invariance > 0.5. **D**: Model invariances are similar for the two groups (p=0.161, Welch's two-tailed t-test approximation) again confirming that whatever differences exist between the two groups are not represented by the linear model.

which class of neuron is present. Unfortunately the distribution of model prediction invariances is somewhat skewed towards large values of $\hat{c}'_d$, making the use of the Welch approximation less ideal, but the sample size is too small for the Mann-Whitney test.

I also compared the discrepancies between the invariance of the linear model and the invariance of the neuron for each cluster. To quantify the improvement for the model, I computed a d' value between the measured invariance and linear model invariance for each unit:

$$d' = \frac{\hat{c}'_d - \hat{c}_d}{\sqrt{\hat{\sigma}'^2_d + \hat{\sigma}^2_d}} \quad \textbf{(1.25)}$$

The values for d' are also the same between the two clusters (Fig. 1.3B), confirming that the linear model improves the prediction just as much for the two groups. The distributions are somewhat skewed, though less so than for the model predictions alone; the extremely high p-value makes it highly unlikely that the means differ.

Discussion

The generation of neurons with invariant responses is not a trivial task since most neurons in lower auditory areas have much shorter integration times and lack the sharp excitation and inhibition along the spectral dimension that I observed here. From comprehensive surveys of tuning properties from neurons in the avian primary auditory cortex (Field L) (Nagel & Doupe, 2008; Woolley et al., 2009), we know that a small number of neurons with similar characteristics exist in these pre-synaptic areas (the slow narrow-band neurons in (Woolley et al., 2009)). Similarly, in the mammalian system, neurons in A1 have been shown to have a range of spectro-temporal tuning similar to that seen in birds but few with the sharp spectral tuning seen here (Depireux, Simon, Klein, & Shamma, 2001; Miller, Escabi, Read, & Schreiner, 2002). Thus it is reasonable to postulate that noise-invariance in NCM (and putatively in mammalian secondary auditory cortical regions) is the result of a series of computations that are occurring along the auditory processing stream. However, it is also known that NCM possesses a complex network of inhibitory neurons and that these play an important role in shaping spectral and temporal response properties (Pinaud et al., 2008). Thus both upstream and local circuitry are almost certainly involved in the creation of noise-invariant neural representations. If neurons in spike shape group 2 represent inputs from other areas, this would indicate that invariant representations are being built up in NCM.

Such an invariant representation could help solve the cocktail party problem. Elhilali et. al. have shown that a neurally-inspired spectrotemporal basis can achieve stream separation (Elhilali & Shamma, 2008). Smaragdis found a similar result using a basis found by decomposing the spectrogram using non-negative matrix factorization (Smaragdis, 2007). Elhilali's work makes very explicit reference to the scales of the spectrotemporal modulations represented in the basis, and refers to a "rate-scale" space. That space is, except for the log-transformation of the frequency scale, identical to the modulation power domain as I have defined it. As I discussed in the introduction, the characteristic features of speech are well constrained in this domain by the physical form of the vocal apparatus.

One possible improvement for future work would be to extend the invariance computation. The PSTH similarity measure of invariance accords well with visual comparison of the spike trains. A more thorough version could extend the metric to use coherence rather than just the correlation, again jackknifing across trials to obtain an error bar. Because the signals are relatively short, the problem is well-suited to the use of the multitaper coherence (Thomson, 1982).

This study also points a way forward in thinking about cortical/forebrain auditory processing. In vision, many higher areas have been shown to hold invariant representations of high-level information, including identity (Quiroga, Reddy, Kreiman, Koch, & Fried, 2005), motion (Priebe, Cassanello, & Lisberger, 2003), and faces (Freiwald & Tsao, 2010). Higher auditory areas, beyond A1, are often seen by physiologists as being unknown territory. Several strong invariance are documented in the behavioral and psychophysical literature, among them speaker identity, pitch, level, and semantic content. Explicit investigation of such features at a neural level could help to shine a light on the function of these areas.

In summary, I have shown that noise-invariant neurons exist in secondary auditory areas and have explained how this invariance can to a large part be explained by the neurons' spectro-temporal modulation tuning and simple non-linearities. I return to this line of reasoning in Chapter 3, where I develop a speech processing system based on these principles.

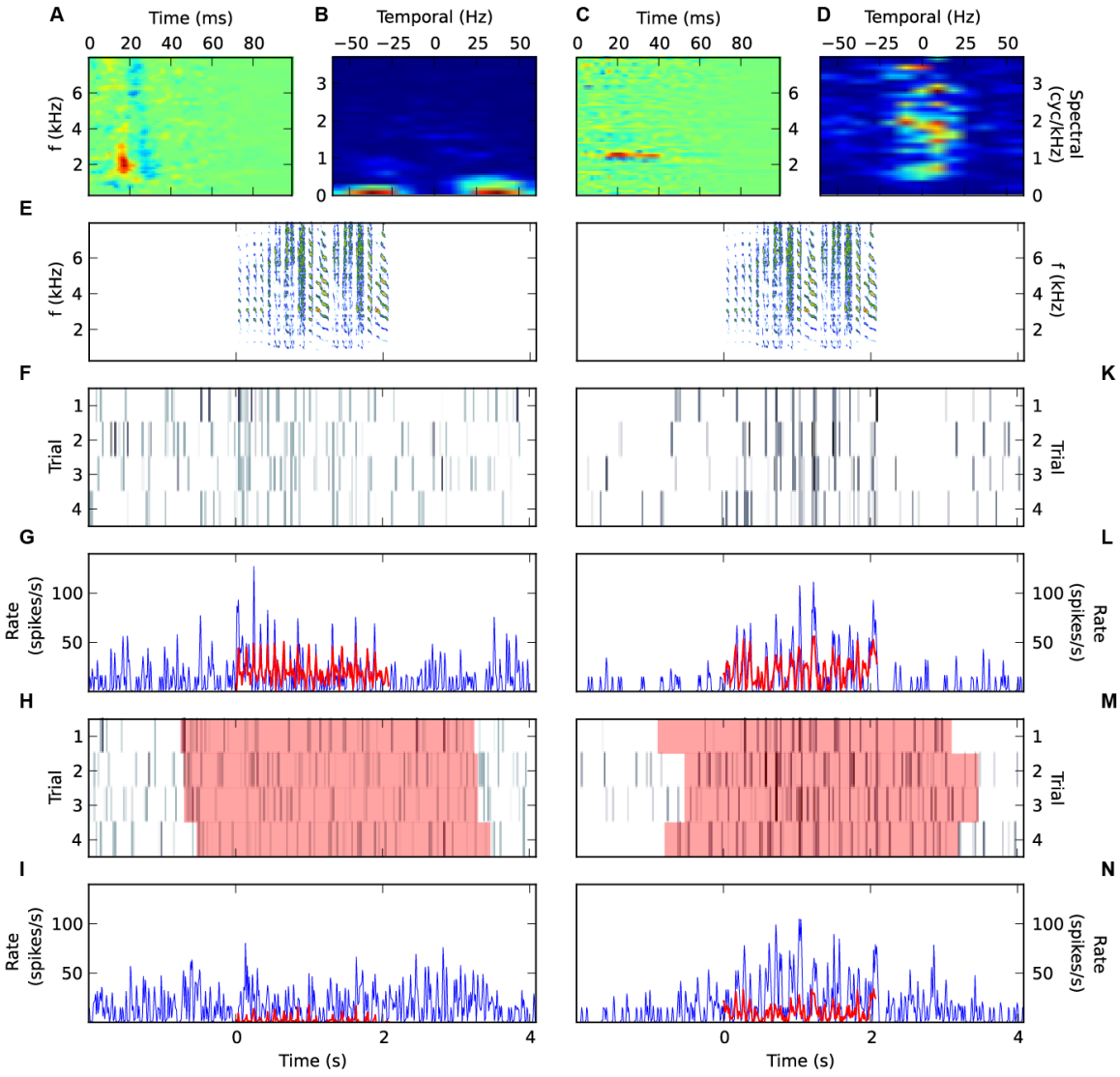


Figure 1.4: Comparison of STRF predictions for two cells. **A**: STRF for a low noise-invariant cell (invariance = 0.25). The STRF predicts an average of 14.6 bits/second of mutual information with the response, 56% of the maximum (defined by the expected mutual information between two trials). **B**: MPTF for the STRF in A. Modulation power is concentrated on the temporal axis near ±35 Hz and 0.1 cycles/kHz, indicating sensitivity to fast, broadband modulations. **C**: STRF for a high noise-invariant cell (invariance = 0.65). The STRF predicts an average of 14.4 bits/second of mutual information with the response, 53% of the maximum possible. **D**: MPTF for the STRF in C. Energy is clustered along the spectral axis between 1.5 and 3 cycles/kHz and less than 20Hz, indicating high spectral modulations and slow temporal ones. **E**: Spectrogram of one ZF song stimulus. **F**: Spike raster for cell from A. **G**: Response of low noise-invariant cell to unmasked song. PSTH shown in blue, prediction of STRF from A shown in red. **H**: Spike raster for response of low noise-invariant cell to masked song. Regions of each trial where masker was present are indicated in pink. **I**: Response of low noise-invariant cell to masked song. PSTH shown in blue, prediction of STRF from A to same stimulus shown in red. Here, both the actual response and the prediction contain significantly less information, indicating a reduction both in rate and in trial-to-trial reliability. The STRF predicts 6.5 bits/second, 71% of the possible mutual information. **J**: Spectrogram of same ZF song stimulus, shown again for clarity. **K**: Spike raster for the high noise-invariant cell shown in C. **L**: Response of high noise-invariant cell to unmasked song. PSTH shown in blue, prediction of STRF from C shown in red. **M**: Spike raster for response of high noise-invariant cell to masked song. Regions of each trial where masker was present are indicated in pink. **N**: Response of cell C to masked song. PSTH shown in blue, prediction of STRF shown in A to same stimulus. The neuron's response is relatively unchanged by the presence of masking noise, although the STRF does not predict as well. This is one of the three cells for which the measured invariance exceeded that of the STRF model.

Chapter 2

Response to Spectrotemporal Modulation Filtering in Avian Auditory Cortex

Abstract

Communication sounds in vocal animals are characterized by spectral and temporal modulations. Psychophysical studies have determined the ranges of modulations that are perceptible by humans, and the subset within this range that are essential for speech comprehension. Here, I examine the neural representation of the vocalization-specific modulations in the songbird model. To determine whether neural responses are tuned to specific modulations, I use two complimentary approaches. First, I estimate the spectro-temporal (STRF) receptive field of each neuron from responses to vocalizations signals. The modulation tuning can then be derived from the STRF by examining the gain of this filter in the Fourier domain. Second, neural responses are obtained to versions of the same vocalizations that have been systematically degraded by removing particular spectral or temporal modulations. This signal degradation is obtained by a novel-filtering method that allows us to perform filtering operations in the modulation domain. If the STRF is a good characterization of the modulation tuning of the neurons, it can be used to predict responses to the degraded signals. Differences between the predictions and the actual responses reveal additional sensitivity for these modulations. I find that the response to spectrotemporal degradation is largely predicted by the features of a linear STRF fit for each neuron.

Introduction

One of the primary physical models to explain sound production in vocal animals is the source-filter model (Gold & Morgan, 2000; Taylor & Reby, 2010). As I discuss in the introduction, the computation of the modulation power spectrum effectively separates generative parameters of speech, the spectral profiles of the source and filter, by performing deconvolution. In the previous chapter, I investigated how these parameters contribute to a noise-invariant representation of vocalizations post hoc, by inferring which parameters of the MTF contribute to noise invariance.

In this chapter, I instead modify the modulation-domain characteristics of the stimuli beforehand, and then examine how this processing changes the responses. Because of the deconvolution shown in equation i.4 and the different spectral modulation scales of the source and filter, modulation filtering in the spectral domain can have the effect of changing the source or filter.

Changes in the modulation domain are perceptible to humans. With white noise carriers, slight temporal amplitude modulations up to 16Hz are easily noticeable, with thresholds of -25dB or about 5% of the amplitude. Higher temporal AM frequencies are more difficult to

detect: above 16Hz, thresholds increase to a plateau of roughly -5dB, or about 50% of the amplitude, at 2kHz. Pure-tone carriers have similar behavior at 10kHz, with higher thresholds (-15dB) but a similar plateau out to 16Hz. Detection thresholds rise, becoming less sensitive, and higher modulation frequencies become less detectable, as carrier frequency decreases (Viemeister, 1979).

For some more spectrally complex carriers, detection performance falls off—i.e., thresholds rise—above about 10Hz (Yost, 1987). Combined spectrotemporal modulations are perceptible throughout a region from at least 0-128Hz and 0-8 cycles/octave, and generally comport with the time-domain studies mentioned above as well as frequency-modulation-only studies (Chi, Gao, Guyton, Ru, & Shamma, 1999).

With speech sounds, modulations can also be characterized by their importance for intelligibility. Modulations below roughly 10Hz are required for understanding or recognition of individual phonemes when either spectral (Drullman, Festen, & Plomp, 1994) or cepstral (Arai, Pavel, Hermansky, & Avendano, 1999) quantities are modulated. Modulations above roughly 16Hz can be attenuated without substantial impact on intelligibility. Spectral smearing up to ½ octave, corresponding to roughly 0.5 cycles/kHz for a 1kHz fundamental, can be performed without degrading comprehension.

In this experiment, I characterize the sensitivity of neurons to spectral and temporal modulation degradations much like the ones reported for the human speech listening experiments mentioned above. By sampling from throughout avian auditory forebrain using a multichannel electrode, I am able to construct a picture of how these modulations are processed throughout the auditory pathway.

Methods

Modulation filtering

To perform modulation filtering, I used code written in our lab and described previously (Elliott & Theunissen, 2009).The filtering consists of five steps. First, we compute the spectrogram (STFT) of the signal per equation 1.17, separating the amplitude and phase components. Second, we compute the complex modulation spectrum by taking the 2D Fourier Transform of the spectrogram amplitude per equation i.6. Third, we apply a gain between 0 and 1 to the modulation spectrum, with 1 for the bands we want to preserve and 0 for the bands we want to attenuate. Fourth, we invert the filtered complex modulation spectrum to produce a filtered spectrogram amplitude. Fifth and finally, we invert the spectrogram, using the filtered amplitudes and the original phase as a first guess for the iterative algorithm described in (Griffin & Lim, 1984).

Stimulus protocol

In these experiments a search protocol (consisting of search stimuli) was not used as recordings were obtained in a systematic fashion with a 16-microelectrode array at depths separated by 100 microns (see below for more details). At each recording location two conspecific songs and two 2s samples of white noise were then used to set voltage thresholds for each electrode for on-line spike discrimination.

After setting the threshold, I played a stimulus protocol designed for STRF estimation and invariance computation. This protocol consisted of a total of 40 sounds: 10, two-second

modulation-limited noise stimuli, generated using the procedure described in Chapter 1; 10 conspecific songs; and two modulation-filtered versions of each song. At each recording position, I used one of two different sets of ten songs.

One modulation filtered version, which I will refer to as "sfilt," was lowpass filtered in the spectral domain below 0.6 cycles/kHz, with a 0.1 cycle/kHz ramp from 0 gain to 1 (Fig 2.1A). Compared to the MPS of the original songs (Fig. 2.1B), these stimuli show the characteristic triangular shape near the origin but are missing the power in the songs that extends along the spectral modulation axis. They are also missing the concentration of power around 1.5 cycles/kHz that corresponds to the harmonic stacks.

The spectrograms for these stimuli (Fig. 2.4A top panel) show temporally sharp features that are smeared in the spectral domain. This preserves the region of modulation power space containing the formants and the syllable edges (Elliott & Theunissen, 2009). Harmonic features are removed, replaced by broadband noise. The perceptual quality of these sounds is similar to those obtained with a noise vocoder or whispered speech (Gold & Morgan, 2000).

The other modulation-filtered version, which I will refer to as "tfilt," was lowpass filtered in the temporal domain below 7 Hz, with a ramp of 1 Hz (Fig. 2.1C). In this case, the only energy is close to the spectral modulation axis, preserving only very slowly varying features. I chose the frequency of 7 Hz to blend information across syllables, based primarily on the temporal statistics of the stimulus. The spectrograms of these stimuli (Fig. 2.4C top panel) show slow features, with some sharp spectral components present where they appear in several syllables in sequence.

I used custom TDT software, mostly the same as what is described in Chapter 1, to play the stimuli in random order. I played ten trials; each trial had each sound played once in random order, with a 5-7 second random delay between stimuli.

Physiology

For this experiment, I recorded 101 single units from 3 birds using a 16-channel tungsten microwire electrode fabricated to order by Tucker Davis Technologies. Each array comprises 16 individual 30µm tungsten wires coated in parylene. The wires are arranged in rows of 8, spaced 250µm apart, for a total width of 1.75mm.

The two rows are mounted on a Printed Circuit Board (PCB) so that they are 375µm apart. The wires in some cases extended as much as 15mm from the edge of the PCB, but in all cases they were potted together with epoxy every 5mm by the manufacturer to prevent buckling. Some electrodes only extended 5mm from the board edge, in which case they were only potted to the board. I added epoxy between the board and the last epoxy land on the 15mm electrodes to improve stiffness.

The individual wires are laser-cut to length with a 45° bevel after being mounted on the PCB. The PCB is connected mechanically and electrically to the headstage by an 18-channel micro connector. For one of the three birds, bird BlaW0603, I used an electrode with tips staggered by 500µm. Within each row, the wires alternated between 4.5 and 5.0mm from the edge of the board.

The electrode also has a reference/ground wire (connected on the PCB to both the ADC reference and the signal ground). To make electrical contact with the CSF, I wrapped this wire around the shank of the screw that fixed the pin to the stereotax and then tightened the screw head down onto the loop.

Surgery

I performed preparatory surgeries quite similar to the ones described in Chapter 1 for pin placement. I removed a substantially larger region of the outer skull over the left hemisphere to make room for the additional size of the 16-channel array. The craniotomy measured roughly 2.5-3.0mm lateral of the midline on the left, across the midline on the right far enough to expose the central sinus, and 1.5-2.0mm rostro-caudal. On one bird, GrayGray1516, I removed a larger region of outer skull (and subsequently of inner skull, dura, and pia), extending the craniotomy roughly 1mm more rostral to allow for a second electrode penetration.

Because the aperture crossed the midline, I placed the pin laterally offset to the right of the craniotomy. With the top layer of skull removed, I marked two fiducial dots: both 0.5mm lateral of the Y-sinus, and 1.0 and 1.2mm rostral of the same.

On the day of recording, I administered the same sequence of urethane injections described in Chapter 1. 30 minutes after the final injection, I placed the bird in the stereotax and fixed the pin in place. I mounted the electrode and microdrive onto the stereotax, and used the coarse vertical adjustment to bring the electrode array near the surface of the inner skull. Under a stereomicroscope, I aligned the array with the fiducial dots on the inner skull and then used the microdrive to move the array up and away from the skull. I then used a scalpel to remove the inner skull as close to the edges of the craniotomy as possible.

I used two custom-formed tungsten minuten pins to remove the dura and pia. The first is a simple 90° turn, the second a 180° hook. The dura has a fine fibrous structure much like muscle, making it relatively easy to tear it along the grain in a rostro-caudal direction. Once I had made a sufficiently long tear, I was able to retract the dura using one of the pins, or to cut away additional portions by holding it with a pin and cutting with a pair of iridectomy shears. As with the inner skull, retracting the dura as close to the edges of the craniotomy as possible made subsequent operations easier. Removal of the pia was somewhat difficult under the 12X magnification of the stereomicroscope. The pia itself is too transparent and too thin to be seen even under these conditions, but the smoothness and the presence of blood vessels are useful indicators. By contrast, the brain surface below appears white and waxy, and the blood vessels are larger and farther between.

After removing the pia, I placed saline over the craniotomy to keep the tissue from drying out. I then used a small forceps to brush crystals of DiI (Invitrogen, catalog #D3911, CAS 41085-99-8) onto the tip of each electrode wire for postmortem histological reconstruction. I advanced the electrode array into the saline using the microdrive and advanced it towards the surface of the brain, then removed the saline to clearly see when the electrodes touched the surface. When they did, I reset the microdrive depth counter to zero, then continued to advance the electrode.

Once all of the electrode wires had entered brain, I advanced the microdrive roughly 300μm further before placing electrode gel onto the craniotomy. I spread the gel out to contact both the exposed metal of the pin and the brain surface, then added a small amount of saline to cover the brain. At this point I moved the stereotax into the sound booth and connected the wires for the microdrive and grounding wires through the port. I adjusted the stereotax so that the bird's head was roughly 20cm from the loudspeaker. I then closed the door to the booth.

When recording with the 16-channel electrode, there is a good chance of getting a single unit on at least one of the channels regardless of where the electrode is positioned, making it unnecessary to move the electrode to find cells or optimize recording quality. At the beginning of the experiment, I moved the electrode down until I saw stimulus-evoked activity on at least one

channel. I then waited at least five minutes before beginning to record. After each site, I advanced the electrode 100μm and waited another five minutes for the electrode to settle.

Analysis

I performed much of the basic analysis for this experiment in the same manner described in Chapter 1, including spike sorting, assessment of responsiveness, and STRF fitting. Spike sort quality was generally not as good, most likely because of the lower impedance of the electrodes. I computed per-class zscores from equation 1.2 and used a cutoff of $\bar{z} \geq 1.5$ to identify potentially responsive units. For each potentially responsive unit, I fit a STRF for each stimulus class using STRFLAB's direct fit algorithm to solve equation 1.18. I assessed predictive power in the same way as in Chapter 1: I considered STRFs which predicted at least 1.2 bits/second and at least 20% of the response information to be predictive.

Jackknife-bias-corrected invariance calculation

Following the method described in Chapter 1, I calculated two invariance metrics for each unit. For each of the 10 song stimuli in the protocol, I compared the response to the control stimulus $\left\{r_{ij}\right\}$ with the response to the sfilt or tfilt version of the stimulus, denoted as $\left\{r_{ij}^{s}\right\}$ and $\left\{r_{ij}^{t}\right\}$, respectively. I will extend this superscript notation, with '*s*' denoting sfilt stimuli and '*t*' denoting tfilt stimuli, throughout this chapter. The mean PSTH $\bar{r}_i$ for the control stimulus is constructed according to equation 1.3, and the mean responses $\bar{r}_i^s$ and $\bar{r}_i^t$ can be calculated in a similar fashion. In analogy to equation 1.5 we can similarly construct set of delete-d jackknife PSTHs for the sfilt and tfilt responses

$$\bar{r}_{idk}^{s} = \frac{1}{n_{trials} - q} w * \sum_{j \in q_k} r_{ij}^{s} \qquad \textbf{(2.1)}$$

$$\bar{r}_{idk}^{t} = \frac{1}{n_{trials} - q} w * \sum_{j \in q_k} r_{ij}^{t} \qquad \textbf{(2.2)}$$

where q_k is again defined per equation 1.6.

We can construct jackknife estimates of the correlation coefficients in analogy to equation 1.7:

$$c_{idk}^{s} = \mathrm{corr}\left(r_{idk}, r_{idk}^{s}\right) \qquad \textbf{(2.3)}$$

$$c_{idk}^{t} = \mathrm{corr}\left(r_{idk}, r_{idk}^{t}\right) \qquad \textbf{(2.4)}$$

Finally, we can compute the estimates $\hat{c}_d^s$ and $\hat{c}_d^t$, errors $\hat{\sigma}_d$, and confidence intervals δ_a from equations 1.12-1.14.

Modulation passband power ratio

To estimate how much degradation of the neuron we should expect, I computed the Modulation Transfer Function (MTF) of each STRF (see Chapter 1) and compared it to the modulation-domain filtering function used for temporal and spectral degradation. The ratio of the power in the passband of the modulation lowpass filter to the total power of the MTF gives a

sense of how much the filter should be affected if the linear model is accurate. I called this the Modulation Passband Power Ratio, MPPR, denoted by ρ:

$$\rho = \frac{\sum_{\omega_s,\omega_t} m(\omega_s,\omega_t) M(\omega_s,\omega_t)}{\sum_{\omega_s,\omega_t} M(\omega_s,\omega_t)} \quad \textbf{(2.5)}$$

where $m(\omega_s,\omega_t)$ is the modulation domain filtering function, and $M(\omega_s,\omega_t) = \left|\mathrm{FT}(h(f,t))\right|$ is the modulation transfer function.

Results

Of the 100 single units, 55 were responsive and had predictive STRFs. Many units were substantially unaffected by the spectral (sfilt) filtering. Invariance $\hat{c}_d^s$ ranged roughly from 0.25 to 0.9. At the high end, units appear to be largely unaffected by the filtering. There is a significant, negative relationship between spectral modulation frequency and invariance to spectral modulation filtering (Fig. 2.2A). This means that units that are sensitive to high spectral modulations are more likely to be affected by the removal of those modulations. The relationship is not absolute, though: some cells with sharp spectral modulation qualities are still quite invariant to the removal of such modulations.

Invariance to spectral modulation filtering is also significantly negatively correlated with increasing temporal modulation frequency (Fig. 2.2B). Cells with long integration times are much more likely to be invariant to spectral modulation filtering. There are substantial populations of units with very slow MTFs, all of which have invariance above 0.5. Similarly, the cells with temporal frequencies above 70 Hz all have invariance below 0.5.

In contrast to the spectral results, I see much less invariance to temporal modulation filtering: values for invariance $\hat{c}_d^s$ are between 0 and 0.5. Even the very slow cells did not have high invariance; in fact, the most invariant cells have temporal modulation frequencies between 30 and 60Hz (Fig. 2.2D). The temporal invariance is not significantly correlated with either the spectral modulation frequency (Fig. 2.2C) or the temporal modulation frequency (Fig. 2.2D).

Passband power

The linear STRF model does predict the degree of invariance to modulation filtering. STRFs having all of their power in the modulation passband, i.e. with $\rho = 1$, should predict exactly the same response with filtering and without, resulting in an invariance $\hat{c}_d = 1$. STRFs having all of their power in the stopband ($\rho = 0$) should predict no response at all and thus an invariance $\hat{c}_d = 0$.

To visualize this, I compared the invariance to spectral and temporal modulation filtering for each unit with the spectral and temporal modulation passband power ratio, respectively. The results should be quite similar, because the MPPR and the center of mass should be highly correlated. This is nearly true for the spectral case, where the MPPR is correlated with the spectral modulation center of mass with r=-0.95. The MPPR for temporal filtering is less well correlated with the temporal center of mass, r=-0.73. This lower number is may be due to the fact that the temporal MPPRs are all so low: if none of the STRFs have much power in the

modulation passband, then the MPPR is relatively constant, while the centers of mass have a wide range.

Because of the correlation between the spectral modulation and the spectral MPPR, plotting the sfilt invariance against the two gives nearly the same picture (compare Fig. 2.2A with Fig. 2.3A). The spectral MPPR correlates roughly as well with the spectral invariance as the spectral modulation frequency does: r=0.32 vs. r=0.31.

With the temporal invariance, however, the arrangement of the corresponding plots is somewhat different. As with temporal modulation frequency, there is no correlation between temporal MPPR and tfilt invariance, but the relationship between the temporal MPPR and the invariance (Fig. 2.3B) is qualitatively different than the relationship between the temporal modulation frequency and the invariance (Fig. 2.2D). This difference arises from the imperfect correlation between the MPPR and the temporal center of mass.

Low temporal invariance units

The units that had low invariance to temporal modulation filtering have some interesting differences. Some neurons, like unit 1497 (Fig. 2.4, middle row), retain substantial responsiveness to filtered songs, but have that response smeared out in time. Others, unit 3018 (Fig. 2.4, bottom row), have their responses virtually abolished by the temporal filtering and retain only an onset response. Both of these neurons have low invariance, 0.13 and 0.06, respectively.

Discussion

We found that many neurons had substantial invariance to spectral degradation, and that this was correlated with the spectral MPPR. For temporal degradation, most neurons were significantly degraded, either by having their response abolished, or by having it smeared in time. While the correlation between spectral center or MPPR and sfilt invariance is expected, it leaves some questions. For one, the trend towards lower invariance with higher modulations or lower MPPRs is not absolute: a significant number of units have low invariance despite having very broad spectral tuning and very high MPPRs.

Additionally, the correlation between the temporal modulations and the spectral invariance is surprising. This could represent a number of things. It could mean that cells with slow integration times are doing computations similar to those covered in Chapter 1, extracting slow pitch features. This poses problems, though, because many of these cells exhibit precisely the type of sharp spectral tuning that we expect to be degraded by the lowpass modulation filtering. Further examination of the exact timing of the responses, and comparison with the model predictions, could provide better answers about this.

The presence of so much invariance to spectral filtering at all is also somewhat surprising. In the case of the temporal degradation, the MPPR is entirely equivalent to the linear model: the STRF is convolutional in time (LTI), meaning that the Fourier transform is multiplicative and thus that absolute temporal modulation phase does not matter, only relative phase. In the spectral domain, on the other hand, the STRF is stationary, meaning that the Fourier transform does not decompose so neatly. In the spectral domain, that is, the absolute spectral phase matters, even though the MTF and MPPR do not account for it. One might expect this approximation to cause a deviation from the model in the spectral domain, but in fact the results show the opposite, that the temporal domain behaves more strangely.

For the spectral domain, one possible explanation would be the absence of out-of-phase harmonic stacks. The stacks that we do see are all in cosine phase in the modulation domain, i.e., the power peaks at integer multiples of the fundamental. The presence of only one single absolute spectral phase could explain why the MPPR, which disregards absolute FM phase, still matches the invariance for higher frequency spectral modulations. For lower frequency modulations, the phase spectrum is more complicated, because unlike the pitch peaks, the formant peaks that these modulations represent are not constrained to have any particular phase.

Low frequency temporal modulations

I chose the spectral and temporal lowpass cutoffs based on the features of the ensemble modulation power spectrum (eMPS) for song (Fig. 2.1B). In fact, while the ensemble modulation transfer function (eMTF) for the STRFs to song matches the eMPS in many ways, it differs in a few crucial ones. Despite the considerable power in the eMPS below 10Hz and extending from the origin to 2 cycles/kHz, the eMTFs for areas L, MLD, and CM show very little tuning in this region of very slow modulations.

Accordingly, the measured temporal MPPRs are quite low, roughly between zero and 0.2. This low range means that the temporal degradations should, in theory, have affected all cells equally: if no modulations in any neuron's preferred range are present, all neurons will have the same nil response. In fact while the range of invariances for the temporal degradation is 0.0-0.5. That is, the temporal degradation did affect all of the cells substantially, but the wide range of invariances means that it did not affect them all equally.

While some cells showed more invariance to temporal modulation filtering, there were also substantial differences among the responses of cells with low temporal invariance, as show in Fig. 2.4. The cell in the bottom row has low invariance because it has its response largely abolished. The cell in the middle row still responds, but because the temporal scale is substantially changed, the invariance is still low. It is possible that a coherence-based invariance metric would capture some similarity between low-frequency response modulations. There could be a high correlation between a low-passed version of the spike trains obtained to the unfiltered stimulus and the response obtained to the tfilt stimulus.

In summary, this experiment shows that the linear STRF partially predicts the response of cells to spectral modulation filtering, and that many auditory neurons are very resilient to this type of processing. Many cells are invariant to spectral degradation, and that performance is well predicted by the linear model. In contrast, the effects of filtering in the temporal domain remain less clear, and some neurons appear to be more invariant than expected.

The temporal result is interesting in light of the psychophysics, because the neurons appear to be very sensitive to the loss of modulations above 7Hz, well within the range for which they are easily detectable for humans (Viemeister, 1979). The timescale for Zebra Finch song is somewhat faster, though in theory this should only push the range of easily-detectable modulations higher. The relative insensitivity to spectral degradations is somewhat at odds with the fine spectral sensitivity shown in behavioral experiments (Cynx, Williams, & Nottebohm, 1990).

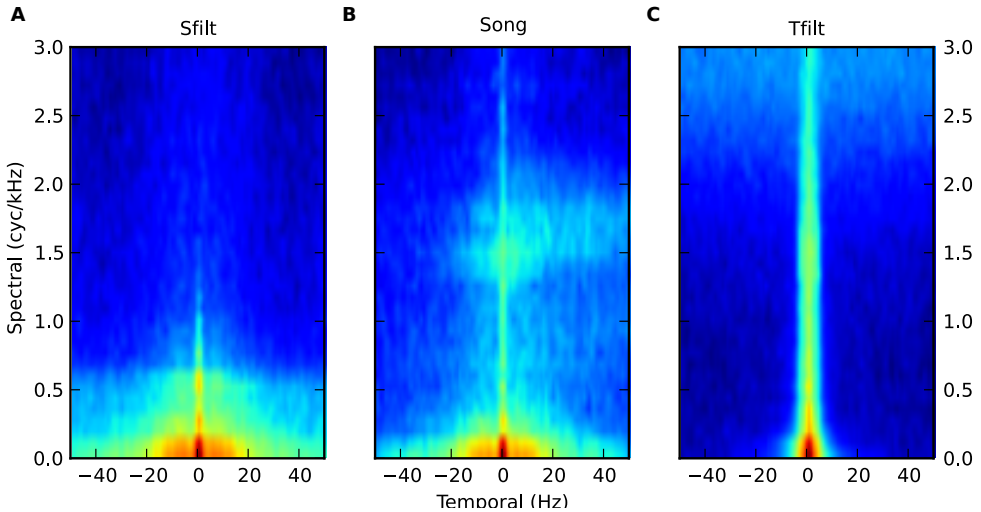


Figure 2.1: Modulation Power Spectra of Stimuli. All MPS are computed from spectrogram samples windowed with a 1 second Gaussian window. Color axis is log scale. **A: MPS of song after lowpass spectral modulation filtering.** Ensemble modulation power spectrum (eMPS) for 20 zebra finch songs that have been lowpass filtered below 0.6 cycles/kHz. Lowpass cutoff is clearly visible, although some leakage at low temporal frequencies is apparent, most likely as a result of the iterative spectrogram inversion process. The high power region between -20 and 20Hz and 0 and 0.5 cycles/kHz contains formant and syllable transitions, and is well preserved in this case. **B: MPS of unfiltered song.** eMPS for the same 20 zebra finch songs before processing. The formant/syllable region in A is apparent, as is the additional concentration of pitch energy around 1.5 cycles/kHz, representing harmonic complexes with a fundamental near 666Hz. There is energy in this region extending past 10Hz, corresponding to chirped syllables around this pitch. Energy is also present above and below the region, but is restricted to be closer to the spectral modulation axis and corresponds to steady notes at other pitches. **C: MPS of song after lowpass temporal modulation filtering.** eMPS for the same 20 zebra finch songs having been lowpass filtered below 7Hz. All of the formant and syllable transitions have been removed, leaving a profile in this region corresponding to the average formant across all 20 songs.

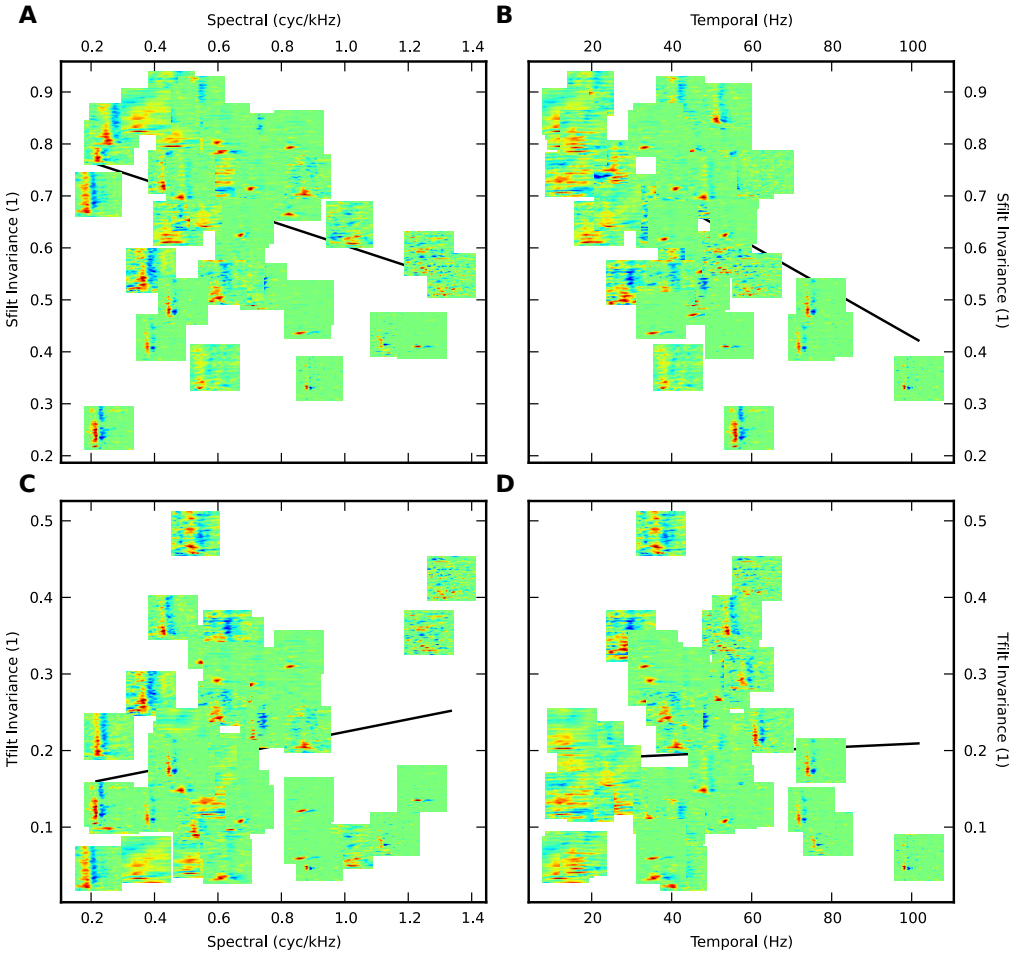


Figure 2.2: Spectral Modulation Center Frequencies for Single Units Predict Modulation Filtering Invariance. For all four plots, units are represented by a small image of their STRF. Modulation center frequencies are computed as the coordinates of the center of mass of the neuron's modulation transfer function. **A: Neurons sensitive to high spectral modulation frequencies are less invariant to lowpass spectral filtering.** The response of neurons with sharp spectral tuning changes when stimuli are filtered to remove sharp spectral features ($r=-0.31$, $p = 0.021$). **B: Neurons sensitive to high temporal modulations are less invariant to lowpass spectral filtering.** This result is unexpected: neurons with long integration times are significantly more resilient to spectral modulation filtering ($r=-0.50$, $p < 0.001$). **C: Changes in neural response to lowpass temporal filtering are unrelated to the spectral modulation characteristics of the cell.** All neurons have their responses degraded by severe temporal modulation lowpass filtering, regardless of their spectral characteristics ($r=0.192$, $p=0.16$). As mentioned in the results, the temporal modulation filtering cutoff of 7Hz is below the response range of the neural ensemble. Despite this, some neurons do retain some information, **D: Temporal center frequencies do not predict degradation by severe temporal modulation filtering.** As in C, the remaining temporal modulations fall outside the range of any of the neurons, and the invariance is entirely uncorrelated with the center frequency ($r=0.042$, $p=0.762$).

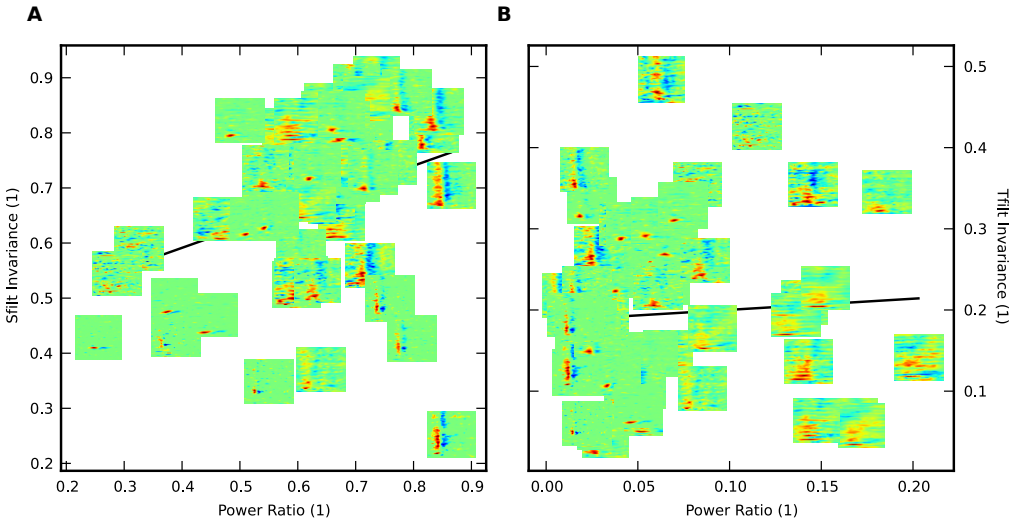


Figure 2.3: Spectral MPPR Predicts Modulation Filtering Invariance. A: Neurons with high MPPR to lowpass spectral filtering are more invariant. Invariance to spectral filtering is correlated with spectral MPPR (r=.321, p = 0.017). Units are distributed in a very similar fashion to Fig. 2.2A. **B: MPPR to lowpass temporal filtering is uncorrelated with invariance.** Invariance to temporal filtering is uncorrelated with temporal MPPR (r=.068, p=0.623). Because temporal MPPR is only partially correlated with temporal modulation frequency for these stimuli, the arrangement of the units differs somewhat from Fig. 2.2D.

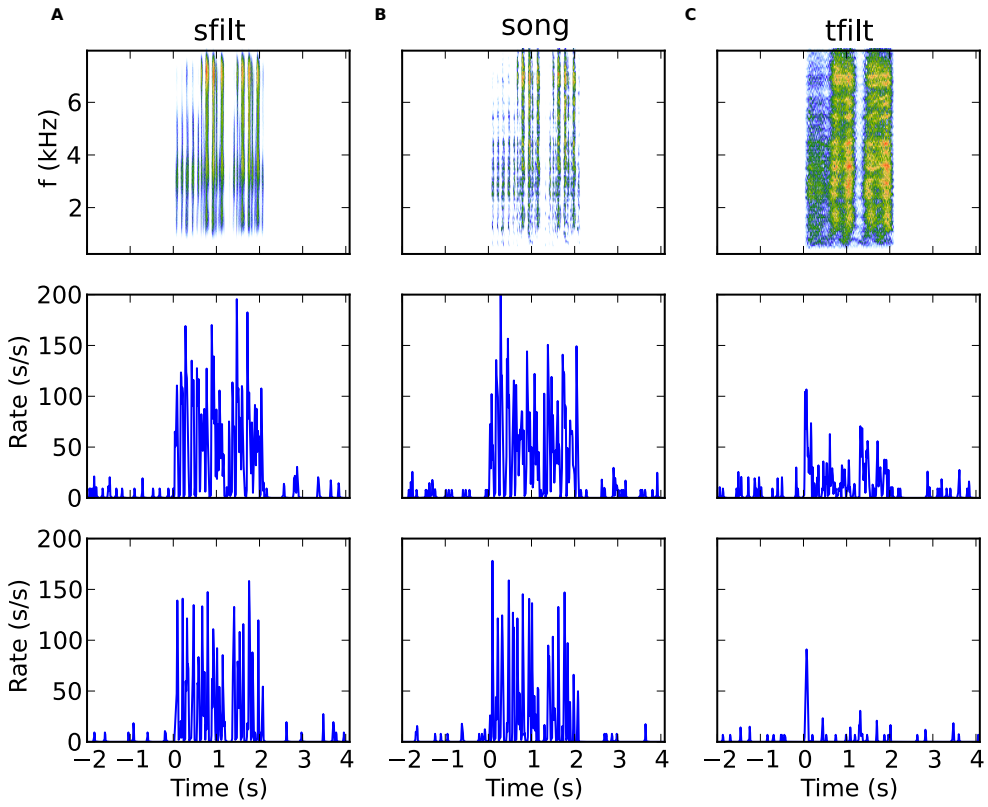


Figure 2.4: Two Single-Units Show Differing Responses to Temporally Filtered Stimuli. Top row: log-power spectrograms of three sounds. Middle row: response of unit 1497 to the sounds in the top row, shown as a PSTH. Bottom row: response of unit 3018 to same sounds. **A: Response to sfilt stimulus.** Spectral features of the sounds are smeared, but sharp temporal features are preserved. Both units respond robustly to the stimulus with a high degree of phase-locking. **B: Response to unfiltered song.** No processing was done for this stimulus, and the spectrogram shows the full range of spectotemporal features. Both units respond similarly to the sfilt stimulus, with similar spike rates and similar phase-locking. **C: Response to sfilt stimulus.** The spectrogram shows the high degree of temporal smearing, but a few strong, narrowband spectral features are visible. Unit 1497 responds much less strongly to this stimulus, but does respond throughout the presentation. Unit 3018, in contrast, shows the onset excitation that characterizes most of the units in this study.

Chapter 3

Modulation-Domain Noise Reduction Using a STRF Basis.

Abstract

Individual neurons and ensembles of neurons can create representations of behaviorally relevant acoustical signals that are invariant to distortions from propagation or noise. Here, we describe a biologically inspired noise-filtering algorithm that can be used to separate song or speech from noise. Because the algorithm uses spectrotemporal receptive fields as a bank of model neurons, we have demonstrated that the computations performed by (STRFs) can indeed explain noise invariance.

Introduction

In this chapter, I will lay out a framework for quasi-real-time noise reduction that combines the findings on invariance from Chapter 1 with a broader knowledge of the statistics of vocalizations. The modeling work was a collaboration with Tyler Lee, who helped code and test the algorithm.

To show that noise invariance and thus noise filtering can be obtained from a modeling implementation of the observed data, we engineered a noise filtering algorithm based on a decomposition of the sound by an ensemble of "artificial" neurons described by realistic STRFs. This ensemble of artificial neurons can be thought of as a modulation filter bank because the response of each neuron quantifies the presence and absence of particular spectro-temporal patterns as observed in a spectrogram and, contrary to a frequency filter bank, not necessarily the presence or absence of energy at a particular frequency band. A similar decomposition has been proposed and used for the efficient processing of speech and other complex signals (Chi, Ru, & Shamma, 2005; Mesgarani, David, Fritz, & Shamma, 2008; Shamma, 2001). To implement noise filtering, we weighted the response of the model neurons to emphasize the representation obtained from the synthetic neurons that were most noise-invariant or, equivalently that provided the best representation to extract the signal from the noise. These weighted model neural responses could then be used to recover the signal from noise by generating a set of time-varying frequency gains.

In creating a noise-reduction algorithm, we are seeking to extract noise-invariant features of the vocalizations and use them to perform a reconstruction with improved SNR. In the context of invariance, we seek a reconstruction that is invariant to background noise. In other words, the goal for constructing a noise-reduction system should simply be that, for inputs consisting of signal plus noise, the output should reproduce the signal.

Implementation

Framework

The basic algorithm comprises three pieces: an analysis filter bank, a gain computation stage, and finally a synthesis of the final waveform. The system is diagrammed in figure 3.1.

Analysis stage

The analysis stage is simply a filter bank. It decomposes the noisy input signal $x(t)$, sampled at frequency f_s, into N separate bands $y_j(t)$, each also sampled at frequency f_s. In our reference implementation, we use 62 Hz band spacing with a Gaussian frequency profile. This profile allows for exact reconstruction of the input signal by adding the bands together.

Gain stage

The gain stage is by far the most complicated. The algorithm computes a time-varying gain between 0 and 1 for each of the N bands. The exact form of this algorithm is described in the next section, but the gain for each band is based on the recent history of the signals in all of the bands. The computed gains are then applied to their respective bands, attenuating the signal at certain frequencies.

Synthesis stage

The synthesis stage combines these N gain-modified time-domain band signals to create the output waveform sampled at frequency f_s. In our reference implementation, we use simple additive mixing, which, as previously mentioned, exactly reconstructs the original signal when the bands are Gaussian and all of the gains are set to 1.

Gain stage (details)

The framework described above is relatively simple, and only differs in small part from previous algorithms. The primary difference is the algorithm used to compute the gains.

STRF-based gains

Our algorithm starts with a time-frequency representation of the signal. For prototyping we have used existing routines to compute a log-power spectrogram $S(f,t)$ sampled at $f_s = 1000\,\text{Hz}$, per equations 1.15-17. In practice, we could also compute an equivalent representation directly from the N band signals. This would involve squaring each signal, low-pass filtering below $\frac{f_t}{2}$, downsampling to f_t, and taking the log. This latter method is likely to be the best choice for an real-time/on-line implementation.

We then convolve this time-frequency representation separately along the time axis with *M* STRFs $h_i(f,\tau)$, that is, time-frequency filters:

$$a_i(t) = h_i(f,\tau) * S(f,t) \quad \textbf{(3.1)}$$

$$\mathbf{a}(t) = \left[a_i(t)\right]$$

$$i \in [1, M]$$

Each STRF has *N* bands, *P* time delays, and is sampled in time at f_t. The output from this stage, called the "activations" $\mathbf{a}(t)$, has *M* channels and is sampled at f_t. This vector represents the projection of the time-frequency stimulus into a "STRF domain", which is intended to approximate the neural representation of sounds.

In this experiment, we used STRFs constructed as the product of two Gabor functions:

$$h(f,\tau) = G(f) \cdot H(\tau) \quad \textbf{(3.2)}$$

$$G(f) = A_f e^{-0.5\left[(f-f_0)/\sigma_f\right]^2} \cdot \cos(2\pi \cdot \Omega_f (f - f_0) + \phi_f)$$

$$H(\tau) = A_t e^{-0.5\left[(\tau-\tau_0)/\sigma_t\right]^2} \cdot \cos(2\pi \cdot \Omega_t (\tau - \tau_0) + \phi_t)$$

The parameters of these Gabor functions (e.g. for time: τ_0, the temporal latency; σ_t, the temporal bandwidth; Ω_t, the best temporal modulation frequency; and ϕ_t, the temporal phase) were randomly chosen using a uniform distribution over the range of those found in area NCM (Chapter 1) and Field L (Woolley, Gill, Fremouw, & Theunissen, 2009). The number of model neurons, M, was not found to be critical as long as the population of STRFs sufficiently tiled the relevant modulation space. M was set to be 140 for the results shown.

For each of the *N* STRFs, we use a constant weight d_i, representing an estimate of how important the *i*th STRF is, represented by a vector $\mathbf{d} = \left[d_i\right]$. For the purposes of reconstruction, the weights are constant and fixed beforehand. The computation of the weights is covered in the next section.

We then construct the time-varying gains $\mathbf{g}(t) = \left[g_j(t)\right]$ by first multiplying the activation $a_i(t)$ by the importance d_i, then projecting back into the frequency domain using a time-frequency kernel $\psi_i(f,\tau)$:

$$g_j(t) = \mathrm{f}\left(\sum_{i=1}^{M} d_i a_i(t+P) * \psi_i(f_j,\tau)\right) \quad \textbf{(3.3)}$$

The offset term *P* is the number of STRF delays, and enforces a causality constraint on the system.The function f was chosen to be the logistic function in order to limit the gains to values

between 0 and 1. For our reference implementation, we have chosen the form of the ψ_i to be frequency-domain only, and to further be the frequency marginals of the STRFs:

$$\psi_i(f,\tau) = \psi_i(f) = \tfrac{1}{P}\sum_{\tau'=1}^{P} h_i(f,t) \qquad \textbf{(3.4)}$$

Using these gains $g_j(t)$, we then synthesized a processed signal $\hat{x}(t)$:

$$\hat{x}(t) = \sum_{j=1}^{N} g_j(t)\cdot\, y_j(t) \qquad \textbf{(3.5)}$$

where $y_j(t)$ is the narrow-band signal from the frequency filter j obtained in the time-frequency decomposition of the song + noise stimulus, $x(t)$.

We learned the weights d_i, was by minimizing the squared error $e^2(t) = (x(t) - \hat{x}(t))^2$ through gradient descent. We generated training stimuli $x(t)$ by summing together a 1.5 s song clip $s(t)$ and a randomly selected chunk $n(t)$ of either ml-noise or zebra finch colony noise of the same duration. To match the experimental results, both the song $s(t)$, and the noise $n(t)$, were first high-pass filtered above 250 Hz and low-pass filtered below 8 kHz, and then resampled to a sampling rate of 16 kHz. The song and noise were weighted to obtain a SNR of 3 dB, although similar results were found with lower SNR's.

We trained the system on all instances of $x(t)$, and determined weights d_i by averaging across values obtained through jack-knifing across this data set ten times with 10% of the data held out as an early stopping set.

We assessed noise reduction performance with by validating on a novel song in novel noise. We computed the cross-correlation between the estimate and the clean signal in the log spectrogram domain. We then took the ratio of this cross-correlation and the value obtained prior to attempting to de-noise the stimulus to obtain a performance ratio, providing a lower bound of 1. We then compared our algorithm to three other spectral subtraction noise algorithms: the optimal Wiener filter (OWF), a variable gain algorithm patented by Sonic Innovations (SINR) and the ideal binary mask (IBM). The optimal Wiener filter is a frequency filter whose static gain depends solely of the ratio of the power spectrum of the signal and signal + noise.

In our implementation, the Wiener filter was constructed using the frequency power spectrum of signal and noise from the training set and then applied to a stimulus from the testing set (of the same class).

The spectral subtraction algorithm for Sonic Innovations uses a time variable gain just as in our implementation. Also, as in our implementation, the analysis step for estimating this gain was based on the log of the amplitude of the Fourier components. However, the gain function itself was estimated not from a modulation filter bank but estimating the statistical properties of the envelope of the signal and noise in each frequency band (US Patent 6,357,395 B1). We used a Matlab implementation of the SINR algorithm provided to us by Dr. William Woods of Starkey Hearing Research Center, Berkeley, CA. Optimal parameters for the level of noise reduction and the estimation of the noise envelope for that algorithm were also obtained on the training signal and noise stimuli and the performance was cross-validated with the test stimuli. The IBM procedure used a zero-one mask applied to the sounds in the spectrogram domain. The

mask is adapted to specific signals by setting an amplitude threshold. Binary masks require prior knowledge of the desired signal and thus should be seen as an approximate upper bound on the potential performance of general noise reduction algorithms. Although these simulations are far from comprehensive, they allowed us to compare our algorithm to a lower bound (the noisy stimulus), to optimal classical approaches for Gaussian distributed signals (OWF), to a very recent state-of-the-art algorithm (SINR), and to an upper bound (IBM).

Results

Noise reduction performance

We assessed the performance of our algorithm by comparing it to 3 other noise reduction schemes: the optimal classical Wiener filter for stationary Gaussian signals (OWF), a state-of-the-art spectral subtraction algorithm (SINR), and the upper bound obtained by an ideal binary mask (IBM). For all cases, we used recordings of undirected zebra finch song as the foreground signal. For the background noise, we used either ml noise, or recordings made in our breeding colony room.

We measured the performance of each algorithm using crosscorrelation (Pearson's r) in the log-power spectrogram domain. As a baseline, we computed the correlation between the signal and the signal + noise. Our presumption is that no noise filtering algorithm should do worse than the noisy signal.

The optimal Wiener filter is a frequency filter whose static gain depends solely of the ratio of the power spectrum of the signal and signal + noise. In a similar approach to ours, the SINR method involves varying the gain in the spectrographic domain while taking into account the structure of noise and signal envelope. In the SINR, the estimation of this variable gain, however, is not based on a biologically inspired analysis of the sounds but instead on engineering and statistical principles. The IBM procedure uses a zero-one mask applied to the sounds in the spectrogram domain. Because the binary mask requires prior knowledge of the desired signal, it is useful only as an upper limit of performance.

In all cases, our algorithm gave significant improvement over the noisy signal (performance ratios greater than 1, Fig.3.2A). Qualitative listening to the reconstructed signals, we verified that the algorithm introduced minimal distortion. Although it was not explicitly a goal, the algorithm also did an excellent job of removing broadband, low-frequency noise from the ventilation system in our colony room recordings.

For the case of song embedded in ml noise, our algorithm performed significantly better than either the classical Wiener filter or the SINR (Fig. 3.2A). Thus, for some background signals, our simple implementation outperforms the state-of-the-art. Listening to the reconstructed songs, we verified that the algorithm does remove much of the background noise, while keeping good sound quality.

When the song was embedded in colony noise, our algorithm performed similarly to the SINR algorithm but worse than the OWF (Fig 3.2A). From listening to the reconstructed signals, the quality is still good, but we found that some vocalizations are present in the gaps in the foreground song. The OWF has static spectral gain, while our algorithm has static modulation domain gains once trained. This may mean that, because the modulation statistics of the vocal background are identical to the foreground song, our algorithm applies less attenuation to them

than the OWF. Explicit detection of silences could improve this, as could a level-dependent gain like the one present in the SINR algorithm.

While our algorithm does not perform nearly as well as the IBM, visual inspection shows that we obtained a spectrographic filtering that is similar to the one that one would obtain from a preset binary mask algorithm (Fig 3.2C bottom).

Discussion

We have shown that a spectro-temporal basis can be used to remove background noise from speech signals, but we are not the first to suggest either a time-frequency or a modulation spectrum basis for speech identification and processing or for noise reduction.

Comparison with other algorithms

All of the noise-reduction models we will treat here fall into the category of *spectral subtraction* algorithms (Boll, 1979). As defined by Boll, this involves computing an approximate noise spectrum and then subtracting it from the noisy speech spectrum. All spectral subtraction/analysis-synthesis methods use a gain computation stage. The crucial difference in ours is in the gain compuatation.

The simplest form of this sort of stream separation is a binary mask, with ones and zeros representing signal and noise, respectively. This can be performed in the time domain alone, but it is commonly extended to the time-frequency domain as well (Brown & Cooke, 1994). The basic challenge involves detecting which temporal and spectral regions should be masked and which should be unmasked. Wang et. al. used a pitch tracker to perform this task, detecting temporal regions of signal from the presence of harmonic pitch and spectral regions from the harmonics of the fundamental (Hu & Wang, 2004). Our noise reduction algorithm does not compute an explicit harmonic pitch or pitch contour, but pitch contours fall within the spectrotemporal modulation space tiled by our receptive fields. Wang et. al. have remarked in their work that detection of non-pitchy vocalization sounds still presents a significant challenge when constructing this sort of algorithm; our approach in principle detects a wider variety of speech sounds, including pitch but also including broad spectral modulations corresponding to formants and sharp, broadband features involved in syllable boundaries and consonants.

One particularly successful speech denoising algorithm is implemented in hearing aids by Sonic Innovations, Inc. (Sonic Innovations, Inc., 2000). The primary advance is that rather than use binary spectrotemporal masking, it explicitly estimates the SNR and uses this to compute a variable gain. This performs two tasks simultaneously. First, it accounts for uncertainty about the actual presence of speech, smoothly lowering the gain in a band according to the amount of noise it contains. The second important task is that it leaves some of the original signal intact even when noise is present; the presence of even pure white noise in spectral gaps has been shown to improve speech intelligibility over the presence of silence. Our noise rejection algorithm resembles the Sonic Innovations algorithm in many ways, employing a similar analysis-gain-synthesis cascade. The primary difference in principle, however, is that we use information from all bands to compute the individual band gains.

Perhaps the most similar algorithm to ours is the basis-decomposition method of Smaragdis et. al. (Smaragdis, 2007). Using NMF to perform blind source separation, the author found features in human speech that closely resemble the modulation features we have previously observed. Unlike Smaragdis' algorithm, we use a basis set that tiles the observed

space of neural tuning rather than using features learned from a limited sample of utterances. In principle, though, our algorithm performs a very similar computation in a way that is feasible for real-time systems.

Although it explicitly performs noise reduction based on the spectrogram, our method implicitly works in the modulation power domain, by means of the de-facto reduced modulation power space spanned by the STRFs. As such, it draws on this group's work on modulations but also on the extensive work done by Shamma's group, both on the importance of spectrotemporal modulations and the use of the STRF basis {Elhilali:2008eo}.

Performance measurement

In order to assess the performance of any noise reduction algorithm, one first needs to determine what sort of errors are important. For the initial work, we have used the correlation in the spectrogram domain, but this is by no means the only option.

Pyschophysical assessment

The gold standard for testing performance of any speech noise reduction algorithm is to have human subjects listen to noise-corrupted sounds and compare some metric of their comprehension with and without the algorithm. The drawback is that this process is extremely time-consuming and expensive compared to automated metrics. Additionally, while this lab is set up to perform this sort of psychophysical experiment, many signal processing labs are not, meaning that only very promising algorithms are ever tested with human listeners, and then only long after they are originally described.

Automatic speech recognition

A faster, cheaper, more accessible alternative to a full psychophysical test is to replace the human listener with an automatic speech recognition (ASR) system. Because it takes linguistic factors into account, an ASR method can provide a better model of comprehension. This method also has the advantage of being easy to standardize across many different experimenters. At the same time, this method is more than merely a poor-man's psychophysical test: improved preprocessing for ASRs is an important application for noise reduction algorithms in its own right.

Ad hoc metrics

Besides the two previous metrics, there are a number of simpler ways to measure performance. I am calling these methods *ad hoc*, because while they are sensible from a signal processing perspective, there is no guarantee that they represent the actual quality for human or machine listeners. A reasonable compromise for using any of these *ad hoc* metrics is to calibrate them against the gold standard of psychophysics. Whatever metric gives best accordance with the psychophysics could then be used for fast training and validation of new routines.

There are two decisions involved in choosing an *ad hoc* metric: the domain of the signal, and the form of the error. By *domain*, I simply mean the representation of the signals, e.g. time-domain, spectrogram, cochleogram, etc.

The simplest choice for the stimulus representation is the time domain, i.e., the sound-pressure waveform. This has the advantage that it is guaranteed to contain all of the information in the waveform. It will, however, represent some features that are imperceptible to listeners (EHMER, 1959) and some that are perceptible but irrelevant to intelligibility (Elliott & Theunissen, 2009).

One solution to the issues with the time domain is to compute a perceptual representation of the various sounds. A spectrogram, or log spectrogram, is a reasonable first approximation. Further transformations, for instance a representation of the auditory nerve fiber bandwidths like the mel or Bark transformations of the frequency scale, could also be useful. Spectral loudness correction, like the equal loudness correction found in Hermansky's PLP framework (Hermansky, 1990), may also help: high-frequency bands have very low absolute power, but can still have reasonable SNR and are important to listener performance. Besides spectrographic representations, explicitly neural models of perception, such as a cochleogram (Lyon, 1982) or model of auditory nerve responses (Patterson et al., 1992), would fall under this rubric.

All perceptual representations address the primary problem with the time-domain representation, because they emphasize the perceptible features of the sound. The forms differ in how much imperceptible and irrelevant information is suppressed. They may also represent features in terms of perceptual salience rather than simple power, as mentioned for the equal loudness curve above, or in terms of cross-band masking, as represented in the MP3 compression algorithm.

The primary problem with any of the perceptual metrics is that they may eliminate perceptible features. Nearly all of the aforementioned transformations, for instance, immediately take spectral amplitude and discard phase; only cochleograms or neurograms retain any phase information. Other features could also, in principle, be lost.

Having chosen a domain for the signal, the next task is to choose a form for the error. The advantages and drawbacks of each are not as clear as for our choice of the signal domain. Any decision would best be handled by comparing the performance to a standard like human listener comprehension.

Far and away the simplest error form is the correlation coefficient (Pearson's r). By extension, we could also use the coherence. Perhaps the most tractable metric is the mean squared error. This metric is easy to implement for fitting weights by gradient descent. The most subtle form is a computation of the SNR. A priori this seems likely to be the best fit to the perceptual data, but could make computation of a gradient difficult.

Human speech

For our initial tests, we have used zebra finch song and either synthetic background noise (ML noise), or recordings of colony noise. Regardless of what error metric we choose, creating a system that performs well on human speech requires realistic recordings both of speech and of background noise both for training and for testing.

Many corpora of speech sounds are available for testing and training speech recognition systems and for performing psychophysical tasks. Most focus either on general recognition of natural speech (TIMIT), on simplified speech (Iowa corpus), or on specific tasks like recognition of spoken numbers (TIDIGITS). As such, these corpora do not address background noise at all, and many, like the Iowa corpus, do not represent natural speech accents or prosody.

Two corpora, however, show special promise for this application. The most applicable is COSINE, COnversational Speech In Noisy Environments, and is more explicitly tailored for measuring speech-in-noise performance. This corpus contains multiple simultaneous microphone recordings of quasi-natural conversations in a variety of noisy natural environments. It is available free from the University of Washington's website: http://ssli.ee.washington.edu/cosine/

The COSINE corpus would be especially useful because it contains clear recordings, noisy recordings, and transcripts of the same recordings from multiple positions. Specifically, each talker wore both a throat mic and a "close-talking" headset mic, the combination of which

provide recordings with a high gain for the talker relative to the background. Additionally, each talker wore a shoulder-mounted mic and a four-microphone array on their chest, the former giving omnidirecitonal recordings and the second giving directional recordings for positional recordings.

The second major corpus for this purpose is the ICSI Meeting Recorder project corpus, a high-quality corpus from a major speech-processing group: http://www.icsi.berkeley.edu/Speech/mr/

The Meeting Recorder corpus is the available via subscription to the UPenn Language Data Consortium, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004T04

Like COSINE, The Meeting Recorder corpus contains simultaneous, multi-microphone recordings of multiple talkers, with transcription. Recordings were made from headset mics and from ambient, tabletop mics. The corpus also contains measurements of the room acoustics. Unlike the COSINE project, though, the context is specifically meetings, is indoors and from only one room.

System modifications

We have a number of avenues available to improve the system.

Choice of STRF basis

A crucial question here is what STRF basis should be used to perform reconstruction/de-noising. For the tests so far, we have used synthetic, time-frequency-separable Gabor filters, but a number of other options present themselves. Many of these are in fact explicitly nonseparable, which could improve noise reduction. Slow FM sweeps are a major component both of birdsong and human speech, and the decomposition of (Smaragdis, 2007) clearly shows a number of this sort of feature.

Because our lab has a large corpus of STRFs computed from birdsong, we could use this exact set as the basis for the decomposition when working with song. The main issue here is completeness: there is no guarantee that this corpus samples the space efficiently enough to give optimal reconstruction. We also cannot easily extend it to process human speech because of the quantitative differences in the modulation power spectrum of those two classes of sound. Although all animal vocalizations can be expected to show the qualitative segregation of features to the regions near the spectral and temporal modulation axes, their exact positions will differ. Zebra finch pitch, for instance, tends to fall in the 400Hz range ($\Omega_f \approx 2.2/kHz$), while human males often have fundamental pitch well below 200Hz ($\Omega_f > 5/kHz$). The formants, too, are heavily influenced by vocal tract length (Taylor & Reby, 2010). As vocal tract length is more than an order of magnitude different between the two species, somewhat different resonances can be expected. Syllable rate, too, differs: finch syllables tend to come at a rate of roughly 10/*s*, while that rate in human speech is more akin to formant transitions, with syllables coming at 2-3/*s*.

A solution to this problem comes from the use of a set of synthetic STRFs with similar properties. For instance, we could define a region of the modulation power domain and synthesize STRFs that tile it. If we define a scaling transformation between the average MPS of human speech and ZF song, we could simply scale the tilling found in finches. We could also simply evenly tile this analogous region of modulation power space.

One can also choose a STRF basis based on the statistical structure of the desired stimulus. For instance, the STRFs could be synthetic but, instead of reproducing the modulation

transfer functions of realistic neurons, they could instead be tuned to the portions of the modulation spectrum where the stimulus has power.

Finally, we do not have to choose a subspace for the STRFs *a priori*. If our STRF basis tiles the entirety of some space, again e.g. the allowed region of modulation power space, then training the importance weights **d** indicate what region of said space is important for reconstruction. We expect that the weights for some combined domain and form of the error should reflect the observed physiological tuning, although that is in itself an experiment.

Finally, we can view the STRFs as a basis set for the stimulus, and choose them based on some optimal reconstruction constraint. For instance, the basis set used by Smaragdis would fall in this category. Another possible choice would be to train a convolutional basis set, with a generative formulation identical to Smaragdis or to the convolutional movie basis of Olshausen (Olshausen, 2002), and train the fields based on a sparsity constraint. The most extreme version of this approach would involve training the STRF set in conjunction with the importance weighting.

All of the above methods rely on predetermined STRF kernels. In principle, however, it may be possible to fit the STRFs as part of the optimization procedure. The feasibility of this depends heavily upon the changes that this makes to the calculation of the gradient; our current system has the gradient available in closed-form, but this change could break that condition.

Choice of output kernel

Our current system performs reconstruction by projecting back into the gain space through the use of frequency-only kernels. As is represented in equation 3.2, the kernel can in principle also have a temporal component. An obvious choice then is to simply use the STRFs themselves, whether separable or not.

Silence detection

We may also be able to improve our algorithm by incorporating an explicit silence detector. The performance measures in figure 3.2A indicate that while our algorithm outperforms both the Sonic Innovations algorithm and the optimal Weiner filter when the background noise is ml noise, the OWF performs better with a background of colony noise and the Sonic Innovations algorithm is the same. Although we would need to replicate this finding for human speech, it is not entirely surprising: ml noise, although it overlaps with zebra finch song in the modulation domain, is less similar to it than the colony background. With the colony noise, our system is effectively faced with a poorly-determined problem: provided that it identifies a sound as a vocalization, does that sound come from the foreground or the background?

Creating an explicit representation of the gaps in the foreground signal, and reducing all of the gains during them, would help with this issue. Close examination of the bottom (gain) panels of Figure 3.2C and D suggests that this is indeed one of the problems for our algorithm when dealing with background voices.

Implications

Our demonstration of the noise-invariance of spectrotemporal feature responses has allowed us to improve upon the state of the art in noise reduction. Background noise, especially from many voices, is a major problem for hearing aid wearers: intrusion of the background can interfere with intelligibility, and worse, many wearers find the output to be so unpleasant in noisy surroundings that they find it preferable to turn off the hearing aid. Clearly, improving intelligibility is a major goal for such systems, but even failing that, improving the pleasantness of the signal would improve the intelligibility in practice.

The main roadblock for hearing aids is the speed at which our algorithm can work. Delays above 80ms are unacceptable in practice, and our STRF basis alone requires 100ms delays. Filter delays and other computational overhead mean that practical feature detection needs to closer to 10ms. Our most invariant STRFs for finch song (Chapter 1) look for modulations in the range $50Hz < \Omega_t < 80Hz$. Human syllable and formant transitions are much slower, though pitch transitions are faster than the formants.

One possible use for a spectrotemporal basis in fast computation, then, is to compute a slower state variable for the faster system. For instance, power detected by slow pitchy STRF features could indicate simply the presence or absence of human speech. Fast features could still be used for processing, effectively suppressing fast transitions that are not broadband.

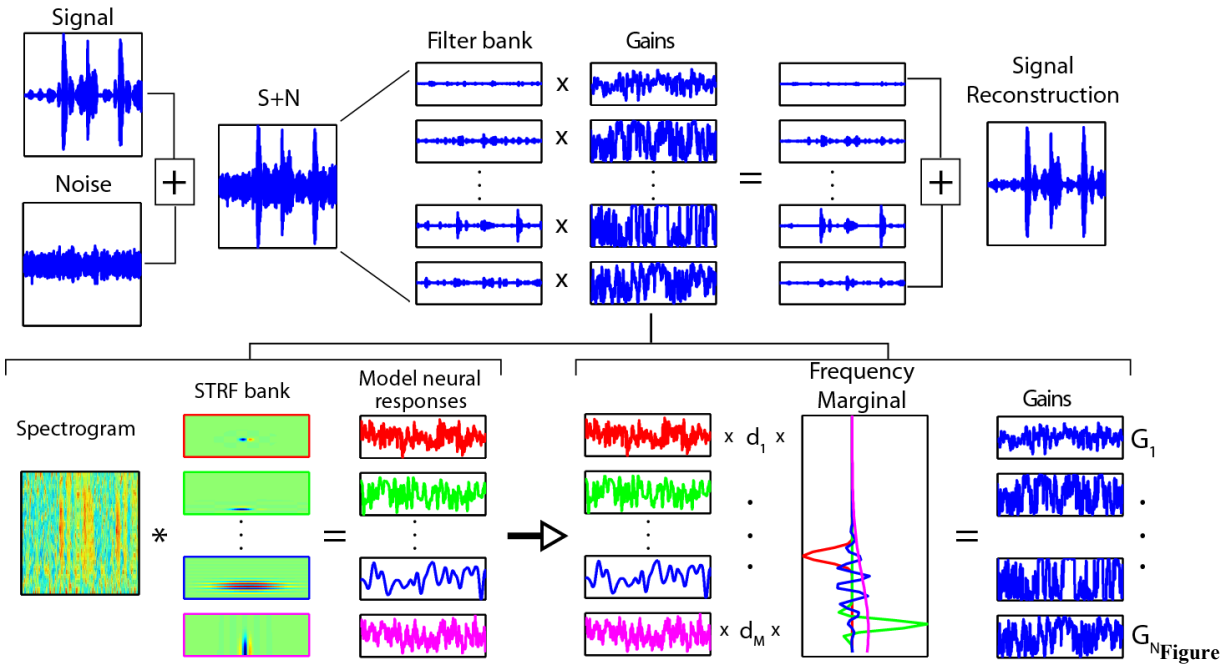
These stringent speed requirements in hearing aids are somewhat relaxed with telecommunications systems, where delays greater than one second are acceptable. Used in a cellular handset, this technology could perform background noise reduction, resulting in better call quality in a noisy environment.

Another appealing potential application is in improving the preprocessing stage in ASR systems. Current algorithms are almost exclusively based on the cepstrum. As discussed in the introduction, the cepstrum does a very good job of taking the spectral modulations into account, but does not represent the interdependency of spectral and temporal modulations. Much of the success of such systems lies in the combination of the cepstral features with an explicit representation of the sequential structure inherent in the language (Gold & Morgan, 2000). Modern ASR systems do take temporal dependencies into account, but without taking the nonseparability of modulation power into account.

Two primary types of temporal information are commonly used in these ASRs. The first is the delta cepstrum, which takes low-order differences among the series of cepstral coefficients (Gold & Morgan, 2000; 2000). This yields a representation of the high-frequency temporal information in the signal, which is only important for the lowest cepstral coefficients. The other temporal representation is the modulation spectrogram concept (Kingsbury, 1998), which is incorporated into the more recent RASTA-based algorithms. This causes an error in the opposite direction, smoothing out the high-frequency information for all frequencies. Low-frequency information is important at all frequencies, and it is important in all cepstral coefficients; but information about fast transitions present in consonants will be lost by this low-pass temporal filtering.

Because our algorithm accounts for the interdependence of spectral and temporal modulations, it is possible that using it as a preprocessor would separate out two different sources of temporal variability. If applied as a preprocessing step, it could account for acoustic dependencies that are intrinsic to vocal production. This might remove variability in the input signal, allowing the linguistic model to fill in regions where the acoustic information is more uncertain.

In conclusion, our work shows promise for a number of fields. Both in the model and in the biological system, given a complete modulation filter bank, the importance weights for a given signal and noise could be learned quickly through supervised learning. Moreover, after learning, the algorithm can easily be implemented in real-time with minimal delay. We therefore propose that this noise filtering approach could be feasible both in engineering and clinical applications.



3.1. Invariant Representations and Noise Reduction. We implemented a biologically inspired noise-filtering algorithm using an analysis/synthesis paradigm (top row) where the synthesis step is based on a STRF filter bank decomposition. The bottom row shows the model neural responses obtained from a sound (spectrogram of noise-corrupted song) using the filter bank of biologically realistic STRFs. These responses are then weighed optimally with weights $d_1,..,d_M$ to select the combination of responses that are most noise-invariant. The weighted responses are then transformed into frequency space by multiplying the weighted responses by the frequency marginal of the corresponding STRF (color-matched on the figure) to obtain gains as a function of frequency. The top row illustrates how these time-varying frequency gains can then be applied to a decomposition of the sound into frequency channels allowing for the synthesis step and an estimate of the clean signal.

Figure

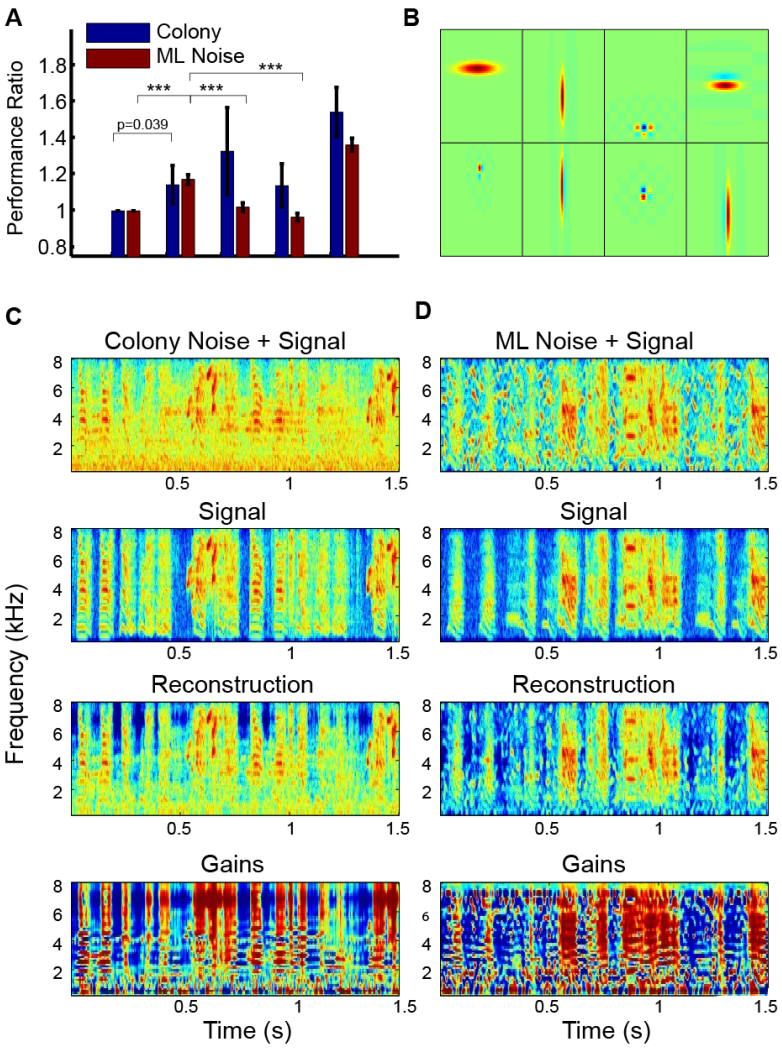


Figure 3.2. Noise filtering with a modulation filter bank. A. Performance of three noise reduction algorithms (STRF, OWF, SINR) and lower and upper bounds (Stim, IBM). The performance ratio (y-axis) depicts the improvement in noise levels over the noise-corrupted signal, as measured by the cross-correlation in the log spectrogram domain. On the x-axis are the models we have tested, where "Stim" is the noise-corrupted signal, "STRF" is the model presented here, "OWF" is the optimal Wiener filter, "SINR" is a spectral subtraction algorithm similar to STRF but based on engineering constructs, and "IBM" is an ideal binary mask. B) The 8 most heavily weighted STRFs. C) Spectrograms of the signal masked with noise from the zebra finch colony, the clean zebra finch song, and our signal reconstruction, followed by the time-frequency gains. D) Same as C but for modulation-limited noise.

Epilogue

Given that the cepstrum is the mainstay of modern speech processing and that the MPS augments it with important information, my research in neurosciences and signal processing shows that it is possible to improve our speech processing algorithms. The spectrogram and the cepstrum are important because they represent intrinsic statistical dependencies in the vocal signal. The modulation power spectrums captures additional dependencies in natural acoustical objects that improve this type of processing in a simple, powerful manner.

The MPS (and the underlying time-frequency representation of sounds by the spectrogram) is of course, not the be-all and end-all of perception. We know that this representation, although complete, encodes particular acoustical features potentially very important for perception in a dense and non-linear fashion.. For instance, the timbre of a synthetic harmonic structure depends upon the relative temporal phase of the constituent tones (PATTERSON, 1987). While this information is recoverable in the spectrogram (and the complete complex MPS), it is in a form that is not easily interpretable or, equivalently, linearly related to perception or neural properties. Thus alternative time-frequency representations that more explicitly represent such phase information might be used in the future to further understand our perception of auditory objects and its neural correlate. For example Gardner and colleagues (Gardner & Magnasco, 2005) have proposed a Hilbert-transform-based which represents amplitude and phase more explicitly. Future work might thus also investigate how perceptual invariance and neural invariance are affected by filtering (or other forms of transformations) in other representations of sounds.

References

Amin, N., Grace, J. A., & Theunissen, F. E. (2004). Neural response to bird's own song and tutor song in the zebra finch field L and caudal mesopallium. *Journal Of Comparative Physiology A-Neuroethology Sensory Neural And Behavioral Physiology*, *190*(6), 469–489. doi:10.1007/s00359-004-0511-x

Arai, T., Pavel, M., Hermansky, H., & Avendano, C. (1999). Syllable intelligibility for temporally filtered LPC cepstral trajectories. *Journal Of The Acoustical Society Of America*, *105*(5), 2783–2791.

Asari, H., Pearlmutter, B., & Zador, A. M. (2006). Sparse representations for the cocktail party problem. *Journal of Neuroscience*, *26*(28), 7477.

Aubin, T., & Jouventin, P. (2002). How to vocally identify kin in a crowd: The penguin model. *Advances In The Study Of Behavior, Vol 40*, *31*, 243–277.

Aulanko, R., Hari, R., Lounasmaa, O. V., Näätänen, R., & Sams, M. (1993). Phonetic invariance in the human auditory cortex. *NeuroReport*, *4*(12), 1356–1358. doi:10.1097/00001756-199309150-00018

Barbour, D. L. (2011). Intensity-invariant coding in the auditory system. *Neuroscience & Biobehavioral Reviews*. doi:10.1016/j.neubiorev.2011.04.009

Bendor, D., & Wang, X. (2005). The neuronal representation of pitch in primate auditory cortex. *Nature*, *436*(7054), 1161–1165. doi:10.1038/nature03867

Bendor, D., & Wang, X. (2006). Cortical representations of pitch in monkeys and humans. *Current opinion in neurobiology*, *16*(4), 391–399. doi:10.1016/j.conb.2006.07.001

Bharucha, J. J., & Mencl, W. E. (1996). TWO ISSUES IN AUDITORY COGNITION:. Self-Organization of Octave Categories and Pitch-Invariant Pattern Recognition. *Psychological Science*, *7*(3), 142–149. doi:10.1111/j.1467-9280.1996.tb00347.x

Billimoria, C. P., Kraus, B. J., Narayan, R., Maddox, R. K., & Sen, K. (2008). Invariance and sensitivity to intensity in neural discrimination of natural sounds. *Journal of Neuroscience*, *28*(25), 6304–6308. doi:10.1523/JNEUROSCI.0961-08.2008

BLUMSTEIN, S., & STEVENS, K. (1979). Acoustic Invariance in Speech Production - Evidence From Measurements of the Spectral Characteristics of Stop Consonants. *Journal Of The Acoustical Society Of America*, *66*(4), 1001–1017.

Bolhuis, J. J., Zijlstra, G., Boer-Visser, den, A., & Van der Zee, E. (2000). Localized neuronal activation in the zebra finch brain is related to the strength of song learning. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, *97*(5), 2282–2285.

Boll, S. F. (1979). Suppression of Acoustic Noise in Speech Using Spectral Subtraction. *Ieee Transactions on Acoustics Speech and Signal Processing*, *27*(2), 113–120.

Boring, E. (1952). Visual perception as invariance *Psychological Review*.

Bronkhorst, A. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acustica*, *86*(1), 117–128.

Brown, G. J., & Cooke, M. (1994). Computational Auditory Scene Analysis. *Computer Speech and Language*, *8*(4), 297–336.

BUUS, S., & FLORENTINE, M. (1991). Psychometric Functions for Level Discrimination. *Journal Of The Acoustical Society Of America*, *90*(3), 1371–1380.

Cariani, P., & Delgutte, B. (1996a). Neural correlates of the pitch of complex tones .1. Pitch and pitch salience. *Journal of neurophysiology*, *76*(3), 1698–1716.

Cariani, P., & Delgutte, B. (1996b). Neural correlates of the pitch of complex tones .2. Pitch shift, pitch ambiguity, phase invariance, pitch circularity, rate pitch, and the dominance region for *pitch. Journal of neurophysiology, 76(3), 1717–1734.*

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. Journal Of The Acoustical Society Of America, 25(5), 975–979.

Chi, T., Gao, Y., Guyton, M. C., Ru, P., & Shamma, S. A. (1999). Spectro-temporal modulation transfer functions and speech intelligibility. Journal Of The Acoustical Society Of America, 106(5), 2719. doi:10.1121/1.428100

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *Journal Of The Acoustical Society Of America*, *118*(2), 887–906. doi:10.1121/1.1945807

Cutting, J. E. (1983). Four assumptions about invariance in perception *Journal of Experimental Psychology: Human Perception and Performance*, *9*(2), 310–317. doi:10.1037/0096-1523.9.2.310

Cynx, J., & SHAPIRO, M. (1986). Perception of Missing Fundamental by a Species of Songbird (Sturnus-Vulgaris). *Journal Of Comparative Psychology*, *100*(4), 356–360.

Cynx, J., WILLIAMS, H., & NOTTEBOHM, F. (1990). Timbre Discrimination in Zebra Finch (Taeniopygia-Guttata) Song Syllables. *Journal Of Comparative Psychology*, *104*(4), 303–308.

Dean, I., Harper, N., & McAlpine, D. (2005). Neural population coding of sound level adapts to stimulus statistics. *Nature Neuroscience*, *8*(12), 1684–1689. doi:10.1038/nn1541

Depireux, D., Simon, J., Klein, D., & Shamma, S. A. (2001). Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *Journal of neurophysiology*, *85*(3), 1220–1234.

Drullman, R., Festen, J. M., & Plomp, R. (1994). Effect of Temporal Envelope Smearing on Speech Reception. *Journal Of The Acoustical Society Of America*, *95*(2), 1053–1064.

Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)* (1st ed. p. 456). Chapman and Hall/CRC.

EHMER, R. (1959). Masking by Tones vs Noise Bands. *Journal Of The Acoustical Society Of America*, *31*(9), 1253–1256.

Elhilali, M., & Shamma, S. A. (2008). A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation. *Journal Of The Acoustical Society Of America*, *124*(6), 3751–3771. doi:10.1121/1.3001672

Elliott, T. M., & Theunissen, F. E. (2009). The Modulation Transfer Function for Speech Intelligibility. *PLoS Computational Biology*, *5*(3), e1000302.

Freiwald, W. A., & Tsao, D. Y. (2010). Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System. *Science*, *330*(6005), 845–851. doi:10.1126/science.1194908

Gardner, T., & Magnasco, M. O. (2005). Instantaneous frequency decomposition: An application to spectrally sparse sounds with fast frequency modulations. *Journal Of The Acoustical Society Of America*, *117*(5), 2896–2903. doi:10.1121/1.1863072

Geffen, M. N., Gervain, J., Werker, J. F., & Magnasco, M. O. (2011). Auditory perception of self-similarity in water sounds *Frontiers in integrative neuroscience*, *5*, 15. doi:10.3389/fnint.2011.00015

Gilinsky, A. (1955). The effect of attitude upon the perception of size. *The American Journal of Psychology*.

Gill, P. R., Zhang, J., Woolley, S. M. N., Fremouw, T., & Theunissen, F. E. (2006). Sound representation methods for spectro-temporal receptive field estimation. *Journal Of Computational Neuroscience*, *21*(1), 5–20. doi:10.1007/s10827-006-7059-4

Gill, P. R., Woolley, S. M. N., Fremouw, T., & Theunissen, F. E. (2008). What's that sound? Auditory area CLM encodes stimulus surprise, not intensity or intensity changes. *Journal of neurophysiology*, *99*(6), 2809–2820. doi:10.1152/jn.01270.2007

Gold, B., & Morgan, N. (2000). Speech and audio signal processing. processing and perception of speech and music (p. 537). Wiley.

Griffin, D. W., & Lim, J. S. (1984). Signal Estimation From Modified Short-Time Fourier-Transform. Ieee Transactions on Acoustics Speech and Signal Processing, 32(2), 236–243.

Griffiths, T. D., & Warren, J. D. (2004). Opinion: What is an auditory object Nature Reviews Neuroscience, 5(11), 887–892. doi:10.1038/nrn1538

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal Of The Acoustical Society Of America*, *87*(4), 1738. doi:10.1121/1.399423

Hsu, A., Borst, A., & Theunissen, F. E. (2004). Quantifying variability in neural responses and its application for the validation of model predictions. *Network-Computation In Neural Systems*, *15*(2), 91–109. doi:10.1088/0954-898X/15/2/002

Hu, G., & Wang, D. (2004). Monaural speech segregation based on pitch tracking and amplitude modulation. *Ieee Transactions On Neural Networks*, *15*(5), 1135–1150. doi:10.1109/TNN.2004.832812

ITO, M., TAMURA, H., FUJITA, I., & TANAKA, K. (1995). Size and Position Invariance of Neuronal Responses in Monkey Inferotemporal Cortex. *Journal of neurophysiology*, *73*(1), 218–226.

Kaufman, L., & Kaufman, J. (2000). Explaining the moon illusion. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, *97*(1), 500–505.

Kingsbury, B. (1998). Robust speech recognition using the modulation spectrogram. *Speech Communication*, *25*(1-3), 117–132. doi:10.1016/S0167-6393(98)00032-6

Knudsen, D. P., & Gentner, T. Q. (2010). Mechanisms of song perception in oscine birds. *Brain and Language*, *115*(1), 59–68. doi:10.1016/j.bandl.2009.09.008

Lohr, B., & Dooling, R. J. (1998). Detection of changes in timbre and harmonicity in complex sounds by zebra finches (Taeniopygia guttata) and budgerigars (Melopsittacus undulatus). *Journal Of Comparative Psychology*, *112*(1), 36–47.

Lutfi, R. A. (2007). *Springer Handbook of Auditory Research*. (W. A. Yost, A. N. Popper, & R. R. Fay, Eds.)Springer Handbook of Auditory Research (Vol. 29, pp. 13–42). Boston, MA: Springer US. doi:10.1007/978-0-387-71305-2_2

Lyon, R. (1982). ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing. In *IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 7, pp. 1282–1285). Presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Institute of Electrical and Electronics Engineers. doi:10.1109/ICASSP.1982.1171644

Malsburg, von der, C., & Schneider, W. (1986). A Neural Cocktail-Party Processor. *Biological Cybernetics*, *54*(1), 29–40.

MARGOLIASH, D., & Fortune, E. S. (1992). Temporal and Harmonic Combination-Sensitive Neurons in the Zebra Finch Hvc. *Journal of Neuroscience*, *12*(11), 4309–4326.

MELLO, C., NOTTEBOHM, F., & Clayton, D. F. (1995). Repeated Exposure to One Song Leads to a Rapid and Persistent Decline in an Immediate-Early Genes Response to That Song in Zebra Finch Telencephalon. *Journal of Neuroscience*, *15*(10), 6919–6925.

Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2008). Phoneme representation and classification in primary auditory cortex. *Journal Of The Acoustical Society Of America*, *123*(2), 899–909. doi:10.1121/1.2816572

Miller, L. M., Escabi, M. A., Read, H. L., & Schreiner, C. E. (2002). Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex *Journal of neurophysiology*, *87*(1), 516–527.

Mooney, R. (2001). Auditory representation of the vocal repertoire in a songbird with multiple song types. *Proceedings of the National Academy of Sciences*, *98*(22), 12778–12783. doi:10.1073/pnas.221453298

Nagel, K. I., & Doupe, A. J. (2008). Organizing principles of spectro-temporal encoding in the avian primary auditory area field L. *Neuron*, *58*(6), 938–955. doi:10.1016/j.neuron.2008.04.028

Narayan, R., Best, V., Ozmeral, E., McClaine, E. M., Dent, M. L., Shinn-Cunningham, B. G., & Sen, K. (2007). Cortical interference effects in the cocktail party problem. *Nature Neuroscience*, *10*(12), 1601–1607.

Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I. H., Saberi, K., Serences, J. T., et al. (2010). Hierarchical Organization of Human Auditory Cortex: Evidence from Acoustic Invariance in the Response to Intelligible Speech. *Cerebral Cortex*, *20*(10), 2486–2495. doi:10.1093/cercor/bhp318

Olshausen, B., & Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609.

Olshausen, B. (2002). Sparse codes and spikes (pp. 257–272). Probabilistic models of the brain: Perception and neural function.

Olver, P. J. (1999). *Classical Invariant Theory (London Mathematical Society Student Texts)* (First Edition. p. 304). Cambridge University Press.

Paavilainen, P., Jaramillo, M., Naatanen, R., & Winkler, I. (1999). Neuronal populations in the human brain extracting invariant relationships from acoustic variance. *Neuroscience Letters*, *265*(3), 179–182.

PATTERSON, R. (1987). A Pulse Ribbon Model of Monaural Phase Perception. *Journal Of The Acoustical Society Of America*, *82*(5), 1560–1586.

Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., & Allerhand, M. (1992). Complex sounds and auditory images. *Simulation*, *83*(1992). Pergamon.

Phan, M. L. (2006). Early auditory experience generates long-lasting memories that may subserve vocal learning in songbirds. *Proceedings of the National Academy of Sciences*, *103*(4), 1088–1093. doi:10.1073/pnas.0510136103

Pinaud, R., Terleph, T. A., Tremere, L. A., Phan, M. L., Dagostin, A. A., Leão, R. M., Mello, C. V., et al. (2008). Inhibitory network interactions shape the auditory processing of natural communication signals in the songbird auditory forebrain *Journal of neurophysiology*, *100*(1), 441–455. doi:10.1152/jn.01239.2007

Priebe, N., Cassanello, C., & Lisberger, S. (2003). The neural representation of speed in macaque area MT/V5. *Journal of Neuroscience*, *23*(13), 5650–5661.

Quenouille, M. (1956). JSTOR: Biometrika, Vol. 43, No. 3/4 (Dec., 1956), pp. 353-360. *Biometrika*.

Rauschecker, J., Tian, B., & Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science*, *268*(5207), 111–114. doi:10.1126/science.7701330

Ribeiro, S., Cecchi, G. A., Magnasco, M. O., & Mello, C. (1998). Toward a song code: Evidence for a syllabic representation in the canary brain. *Neuron*, *21*(2), 359–371.

Sadagopan, S., & Wang, X. (2008). Level invariant representation of sounds by populations of neurons in primary auditory cortex. *Journal of Neuroscience*, *28*(13), 3415–3426. doi:10.1523/JNEUROSCI.2743-07.2008

Schreiber, S., Fellous, J., Whitmer, D., Tiesinga, P., & Sejnowski, T. J. (2003). A new correlation-based measure of spike timing reliability. *Neurocomputing*, *52*, 925–931.

Shamma, S. A. (2001). On the role of space and time in auditory processing. *Trends in Cognitive Sciences*, *5*(8), 340–348.

Singh, N. C., & Theunissen, F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *Journal Of The Acoustical Society Of America*, *114*(6), 3394–3411. doi:10.1121/1.1624067

Smaragdis, P. (2007). Convolutive speech bases and their application to supervised speech separation. *Ieee Transactions On Audio Speech And Language Processing*, *15*(1), 1–12. doi:10.1109/TASL.2006.876726

Smith, E., & Lewicki, M. (2006). Efficient auditory coding. *Nature*, *439*(7079), 978–982. doi:10.1038/nature04485

Sonic Innovations, Inc. (2000, January 12). Noise reduction apparatus and method. US Patent Office.

Stevenson, J. G., Hutchison, R. E., Hutchison, J. B., Bertram, B. C. R., & Thorpe, W. H. (1970). Individual Recognition by Auditory Cues in the Common Tern (Sterna hirundo). *Nature*, *226*(5245), 562–563. doi:10.1038/226562a0

Stripling, R., Volman, S., & Clayton, D. F. (1997). Response modulation in the zebra finch neostriatum: Relationship to nuclear gene regulation. *Journal of Neuroscience*, *17*(10), 3883–3893.

Taylor, A. M., & Reby, D. (2010). The contribution of source-filter theory to mammal vocal communication research. *Journal Of Zoology*, *280*(3), 221–236. doi:10.1111/j.1469-7998.2009.00661.x

Terleph, T. A., Mello, C., & Vicario, D. S. (2006). Auditory topography and temporal response dynamics of canary caudal telencephalon. *Journal of Neurobiology*, *66*(3), 281–292. doi:10.1002/neu.20219

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

Thomson, D. J. (1982). Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, *70*(9), 1055–1096. IEEE. doi:10.1109/PROC.1982.12433

Tukey, J. (1958). *Bias and confidence in not quite large samples*. Annals of Mathematical Statistics.

VANGOOL, L., MOONS, T., PAUWELS, E., & OOSTERLINCK, A. (1995). Vision and Lies Approach to Invariance. *Image and Vision Computing*, *13*(4), 259–277.

Viemeister, N. F. (1979). Temporal modulation transfer functions based upon modulation thresholds. *Journal Of The Acoustical Society Of America*, *66*(5), 1364. doi:10.1121/1.383531

Viemeister, N. F., & Bacon, S. P. (1988). Intensity discrimination, increment detection, and magnitude estimation for 1-kHz tones *Journal Of The Acoustical Society Of America*, *84*(1), 172–178.

Vignal, C., Attia, J., Mathevon, N., & Beauchaud, M. (2004a). Background noise does not modify song-induced genic activation in the bird brain. *Behavioural Brain Research*, *153*(1), 241–248. doi:10.1016/j.bbr.2003.12.006

Vignal, C., Mathevon, N., & Mottin, S. (2004b). Audience drives male songbird response to partner's voice. *Nature*, *430*(6998), 448–451. doi:10.1038/nature02645

Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, *14*(4), 715–770.

Woolley, S. M. N., Fremouw, T., Hsu, A., & Theunissen, F. E. (2005). Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nature Neuroscience*, *8*(10), 1371–1379. doi:10.1038/nn1536

Woolley, S. M. N., Gill, P. R., Fremouw, T., & Theunissen, F. E. (2009). Functional Groups in the Avian Auditory System. *Journal of Neuroscience*, *29*(9), 2780–2793. doi:10.1523/JNEUROSCI.2042-08.2009

Yost, W. A. (1987). Temporal changes in a complex spectral profile. *Journal Of The Acoustical Society Of America*, *81*(6), 1896. doi:10.1121/1.394754